2023

Machine Learning Models to automate Radiotherapy Structure Name Standardization

Priyankar Bose
Virginia Commonwealth University

THESIS

MACHINE LEARNING MODELS TO AUTOMATE RADIOTHERAPY STRUCTURE NAME STANDARDIZATION

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

by

PRIYANKAR BOSE

B.Tech., Kalinga Institute of Industrial Technology - July 2014 to May 2018

Director: Thesis Preetam Ghosh,

Interim Chair and Professor, Department of Computer Science

Virginia Commonwewalth University

Richmond, Virginia

July, 2023

Acknowledgements

I am obliged to my advisor, Dr. Preetam Ghosh, for his constant support, valuable suggestions, and sensible criticism to inspire and improve my work.

I want to acknowledge Dr. Thang Dinh and Dr. William C. Sleeman IV for serving on my thesis committee.

My sincere thanks to my collaborators Dr. Jatinder Palta, Rishabh Kapoor and Dr. Sriram Srinivasan for their encouragement, suggestions, and valuable guidance for improving my understanding of the radiation oncology domain.

I am also immensely thankful to my labmates especially Dr. Pratip Rana and Dr. Syed Khajamoinuddin for helping me understand both the technology and the radiation oncology domain. I have been extremely fortunate in having lab mates like Andrii Cherevko, Joseph Pateras and Ahmad Al Musawi. Mridula Sarkar and Dr. Dipankar Bandopadhyay were a source of constant support and motivation which was of immense help to my work. Their helping hand was evident at every stage of my stress, anxiety, and venture in my research.

Finally, it is an immense pleasure to express sincere gratitude and heartfelt respect to my parents, Pinaki Bose and Debolina Bose, my grandparents and my relatives and friends for their boundless love and motivation. Without their affection and moral support, I would not have come up to this stage.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Abstract

THESIS

MACHINE LEARNING MODELS TO AUTOMATE RADIOTHERAPY STRUCTURE NAME STANDARDIZATION

By Priyankar Bose

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

Virginia Commonwealth University, 2023.

Director: Thesis Preetam Ghosh,
Interim Chair and Professor, Department of Computer Science

Structure name standardization is a critical problem in Radiotherapy planning systems to correctly identify the various Organs-at-Risk, Planning Target Volumes and 'Other' organs for monitoring present and future medications. Physicians often label anatomical structure sets in Digital Imaging and Communications in Medicine (DICOM) images with nonstandard random names. Hence, the standardization of these names for the Organs at Risk (OARs), Planning Target Volumes (PTVs), and 'Other' organs is a vital problem. Prior works considered traditional machine learning approaches on structure sets with moderate success. We compare both traditional methods and deep neural network-based approaches on the multimodal vision-language prostate cancer patient data, compiled from the radiotherapy centers of the US Veterans Health Administration (VHA) and Virginia Commonwealth University (VCU) for structure name standardization. These de-identified data comprise 16,290 prostate structures. Our method integrates the multimodal textual and imaging data with Convolutional Neural Network (CNN)-based deep learning approaches such as CNN, Visual Geometry Group (VGG) network, and Residual Network (ResNet) and shows improved results in prostate radiotherapy structure name standardization. Our proposed deep neural network-based approach on the multimodal vision-language prostate cancer patient data provides state-of-the-art results for structure name standardization. Evaluation with macro-averaged F1 score shows that our CNN model with single-modal textual data usually performs better than previous studies. We also experimented with various combinations of multimodal data (masked images, masked dose) besides textual data. The models perform well on textual data alone, while the addition of imaging data shows that deep neural networks achieve better performance using information present in other modalities. Our pipeline can successfully standardize the Organs-at-Risk and the Planning Target Volumes, which are of utmost interest to the clinicians and simultaneously, performs very well on the 'Other' organs. We performed comprehensive experiments by varying input data modalities to show that using masked images and masked dose data with text outperforms the combination of other input modalities. We also undersampled the majority class, i.e., the 'Other' class, at different degrees and conducted extensive experiments to demonstrate that a small amount of majority class undersampling is essential for superior performance. Overall, our proposed integrated, deep neural network-based architecture for prostate structure name standardization can solve several challenges associated with multimodal data. The VGG network on the masked image-dose data combined with CNNs on the text data performs the best and presents the state-of-the-art in this domain.

CHAPTER 1

INTRODUCTION

1.1 Background

Radiation therapy (RT) is an effective cancer treatment therapy where high-intensity radiation beams are used to kill cancerous tissues and cells, decreasing the size of the malignant tumor. In the RT treatment workflow, radiation oncologists use the images based on a Computed Tomography (CT) or Magnetic Resonance (MR) dataset saved in the Digital Imaging and Communications in Medicine (DICOM) files to delineate or contour the various anatomical regions or structures of the organ of interest in these imaging datasets and provide appropriate structure names. These physician-identified structures are either Organs at Risk (OARs), Planning Target Volume (PTV), Clinical Target Volume (CTV), Gross Tumor Volume (GTV), or 'Other' (all the remaining structures). Based on the particular disease site such as prostate or lung cancer, the radiation oncologist contours all neighboring OARs such as bladder, rectum, bowel, femurs, etc., for prostate cases and heart, spinal cord, both lungs, ribs, etc., for the lung cases. While defining these contours and naming them, we observe a high level of variability in the recorded structure names, which makes it hard to consistently gather data for the same structure contour type across a large population of patients. Inconsistencies in the physician-given structure names are primarily due to the personal choice of the physicians coupled with the variation in policies and systems at different RT clinics.

This issue of disparity between the physician-given structure names is addressed by the American Association of Physicists in Medicine (AAPM), and the American

Society for Radiation Oncology (ASTRO)[1, 2, 3]. It mainly addressed the key challenges in the Radiation structure name standardization process and has released a Task Group 263 (TG-263) report where the standard names for the structures are mentioned. With the availability of the standard structure names, there rises the need to automate the standardization of the structure names. It takes huge amounts of time and labour to manually standardize the structure names which presents a challenge in the clinical world that requires rapid decision-making depending upon the criticality of cancer patients. Hence, automatic prediction of standard structure names is a vital problem to solve both from a clinician's and an informatician's point of view. However, there have been limited attempts to automate the structure name standardization process using artificial intelligence (AI) and machine learning (ML) related techniques. Extensive experimentation with various data models and networks is required to elevate the current state-of-the-art in this domain.

From a clinical perspective, our framework has the potential to enable the construction of data pooling tools that can reuse retrospective patient imaging and contouring datasets for tracking patient outcomes, building data registries and clinical trials. It can be integrated with the integrated data abstraction, aggregation, storage, curation and analytics software,the Health Information Gateway and Exchange (HINGE) [4]. Standardized structure names help ensure that all members of the radiation oncology team, including physicians, dosimetrists, and therapists, are using consistent and accurate terminology when identifying and contouring anatomical structures. Furthermore, consistent and accurate contouring of anatomical structures is critical for achieving optimal treatment outcomes in radiation oncology. Standardized structure names can help ensure that all team members are working on the same page, which can help improve treatment accuracy and efficacy.

Multimodal learning can be considered as a branch of deep learning where the learning process is on multiple modalities of data such as visual, language, auditory, etc. Multimodal data can be both homogeneous and heterogeneous. Homogeneous multimodal data comprises multiple similar modalities like CT-image and X-Ray image. Heterogeneous multimodal data consists of different modalities like CT-image and language/text data. Some multimodal data can combine both homogeneous and heterogeneous datasets where multiple similar and dissimilar modalities are present, like CT-image, X-Ray image, and text. Although multimodal learning on multimodal data can improve performance, combining datasets with varying levels of noise and conflicts for learning tasks is more challenging on heterogeneous multimodal data than the homogeneous counterpart. However, heterogeneous multimodal data is more commonplace. For example, images are often accompanied by tags or text explanations and vice-versa. The most common way of handling heterogeneous multimodal data is to combine the features extracted from the different inputs by concatenating or fusing them for solving classification, segmentation, or question-answering problems with multimodal learning. The underlying multimodal learning methods initially used CNN-based or a combination of CNN-based and other networks. Later, some architectures replaced the CNN in their model with advanced and efficient networks such as Region-Based Convolutional Neural Network (R-CNN), ResNet, etc. Currently, transformer-based and graph-based models are very popular in this domain. With more data, vision-language deep learning models have received a lot of attention in the recent past. They have a range of potential applications such as, multimodal reasoning, question answering, visual generation, or captioning (discussed in Section 2). These tasks can be broadly viewed as image/text generation/retrieval or classification problems. Stuart et al.[5] presented a multimodal classification model using images and texts by simple concatenation of the image, description, and title features.

1.2 Related Works

AI/ML is a popular topic in many clinical and biomedical processes, including Radiation Oncology [6, 7, 8]. Natural Language Processing (NLP)-based tasks on clinical and biomedical texts have also gained immense popularity [9, 10, 11, 12]. ML models can be used for automation, value prediction, classification, or other tasks in radiation oncology. A few prior works in standardizing the structure names for organs such as prostate, lung, and head and neck have proposed automated ML models. It has been shown that the standardization of structure names can be done reliably by using neural networks on the head and neck imaging data [13, 14]. This model reported good results but only considered a limited number of OARs for prediction; they also did not consider the non-OARs. Hence, this method is unsuitable for real-world clinical datasets as non-OAR structures usually form most of the structure-naming datasets.

Handcrafted 1D features with reduced dimensions were extracted from the imaging data with singular value decomposition (SVD) [15] based on the bony, non-bony, and combined anatomy. These 1D features were used to build a model that identifies the TG-263 labels [16] where automated ML methods were proposed. The methods were evaluated with weighted F1-Score (best performances of 87.38% on VHA and 90.10% on VCU non-curated datasets) which skewed the performance towards the majority class. Furthermore, an ML model was built based on the textual physician-given structure name data by using a supervised FastText algorithm to create a disease-dependent structure name standardization model [17]. However, both the handcrafted imaging features and the text data were considered together for the first time with traditional ML-based algorithms where two different integration techniques were discussed [18]; it showed decent performance (macro-averaged F1-Score of 87.9%

on VHA data and F1-Score of 75.4% on VCU data with intermediate integration). All these approaches have only applied traditional ML algorithms as they used handcrafted 1D vectors with reduced dimensions from the geometric data. However, it is important to train the whole 3D vision-dose dataset with a learning model for minimum information loss. Due to the huge dimensionality of the vision-dose data, traditional methods face challenges. This motivates the use of deep learning (DL)-based models for automation on the 3D image and dose data with/without text. Furthermore, combining both the VHA and VCU datasets provides additional avenues for performance enhancement which has not yet been explored. DL models are gradient-based computational methods with many processing layers to learn data representation with multiple levels of abstraction [19]. Hence, DL algorithms have the potential to serve as better learning algorithms than standard ML algorithms for the structure name standardization problem. DL methods on this dataset were first proposed by Bose et al. [20] and Sleeman et al. [21] earlier in 2021. Sleeman et al.(2021) [21] proposed a DL-based approach in this context while considering the multimodal geometric data and the radiation dose data where both the data types are numbers. Bose et al. [20] proposed a CNN architecture on the text data and handcrafted geometric features showing improved performance over the previous ML-based network on geometric and FastText-based textual features. ChemProps [22] was introduced in 2021 for composite polymer name standardization. Delineated organs at risk and target Standardization was done for prostate[23] and pelvic[24] cancer patients in 2021 and 2023, respectively. Furthermore, organs at risk delineation and standardization in radiotherapy were studied in 2021 [25] and 2023 [26, 27]. However, for radiotherapy breast structure standardization [27], accuracy was mostly used to evaluate the performances which is biased towards the majority class.

1.3 Summary

To summarize, radiation oncologists use CT or MR images in DICOM files to contour the different anatomical regions in prostate cancer related dataset which are either OARs, PTV, CTV, GTV or other structures. The random physician-provided structure names vary widely due to their personal choices; this makes it very hard to consistently gather data for the same structures across various patients. AAPM and ASTRO addressed this key challenge of disparity across structure names by releasing a TG-263 report which mentions the standard names of the structures. A few prior works in automated standardization of the structure names for organs such as prostate, lung, and head and neck have proposed AI/ML based models. Handcrafted 1D features with reduced dimensions were extracted from the imaging data with SVD based on the bony, non-bony, and combined anatomy and traditional ML methods were used on that data to predict the TG-263 labels. Furthermore, ML models were built based on the textual physician-given structure name data by using a supervised FastText algorithm to create a disease-dependent structure name standardization model and both the handcrafted imaging features and the text data were considered together for the first time alongside traditional ML-based algorithms. As these approaches have only applied traditional ML algorithms with reduced 1D vectors from the geometric data, the obvious next step is to train the whole 3D vision-dose dataset with a learning model for minimum information loss. This motivates the use of DL- or heterogeneous multimodal learning-based models for automation of structure name standardization on the 3D image and dose data with/without text.

CHAPTER 2

VISION-LANGUAGE TASKS AND METHODS WITH EVALUATION METRICS

The availability of vast amounts of curated structured and unstructured data over the past decade has led to the popularity of machine learning, and deep learning based approaches for solving various data-driven problems. With time, the difficulty of these problems has increased, which led to the curation of more annotated datasets. As a result, multimodal data has become the norm nowadays as learning from multiple data modalities has various advantages over any single modality with lesser information. Tasks involving vision-language multimodal data have gone up in large numbers over the last few years. Since image data is numeric, it is considered structured and straightforward to use. On the other hand, textual data is linguistic, and this type of unstructured data needs preprocessing before usage; such preprocessing involves the conversion of the textual data into numeric vectors. Several traditional machine learning approaches have been introduced for predictive modeling on both structured and unstructured datasets. However, the difficulty of the prediction problems and the size of datasets grew with time, leading to the emergence of deep neural network (DNN) based architectures as more promising alternatives.

Multimodal datasets can be paired where samples from each modality are dependent on each other and observations refer to the same event[28]. In the real world, it is challenging to record paired multimodal data; hence, unpaired multimodal data is generally prepared by designing a variety of algorithms where each modality is recorded independent of each other. Apart from that, like other datasets, multimodal data can be noisy or some samples can get corrupted during data retrieval. Existing datasets for image-text analysis specifically in the medical domain are not large enough and this creates a strong requirement for large scale datasets[29]. Multimodal data curation is costly both in terms of labor and time needed for annotation. Hence, the community often uses existing standard datasets which is more practical and convenient to curtail labor and time expenses[30]. Researchers often use standard evaluation metric to compare their methods on these popular datasets. Also, proper dataset selection can be challenging as at times, they may lack proper description from the domain specialists. Hence, comprehensive description of datasets can be significantly helpful[30] for future research on multimodal data. Also in many cases, realtime datasets can be imbalanced. The class imbalance with multimodal data significantly lower the performance of DNN models and this challenge can be dealt with either using data-level methods like sampling and augmentation or algorithmic-level methods like the introduction of different class weights to the classifier[31]. Some standard multimodal vision-language datasets are MS-COCO [32], COCO-QA [33], RefCOCO/RefCOCO+ [34], DAQUAR-ALL and DAQUAR-REDUCED [35], Visual Genome [36], Visual7W [37], Getty Image [38], Natural Language for Visual Reasoning (NLVR^2), amongst others. Considering image and text-based multimodal data, deep learning based techniques have been widely adopted to build attention-based networks to learn the joint cross-modal semantic relations. Most of these attention-based networks are either built with CNNs or with both CNNs and recurrent neural networks (RNNs) [39]. Recently, graph neural networks (GNNs) and transformers have become extremely popular for attention-based approaches on image-text multimodal tasks [39]. The attention-based mechanisms built with the above mentioned deep learning networks present an interesting range of methodologies that exhibit improved model performance.

2.1 Vision-language Multimodal Tasks

Next, we revisit some popular machine learning tasks for multimodal data analysis. These methods use information extracted from text to tackle vision-related tasks or vice-versa. At a higher level, this is achieved by mapping visual and textual data to the same latent space to leverage the overlapping information of both data modalities to improve the model performance. Some of the popular tasks that use image and text data jointly are listed as follows:

2.1.1 Vision-language Multimodal Classification

These tasks aim to classify or separate various classes in the dataset. Classification can be both Bi-class or Multiclass. Vision-language multimodal classification has been performed in the past for emotion classification in Twitter data [40] or hate-speech classification [41].

2.1.2 Vision-language Multimodal Visual Segmentation

Segmentation refers to the process of separating out objects or bounding boxes in an image. Vision-language multimodal segmentation [42] refers to image segmentation based on queries or referring expressions. Several attention-based architectures have been built for image-text visual segmentation on datasets like ReferIt [43], UNC[34], UNC+[34], G-Ref[44], etc.

2.1.3 Vision-language Multimodal Question Answering

These tasks provide an answer to a question based on a vision-language input [45]. For example, answering a question ”What is the color of her eyes?” from the image of a girl falls under question-answering tasks. Several attention mechanisms have been developed on image-text data like VQA[45], VQA 2.0[46], VisDial[47], VCR[48], Med-VQA[49] dataset, etc, for vision-language multimodal question answering.

2.1.4 Vision-language Multimodal Visual Captioning

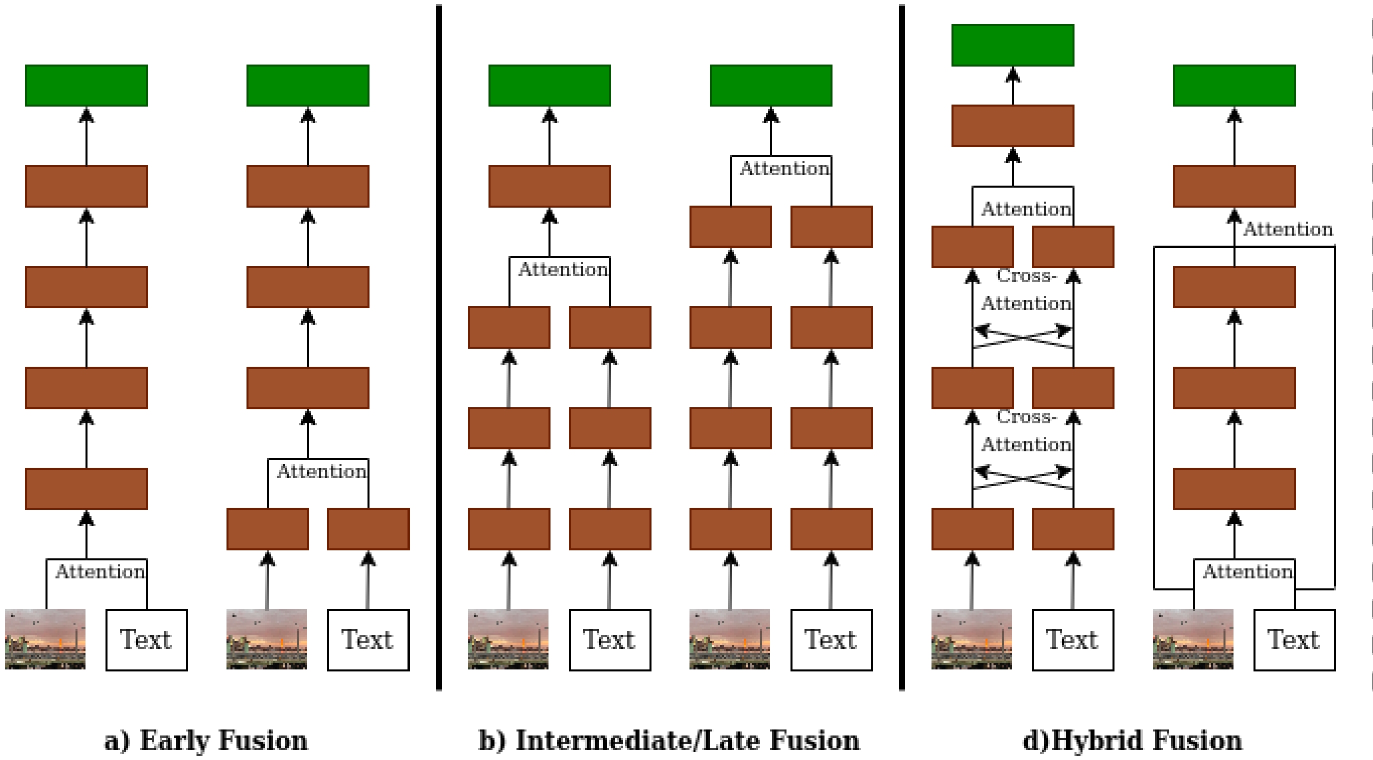
These tasks are responsible for writing descriptions of images using text. DimBERT [50] is a transformer model that has been tested on the generation tasks of Flickr30K[51] and MS-COCO[32] datasets.

2.1.5 Vision-language Multimodal Image-Text Retrieval

These tasks are responsible for retrieving images or texts based on multimodal image-text data [52] and attention mechanisms have been developed previously on Kitchen[53], Flickr30k[51], MS-COCO[32], RefCOCO/RefCOCO+[34], etc. datasets.

2.2 Vision-language Multimodal Methods

The objective of multimodal learning is to learn the semantic alignment between the sub-parts of instances across different data modalities. For example, in the case of vision-language modalities, it is vital to learn the mapping between the image patches and their corresponding textual sub-parts. Semantic alignment learning has already proven to be successful by allowing the sub-parts across different modalities to densely associate. As a result, a particular sub-part of one modality is aligned with all the sub-parts from another modality, where most of the sub-parts are trivial and hence, are irrelevant sub-elements [54]. The irrelevant sub-parts need to be attended to eliminate unnecessary information that distracts the semantic alignment. In this context, attention mechanisms are very important in multimodal learning as they can effectively align one modality's sub-parts with the other's relevant sub-parts.



a) Early Fusion b) Intermediate/Late Fusion d)Hybrid Fusion

Fig. 1.: Comparison of the first three attention-based Fusion methods on images and texts as inputs.

2.2.1 Simple Fusion Models

Multimodal fusion, which involves a combination (addition, averaging, weighted voting, etc.) of information from different data sources like text and images, has emerged as a very popular method for various analytical tasks. Over the past few years, the computational semantics community has popularized the use of extended distributional semantics with vision-language multimodal data to integrate features extracted from pictures and language. This resulted in the design of flexible and general frameworks to integrate the visual and textual features that involves mapping the visual representations of the word-level semantics for improved learning performance.

However, because images and documents represent distinct modalities, multimodal fusion requires the conversion of visual features into a discrete space, like word units. Images are commonly represented by a continuous feature space like shape, color, etc., whereas words in a sentence/phrase are discrete. Multimodal fusion can be done at any phase of the model: the features can be fused at the beginning of the model, or the outputs of each modality can be fused by following some statistical equations for the final prediction. The different fusion techniques are summarized below[55] and the first three attention-based fusion strategies are illustrated in Fig.1 [56]:

- **Early Fusion**: This method fuses data or features from different modalities to form a single feature space. Many segmentation architectures can be adapted for the early fusion strategy. Moreover, cross-modal interactions throughout the encoding stage fall under the early fusion class; this includes feature-level fusion, which considers such cross-modal interactions during encoding. Such fusion of data from multiple modalities right at the beginning is straightforward and meaningful in the case of homogeneous modalities. However, simple concatenation of feature vectors from different modalities is not the most meaningful way of representing the model input in most cases; with high dimensional data, this adversely affects the model's performance [18, 57].

- **Intermediate/Late Fusion**: This technique combines the data sources to form a common or reduced feature space. Here, high-dimensional features are transformed into lower-dimensional features by passing the input features through some networks. The reduced feature vectors can be combined to form the feature vectors that represent the model input. Also, the reduced feature vectors can be used as separate model inputs, and a combination of features can be done in the middle stages of the model. Various combinations of integration techniques are possible in the case of intermediate integration. One such combination is passing each modality's inputs through varying networks, either in parallel or in series, and ultimately fusing them. Another intermediate integration technique can be the fusion of features in the intermediate step and then reconstructing the features from the fused features, like in the case of an encoder-decoder network. Late fusion methods integrate multimodal feature maps at the decision level.

- **Hybrid Fusion**: Hybrid fusion methods can leverage the advantage of both early and late fusion strategies. Generally, the segmentation network accesses the data through the corresponding branches. These models have a generative component which learns the features from the low-level inputs in separate branches where each branch individually gathers the information from each modality and processes them forward in the branch. These models have a discriminative component where the features from individual branches are processed together for higher-level reasoning. Therefore the hybrid fusion networks are capable of adaptively generating a joint feature representation over multiple modalities, resulting in an improved performance in terms of accuracy and robustness [55].

- **Statistical Fusion**: An alternative fusion approach, called statistical fusion, has rarely been used for fusing multimodal data. These models are easier to use in the case of homogeneous multimodal data but are tricky for heterogeneous multimodal data. To reduce the model uncertainties during decision-making, Blum et al. [58] introduced statistical fusion methods for deep-learning based segmentation. Methods like Bayes categorical fusion and Dirichlet fusion were performed on data across two homogeneous modalities. They also provide a framework to fuse heterogeneous multimodal data with deep learning.

2.2.2 Simple Stacked Models

A stacked model involves the concatenation of information from multiple modalities at either the data level or the feature level. The initial attempt of stacking was to concatenate the raw data from different modalities into multiple channels. In the Stacked model, vectors from the different modalities are simply stacked on top of each other at any stage of the model. As with fusion models, the goal of stacked models is also a refined performance capable of mapping the visual representation of the word semantics. Stacked models present a flexible framework that requires converting visual features into discrete spaces like word units. Learning the visual representations from natural language supervision [59] became popular after 2016 and slowly replaced the stacked models. Stacking of vision and language features can be done in any stage of the model architecture, e.g., early stacking, intermediate/late stacking, or hybrid stacking, thereby bearing much resemblance with simple fusion models.

2.2.3 Complex Hybrid Dynamic Models

Dynamic models are complex models made up of multiple modules that help in dynamic learning with attention. Dynamic Memory Network (DMN) was first introduced with a Gated Recurrent Network (GRU) base in 2016 by Ankit et al.[60] in tackling question-answering problems. In this network, the vector representation of the question triggers an iterative attention mechanism that retrieves relevant facts by searching the input vectors. Then, the memory module in DMN reasons over retrieved facts and generates the answer by an answer module that comprises the vector representation of all relevant information. Dynamic models with attention are often bidirectional, i.e., they are capable of cross-modal information retrieval. These models learn the information dynamically through the presence of deep dynamic layers. Some of the multimodal dynamic models include a combination of stacking and fusion operations, and some take cognizance of inter-modality and intra-modality relations. Since 2015-16, when the community started taking an interest in attention-based vision-language multimodal problems, several dynamic model architectures have been proposed.

2.2.4 Transformer Models

Information from image-text multimodal data can also be learned attentively by using an encoder-decoder based Transformer Model. Many methods have been tried in the past that have used attention-based transformer networks to perform some tasks with vision-language multimodal data. Like dynamic models, most multimodal transformer models are usually bidirectional as transformer networks usually contain joint information representation modules, thereby helping in cross-modal information retrieval. In addition, recurrent networks like LSTMs form the building blocks of a transformer.

Language models have shown improved performance in many NLP tasks using contextual information to represent the features. It is also a supervised learning model, as each instance's inputs are well-defined. The language models popular in NLP tasks are ELMO, ULMFit, BERT, etc. Out of these, Bidirectional Encoder Representations from Transformers or BERT [61], introduced by Google in 2019, has become extremely popular for various NLP tasks. Its breakthrough has improved performance in many NLP tasks because of its strong ability to pre-train deep bidirectional representations of any unlabelled text by conditioning on its context on both sides in all the 12 transformer layers. Pre-trained representations are increasingly becoming crucial for multiple NLP and perception tasks. Likewise, for solving vision-language multimodal problems, several variants of BERT and other transformer models have been introduced that can attend to the data across two modalities. Transformer model-based networks are the most popular for solving multimodal vision-language problems with attention following the inception of BERT in 2019. Recently, the transformer-based cross-modal interactions have been classified into six categories [62]: a) early summation, b) early concatenation, c) hierarchical attention (multi-stream to one-stream), d) hierarchical attention (one-stream to multi-stream), e) cross-attention, f) cross-attention to concatenation, out of which either the hierarchical attention models or the cross-attention models are of great significance to us. As discussed below, transformer-based models have recently gained immense popularity in vision-language multimodal data modeling and span multiple application domains.

2.2.5 Graph based Models

A Graph neural network (GNN) is an artificial neural network applied to data in graph format [63]. The pairwise message passing enables graph nodes to update their representations through information exchange with their neighbors iteratively. The various tasks that have been solved over time with GNNs involve 1) Graph-level tasks, 2) Node-level tasks, and 3) Edge-level tasks. The graph-level tasks are similar to image classification or text sentiment analysis. In contrast, the node-level tasks are analogous to image segmentation or predicting the parts of speech of each word in a sentence. The edge-level tasks are similar to predicting the connections between objects in the case of images and relationships between entities when textual data is used. In multiple areas of science and engineering, such as pattern recognition, computer vision, NLP, molecular biology, etc., the bonding among data can be represented by a graph. GNNs have earlier been used with images and text separately or with image-text multimodal data. The first graph-based attention mechanism on image-text data emerged around 2019 and became extremely popular. The different graph based fusion methods for heterogeneous multimodal data came up in 2020 [64], but they were not restricted to images and texts only. They have provided a general framework for graph based approaches with attention on image-text multimodal data and were applied to the following domains.

2.3 Common Evaluation Metrics

We next outline the popular evaluation metrics in the context of vision-language multimodal tasks. F1-Score is a popular evaluation metric for classification tasks in general. Comparisons can be classified as exact or relaxed matches [65]. Relaxed match only considers the correct type and ignores the boundaries as long as there is an overlap with ground truth boundaries. In the case of an exact match, it is expected that the entity identified correctly should also detect boundary and type correctly at the same time [65]. Precision is the fraction of relevant points among all retrieved points. Recall is the fraction of retrieved points among all relevant points. In case of segmentation tasks, Intersection over Union (IoU) and Dice Similarity Coefficient (DSC) are very popular evaluation metrics that are both straightforward and effective. IoU is measured as the area of overlap between the predicted and ground truth values divided by the area of union between them. DSC is calculated as twice the area of overlap divided by the total number of pixels in both images. The following keys are used to calculate the F1-Score, IOU, DSC, precision, and recall.

- True Positive (TP): A perfect match between the entity obtained by the system and the ground truth when the ground truth is a positive class.

- False Positive (FP): A mismatch between the entity detected by the system and the ground truth when the ground truth is not a positive class.

- False Negative (FN): A mismatch between the entity detected by the system and the ground truth when the ground truth is not a negative class.
- True Negative (TN): A perfect match between the entity obtained by the system and the ground truth when the ground truth is a negative class.

They are calculated as follows:

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

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$DSC = \frac{2 * TP}{2 * TP + FP + FN}$$

$$IoU = \frac{TP + FP + TN}{TP + FP + FN}$$

2.4 Summary

Tasks involving vision-language multimodal data have gone up in large numbers over the last few years. Vision data is numeric and straightforward to use whereas textual data is linguistic and needs conversion into numeric vectors before use. We have discussed the various multimodal vision-text tasks, methods along with the relevant evaluation metrics in this chapter. The different vision-language tasks involve vision-language multimodal classification, vision-language multimodal visual segmentation, vision-language multimodal question-answering, vision-language multimodal visual captioning, vision-language multimodal image-text retrieval, etc. while the various multimodal methods are simple fusion models, simple stacked models, complex hybrid dynamic models, transformer models and graph based models. The popular evaluation metric for classification tasks in this domain is the F1-Score where as IoU and DSC are more popular for segmentation tasks.

CHAPTER 3

PROSTATE STRUCTURE NAMING DATASET AND PRE-PROCESSING

The information can come from various input channels; for example, images are made up of tags and captions, videos are associated with visual and audio signals, and so on [66]. Modality of data is often used in data science to refer to the measurement method used to obtain the data. Each modality, whether independent of other modalities or dependent on other modalities, has unique information that, when added together, may improve model performance. Although combining complementary data from multiple modalities may improve the performance of learning-based approaches, they are accompanied by practical challenges of fully leveraging the various modalities such as noise, conflicts between modalities, etc. [67]. A multimodal dataset contains data of different modalities where the data-types are similar across different modalities, i.e., homogeneous multimodal data or the data-types can vary across different modalities, i.e., heterogeneous multimodal data. For example, multimodal neuroimaging data consisting of magnetic resonance imaging (MRI) data and positron emission tomography (PET) data used for effective Alzheimer's disease diagnosis [68] is a multimodal data system where the data types of the various modalities are the same as images are either 2-D/3-D arrays of numbers. On the other hand, images and text were both considered for learning with multimodal data where the data types vary; images are number arrays, whereas text is strings [66, 18]. Multimodal learning has been addressed many times across various domains [69, 70, 71] including the clinical and the biomedical field [72, 73, 74]. However, in most of these cases, the data types are similar across different modalities. The additional modalities were either originally present in the dataset or have been synthesized from a particular modality. In such cases, the conflicts between modalities is likely to be less unlike multimodal data with varying data-types where the data for various modalities are physically collected. Our multimodal vision-text dataset for prostate structure naming and pre-processing techniques are discussed in the following sections.

3.1 Prostate Structure Naming Dataset

The textual physician-given structure names [17] and the 3D DICOM CT geometric data [16] are considered in our heterogeneous multimodal dataset. The DICOM CT image data show that physicians identify the anatomical structures of interest that should be irradiated or avoided during treatment. The physicians then use a Treatment Planning System (TPS) to delineate the border around these structures. This process was implemented for all the relevant imaging slices, producing several enclosed polygons for each structure. These structure data for a particular patient were stored in DICOM format. The clinical dataset used here was collected from 759 prostate cancer patients by the VHA RT centers and VCU radiation oncology department. The count of the various organ structures for prostate cancer patients is shown in Table 1. Out of the 759 prostate cancer patients, the total count of the physician-given structure names was 16,290, and the standard structure names consisted of 6 OARs (Femur_L, Femur_R, Bowel_Large, Bowel_Small, Bladder, Rectum), Target (PTV) and 'Other' (all the remaining) prostate structures. From the original dataset, there was some data loss with time due to corruption, a shift in technologies and platforms, etc. Finally, the dataset consists of 9723 samples or structures, out of which 7803 samples were used for training and 1920 samples for testing.

- *VHA Dataset:* There are 40 RT centers under VHA that are spread nationwide. Hence, there is a need to evaluate the quality of treatments across these centers. To ensure this, VHA had implemented a clinical informatics initiative called the Radiation Oncology Quality Surveillance Program (VA-ROQS) [75]. The maximum number of prostate cancer patients considered for each center was 20. The patients were selected per the criteria mentioned in Hagan et al. [75] that ultimately helped store the data of 709 prostate cancer patients for analysis. Next, the physicians manually labeled the organ structures using TG-263 nomenclature for building the models.

- *VCU Dataset:* A dataset was prepared from the DICOM CT geometric data from a random cohort of 50 prostate cancer patients from the Radiation Oncology department at VCU. The physicians manually labeled the structures, similar to the VHA dataset.

Standard Names	**VHA Physician Given Name Counts**	**VCU Physician Given Name Counts**	**Total Physician Given Name Counts**	**Available Given Name Counts**
Bladder	609	50	659	519
Rectum	719	50	769	517
PTV (Target)	714	38	752	522
Femur_L	694	29	723	508
Femur_R	700	29	729	515
SmallBowel	250	49	299	145
LargeBowel	341	0	341	234
‘Other’	11,038	980	12,018	6763
Prostate Total	**15,065**	**1225**	**16,290**	**9723**

Table 1.: Distribution of the Organ structures for the Prostate Cancer Patients.

3.2 Data Pre-Processing

In this case, the multimodal data consisted of numeric vision-dose data and the randomized physician-given textual structure names. The textual names, being unstructured, have to be first converted into numbers before they are fed into the learning framework. On the other hand, vision-dose data are 3D arrays of numbers in each case. Due to the varying nature of the input data modalities, both the modalities require different types of pre-processing. We next discuss the data pre-processing techniques for the different data modalities.

3.2.1 Textual Data

The textual features are the physician-given structure names. The maximum length of the given names and the characters used in them depends upon the system used by the particular vendor. The distribution of the physician-given structure names of three random prostates is shown in Table 2. Notably, in the case of the 'Other' structures, a wide variation in the given names is observed for prostate cancer patients. Furthermore, some physicians annotate some 'Other' type structures as PTV, but the term 'PTV' is always associated with the Target type structures. This is a challenge for any ML algorithm to predict whether the term 'PTV' falls under the Target class or 'Other' class. The inconsistencies in the physician-given names of the structures are shown in Tables 2. Although a wide variation can be noticed in the structure naming procedures in general, the overall character set is limited. Since the 'Other' class consists of all the contoured structures except the OARs and the Target, it is the highest occurring structure. A high level of data imbalance was also observed between the 'Other' class and all the remaining classes.

Text preprocessing techniques need to be wisely chosen so that important details

Table 2.: Distribution of the Physician Given Structure Names for the Prostate Cancer Patients.

Structure Type	**Standard Name**	**Patient 1**	**Patient 2**	**Patient 3**
OAR	LargeBowel	Colon_Sigmoid	-	-
OAR	Femur_R	Femur_Head_R	RtFemHead	Hip Right
OAR	Femur_L	Femur_Head_L	LtFemHead	Hip Left
OAR	Bladder	Bladder	bladder	Bladder
OAR	Rectum	Rectum	rectum	Rectum
OAR	SmallBowel	-	bowel	-
Target	PTV	PTV_7920	PTV45Gy	PTV 2
‘Other’	“Other”	z post rectum	ptv4cm	Rectum - PTV
‘Other’	“Other”	Body	nodalCTVfinal	Prostate + SV
‘Other’	“Other”	CTVp	NONPTVBlad	PTV 1
‘Other’	“Other”	CouchInterior	CTVProsSV	Bladder - PTV
‘Other’	“Other”	PenileBulb	External	Seminal Vesicles
‘Other’	“Other”	Prostate	FinalISO	Seed Marker 1
‘Other’	“Other”	z_rectuminptv	MarkedISO	Dose 104[%]
‘Other’	“Other’	z_dosedec	CTVBst	Seed Marker 3

are not missed, which may result in poor model performance. To avoid this, we restricted ourselves to minimal text-based preprocessing, which consisted of lowercasing all the alphabets.

It is vital to choose the precise tokenization algorithm so that most of the terms are present in the vocabulary of the tokenizer. Our dataset contains clinical data; hence, selecting a medical domain tokenizer is very relevant. Here, we have used a recent tokenizer that is strong in the biological domain, BioBERT [76]. This tokenizer breaks up a single word into multiple tokens. For example, after preprocessing, the BioBERT tokenizer tokenizes the physician given names *’nodalCTVfinal’* and *‘z_dosedec* into *‘nod’, ‘##al’, ‘##CT’, ‘##V’, ’##final’* and *‘z’, ‘_’, ‘dose’, ‘##de’, ‘##c’*, respectively.

After tokenization, the next goal is to produce the feature vectors from the text. We have followed two ways of generating the feature vectors: (a) based on our corpus and (b) using the pre-trained word embeddings. The tokens were converted into the token-ids, and the feature vector was generated based on our corpus. Corpus-based embedding creates the embedding vector based on the count of a particular word in our corpus and we then train the weights to the embedding vector. In our case, corpus-based embedding performed better than some of the pre-trained word embeddings. We varied the embedding dimensions keeping the network unchanged and the embedding dimension of 256 produced the best results for prostate cancer patients and that layer was fed to a 1D CNN with 256 filters. BioBERT-based pre-trained word embeddings are also generated. Both these embeddings represent contextualized word embeddings that train a BERT [77] based model over a biomedical and clinical corpus. In our case, the feature vectors based on our corpus provided excellent results compared to the pre-trained word embeddings, as the physician-given structure names are not context-dependent. These word embeddings are used as input to the deep learning model.

3.2.2 Vision-Dose Data

As part of the treatment planning process, physicians annotate regions of interest in the planning CT image using a manual or semi-automated contouring tool. These annotations are saved in the DICOM-RT structure set format, which includes the name of each contoured structure and the 3D coordinate location of each drawn point.

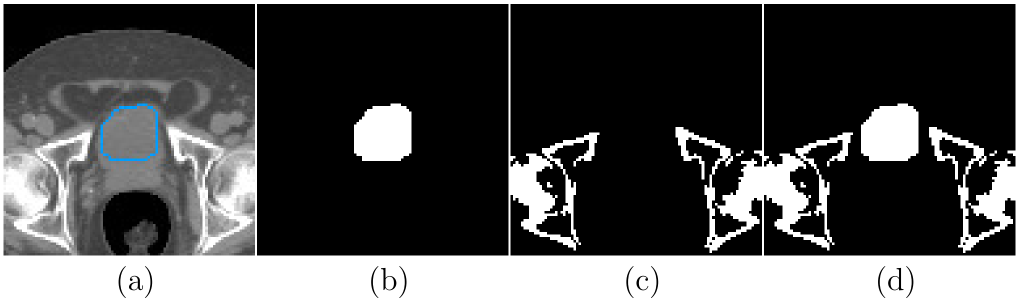


Fig. 2.: Image and structure set data from a single CT slice: (**a**) Delineation of a bladder (in blue) over the corresponding planning CT image (**b**) Bitmap representation of the bladder (**c**) Bony anatomy of the same CT, created with a density-based filter (**d**) Combination of the structure set and bony anatomy data

Custom software used in our prior work by Sleeman et al. [16] who extracted these individual points from the training structure set files and connected them for each corresponding CT image slice to create a number of 2D hollow bitmaps. Each bitmap was then made solid with a flood fill algorithm and then combined to create a single volumetric bitmap for each delineated structure. Each planning image was resized to 96 × 96 × 48 voxels, and the resulting structure bitmaps were interpolated on this same grid. In addition to the structure set data, the corresponding planning CT image was filtered to create a bitmap of the bony anatomy, which was also converted into a feature vector. Figure 2a shows the delineation of a bladder over the corresponding planning CT image, and Figure 2b–d shows the resulting bitmap representations. The dose data are also introduced by recording the dose values in the respective voxels for the particular structure set in the organ that were too interpolated on the same grid as that of the structure sets. The voxels inside PTV reasonably receive the highest amount of dose, whereas OARs and 'Other' structures receive a much smaller amount of dose. Thus, the dose values provide significant information on top of the delineated images for pointing out the structure-classes based on the magnitude of the dose received.

3.3 Summary

The textual physician-given structure names and the 3D DICOM CT geometric data are considered in our heterogeneous multimodal dataset. There are 40 nationwide RT centers under VHA, storing the data for 709 prostate cancer patients that considered a maximum number of 20 prostate cancer patients for each center. The VCU dataset considered a random cohort of 50 prostate cancer patients from the Radiation Oncology department, thereby making the total count of 759 patients in the clinical dataset that comprise 16,290 physician-given structures including 6 OARs (Femur_L, Femur_R, Bowel_Large, Bowel_Small, Bladder, Rectum), Target (PTV) and 'Other' (all the remaining). As a result of some data loss due to corruption, a shift in technologies and platforms, the final dataset consisted of 9723 samples or structures, out of which 7803 samples were used for training and 1920 samples for testing. Text pre-processing involving lowercasing all the alphabets and data tokenization using BioBERT were performed on the textual datasets. For the vision data, each volumetric bitmap for each delineated structure was resized to 96 × 96 × 48 voxels and the resulting structure bitmaps were interpolated on this same grid. In addition to the structure set data, the corresponding planning CT image was filtered to create a bitmap of the bony anatomy, which was also converted into a feature vector. The dose data is also introduced by recording the dose values in the respective voxels for the particular structure set in the organ that were also interpolated on the same grid.

CHAPTER 4

PROSTATE STRUCTURE NAME STANDARDIZATION METHODS AND RESULTS

We used CNN, or CNN-based architectures with intermediate stacking of multimodal features, on the 3D images and doses as CNNs have previously demonstrated an edge over other deep neural networks (DNN) on vision data [78]. We build a naive 1D DNN on the reduced features of the geometric data as shown in Section 4.1 and 3D CNN, Residual Network (ResNet), and Vision Geometry Group (VGG) Network on the masked vision-dose data as show in Section 4.2 to compare their performances. We have restricted ourselves to these networks that we have customized for our purpose. Although some other advanced and computationally heavy models have been proposed in recent years, the community is still carrying out meaningful experiments using the models that we have proposed here. In the future, we would explore such advanced models including DenseNet [79], Squeeze Net [80], ENet [81] besides also some vision transformers [82] and compare their performances. We have used 1D CNN on the pre-processed textual data as CNN performed the best on the texts when compared with the performance of Recurrent Neural Networks (RNNs). In this chapter we discuss the various DL methods on different representation of the prostate structure naming dataset.

4.1 Deep Neural Network Architectures on Reduced Featured Geometric and Text Data

Each planning image consists of 96 x 96 x 48 voxels and the resulting structure bitmaps were interpolated on this same grid. In addition to the structure set data, the corresponding planning CT image was filtered to create a bitmap of the bony anatomy only which was also converted into a feature vector. The structure set and bony anatomy bitmaps were then converted into 1-D vectors, concatenated, and feature reduction was performed using truncated singular value decomposition (SVD) [15] following the previous work[16].

4.1.1 Methods

Earlier, no deep learning models were considered for analyzing the textual features but these algorithms seem to have the potential of providing the state-of-the-art results on this textual dataset. We build a CNN-based deep learning model to train and test the textual data.

4.1.1.1 SVD

Matrix decomposition or matrix factorization is performed in case of high-dimensional data which involves describing a given matrix using its constituent elements. Singular Value Decomposition, or SVD is one of the most common matrix decomposition methods. Data reduction with SVD is more stable than other methods like eigen decomposition and hence, it is widely applied for a variety of applications such as compressing, denoising, and data reduction.

In our case, we fit the 3D vision-dose data consisting of delineated structure sets, dose data and organ images with SVD algorithm and transformed the 3D data with three channels into a singular dimension consisting of 2304 features. By doing this, we significantly reduced the size of the input data which resulted in a considerable speed-up of the training process. This resulted in 99.86% reduction in the number of features that requires significantly less storage space. The multi-view data is given as input into the parallel CNN layers as shown in the following subsections.

4.1.1.2 Deep Neural Network architecture

A CNN is an artificial deep neural network (DNN) originally introduced in [83]. It uses convolution on the input and passes down the result to the next layer. Although multi-layer perceptron neural networks can learn from textual data, they are often outperformed by CNNs [84] which can slide a window on the input data. CNN was first used for sentence classification in 2014 [85]. The hyper-parameters that a CNN uses are convolutional kernels, filters and the number of input channels and output channels. A CNN takes into account the words around an individual word while making predictions and this can be very effective for classification.

We represented the features and embedded them from the textual data and geometric data as mentioned earlier. The size of the window is dependent the number of filters and window-size determines the amount of data of a particular data instance to be processed by the model at once. We use two layers of 1D CNN with 1D max-pooling to feed the textual features after creating the embedding matrix using an embedding layer. The number of filters used in the consecutive 1D CNNs in the case of textual prostate data was 256 and 128, respectively. Since, the geometric data is reduced from the original 3D vision-dose data by SVD, there is not much contextual information. Hence, we use a dense feed-forward linear layer to feed the geometric data with 1152 units, followed by feed-forward layers of 576 and 288 units, respectively. Next, we concatenate the parallel textual and geometric features and pass it through dense layers of 256, 128, 64 and 32 units, respectively. At last, we feed the data through the last dense layer of 8 units for classifying the data into the 8 above mentioned classes. The activation function used in all the layers except the last layer is Rectified Linear Unit (ReLU) whereas Softmax is used in the last layer for multi-class classification. “Categorical cross-entropy” was used as the loss function with the “adam” optimizer. For our model, we used precision, recall, and accuracy as metrics which are macro-averaged which is explained in details in the next Section 4.2.

Since, the representation of the majority ’Other’ class is quite large compared to other classes, it skews the performance of the model towards the majority class. Hence, undersampling the majority class samples is of great significance. We have undersampled the majority class samples at 500, 100, 1500, 2500, 3500 and 4500 samples which are explained in details in the next Section 4.2. Repeated experimentation with undersampling the majority class samples shows the variation in the performance of the DNN architecture and shows that a little amount of undersampling can be instrumental in boosting the network performance.

4.1.2 Results

At first, the majority class was undersampled at samples of 500, 1000, 2500, 3500 and 4500 and the performance of the feed-forward neural network on the 1D geometric data and 1D CNN on the textual data individually. The performances of these architectures with majority class undersampling are recorded in Table 3. The performance of the DNN on the feature reduced geometric data is very poor, achieving macro-averaged F1-Scores around 40%. F1- Scores on the geometric data vary between 38.08% and 42.2% and the majority class undersampling doesn’t result in a significant performance improvement. The information loss in the vision-dose

Table 3.: Model performances for the Prostate cancer patients with variation in the majority class samples on either reduced feature geometric or text data.

Total Samples from 'Other' Class	Data Modality	Method	Precision (in %)	Recall (in %)	F1-Score (in %)
500	Geometric	1D DNN	36.35	53.8	39.99
500	Textual	1D CNN	83.45	97.39	88.87
1000	Geometric	1D DNN	40.28	49.56	39.77
1000	Textual	1D CNN	84.63	97.02	89.74
1500	Geometric	1D DNN	39.67	45.93	42.2
1500	Textual	1D CNN	86.09	97.37	90.55
2500	Geometric	1D DNN	42.66	42.19	41.42
2500	Textual	1D CNN	89.78	95.07	92.23
3500	Geometric	1D DNN	43.95	42.1	42.18
3500	Textual	1D CNN	90.1	94.86	92.34
4500	Geometric	1D DNN	44.0	40.85	41.04
4500	Textual	1D CNN	90.23	94.85	92.42
5432 (No Sampling)	Geometric	1D DNN	43.9	36.7	38.08
5432 (No Sampling)	Textual	1D CNN	91.16	94.75	92.88

data because of feature reduction is so significant that the feed forward network fails to learn the information from the data to a great extent. Therefore, there is a need to develop a model that can take in the full vision-dose data without feature reduction.

On the other hand, it is evident from the data in Table 3 that the 1D CNN on the textual data alone performs very well. The macro-averaged F1-Scores are around 90% for all the cases of majority class undersampling. The F1-Scores vary between 88.87% and 92.88%. The majority class undersampling has little influence in the performance of this model but higher degrees of majority class undersampling results in a slight performance deterioration. It is evident from the results that the information from textual data is of paramount importance for successful radiotherapy structure name standardization. Hence, it is vital to experiment with a network that takes in both the data modality; textual and geometric.

Table 4.: Model performances for the Prostate cancer patients with variation in the majority class samples on reduced feature geometric and textual data.

Total Samples from 'Other' Class	Method	Precision (in %)	Recall (in %)	F1-Score (in %)
500	1D DNN and 1D CNN	77.37	95.66	84.46
1000	1D DNN and 1D CNN	78.7	96.09	85.1
1500	1D DNN and 1D CNN	87.01	97.41	91.37
2500	1D DNN and 1D CNN	87.54	96.99	91.58
3500	1D DNN and 1D CNN	89.56	91.22	90.22
4500	1D DNN and 1D CNN	89.38	95.01	92.01
5432 (No Sampling)	1D DNN and 1D CNN	91.46	93.22	92.23

The performance of the 1D fully connected DNN and 1D CNN network on the geometric and textual data together are reported in Table 4. The performance of this multimodal data exhibits a significant improvement in performance over the unimodal geometric data model but simultaneously, it exhibits a slight drop in performance over the unimodal textual model. Although the macro-averaged score goes up to 92.23% with undersampling the majority class at 4500 samples, it drops to 84.46% when the majority class is undersampled at 500 samples. Similar to the unimodal textual model, the performance drops with higher amount of undersampling. Also, it is important to note that the presence of the feature reduced geometric data causes a slight deterioration in performance over the unimodal textual model. Hence, there is a significant need for a deep network to learn from the 3D vision-dose and textual data. Hence, in the next Section 4.2, we discuss the about architecture and extensive experimentation of our DNN on the vision-dose and textual data.

4.2 Deep Neural Network Architectures on 3D Vision-Dose and Text Data

We masked each planning CT image with their respective bitmap structure representations for our final architecture using simple multiplication. Thus, the 3D integer image arrays were converted to 3D float arrays after masking. Similarly, we also masked the planning doses with their respective bitmap representation to obtain masked dose data. For each structure set, the image and the dose arrays were concatenated into separate channels, thereby stacking the image and dose information in a 3D array with two channels. Masking the images and doses with their structure bitmaps is more memory efficient than using both images and doses along with their respective structure bitmaps. Using all these three modalities together requires more memory space and time to create, store and train the data, with training being particularly computationally expensive. However, the masked data contain all the information present in the three modalities and is significantly less computationally expensive for training and storing. Hence, masking is highly recommended in case of performance improvement or in case of marginal drop in performance. In our case, masking improves the model performances to some extent, as shown in the following sections. The computationally easy step of masking the images and doses with the help of the corresponding structure bitmaps is shown in Figure 3a,b, respectively.

4.2.1 Methods

Our network architectures with 1D CNN on the text and 3D CNN, 3D VGG network, and 3D ResNet on vision-dose are illustrated in Figure 4. Although, the architectures of the three different networks are somewhat similar, it is interesting to visualize the particular sequence of layers in each case. It provides a reference to

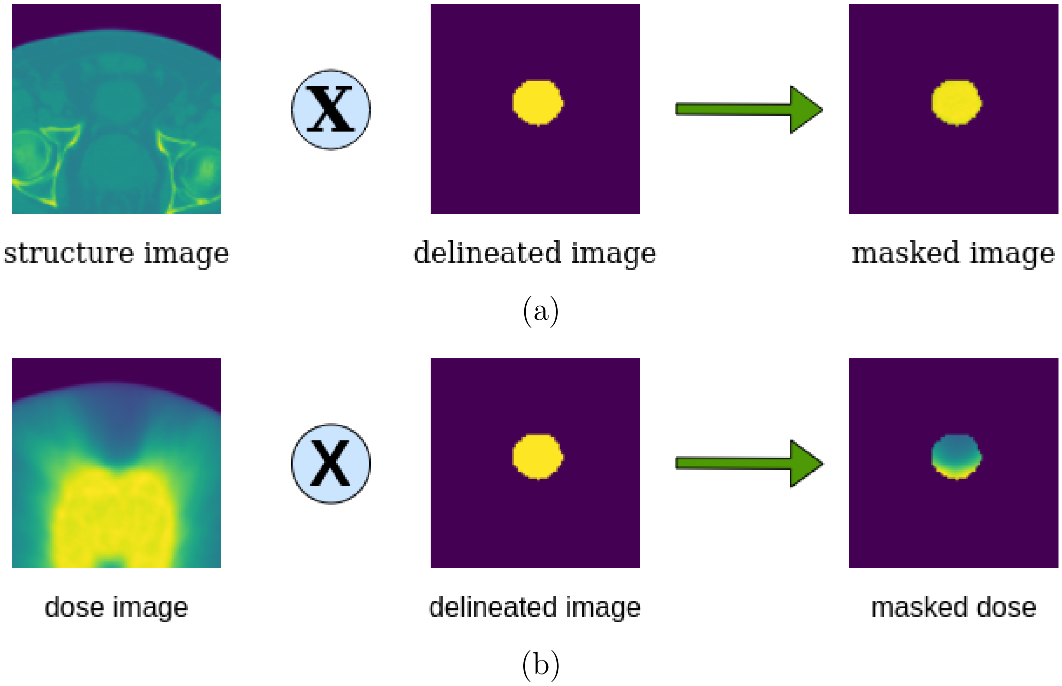
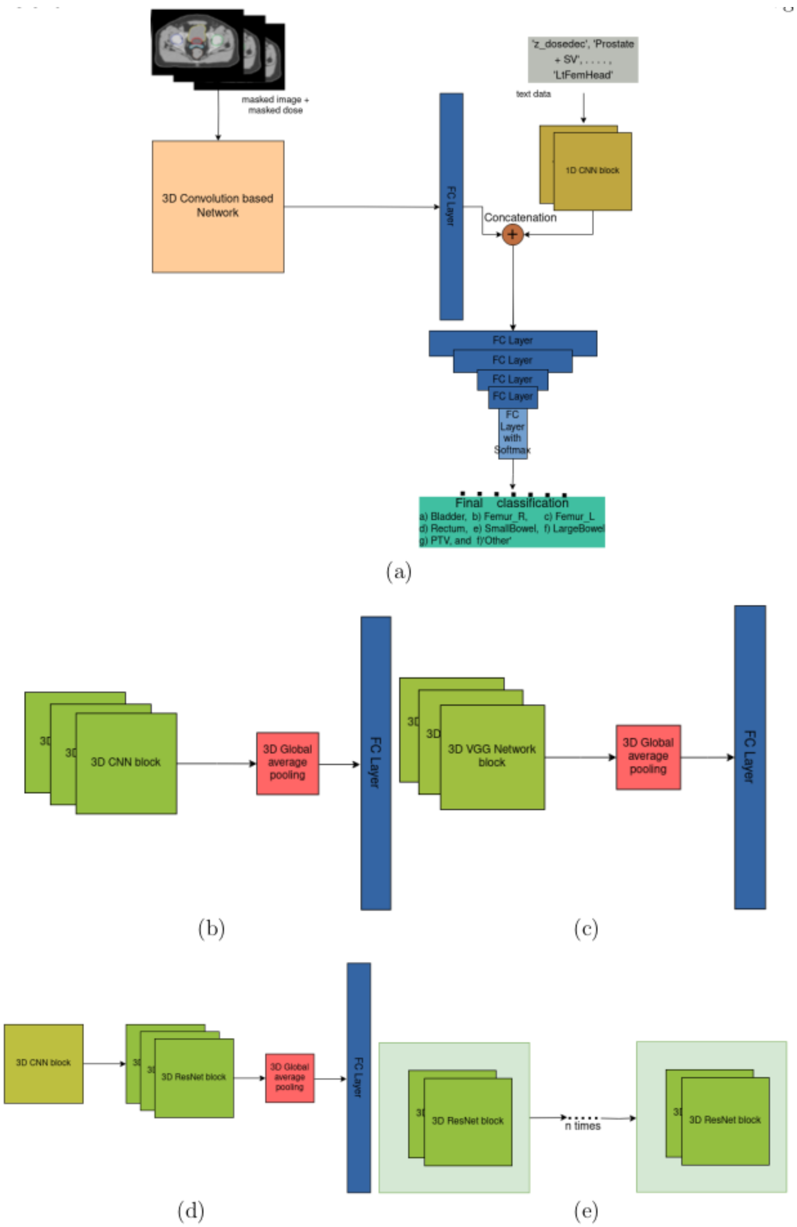


Fig. 3.: Pictorial Representation of our Masking Step in the case of (**a**) images, and (**b**) doses of Prostate RT Patients.

the readers for future replication purposes and enhances the clarity of the detailed architecture. A CNN [83] is a DNN that uses convolution on the input and directs the result of convoluting to the next layer. Although multi-layer perceptron neural networks can be trained on textual data, their performances are often overshadowed by CNNs [84] that can slide a window of user-defined size on the input data. CNN was first used for sentence classification, i.e., a particular type of text classification task, in 2014 [85]. The hyper-parameters used in CNNs are the number of input and output channels, convolutional kernels, and filters. RNNs such as Simple Recurrent Unit (SRU), Long Short Term Memory (LSTM) [86], etc., are a class of DNNs that work on the cyclical connections between nodes, exhibiting temporal dynamic behavior. They are capable of using their internal state or memory to train inputs of varying length sequences [87].



(a)

(b)

(c)

(d)

(e)

Fig. 4.: Overview of our DNN architecture: (**a**) General architecture of our 3D Convolution-based network on vision-dose data and 1D CNN on text data, (**b**) Customized 3D CNN on vision-dose, (**c**) Customized 3D VGG network on vision-dose, (**d**) Customized 3D ResNet on vision-dose, (**e**) Stacked customized 3D VGG Network blocks with nested 3D ResNet blocks inside each block on vision-dose data.

Recurrent Networks perform well on contextual data as these learn to remember the previous steps. Since, the textual data in our case contains random physician-given names of the structures, there is not much context in our textual data. Due to the absence of context, CNNs are more effective here. Hence, we have chosen CNNs over recurrent networks in this case. We build a DNNs on these multimodal datasets with the mentioned networks where the two features (vision-dose and text) are combined, intermediately and fed through hidden layers before classifying the multi-classes using a classifier in the end. We have used a batch size of 32, categorical cross-entropy as loss, 200 number of epochs (except 50 epochs for the architectures with LeakyReLU activation as LeakyReLU converges faster than ReLU [88]), and Adam as optimizer with an initial learning rate of 0.001 and staircase decay steps of 10000 at 0.96 decay rate, in training all our DNNs.

The general architecture of our model is shown in Figure 4a, where we vary the 3D convolution-based network on the vision-dose data while keeping the other parts unchanged. We used two consecutive 1D CNN layers of 256 and 128 filters, respectively, with Rectified Linear Unit (ReLU) [89] activation function on text embeddings. We have used 1D max pooling after each convolution layer with a constant kernel size of 8. After these two layers, the features are flattened into a 1D vector and concatenated with vision-dose features for the multimodal model. In all the architectures, we have concatenated the textual and vision-dose features in this intermediate stage.

4.2.1.1 CNN

Our 3D CNN architecture on vision-dose data consists of several convolution blocks in sequence, as shown in Figure 4b. Each convolution block consists of 3D convolution layer with ReLU activation function, 3D max-pooling layer, a spatial 3D dropout layer with 20% dropout, and finally, a batch-normalization layer. The three convolution blocks first consist of a 3D convolution layer where the number of output channels or convolution filters increases with the increase in the number of convolution blocks, i.e., the increase in the model depth. The convolution kernel size is chosen as 3 except for the first block, where the kernel size is 9. A global 3D average-pooling is performed at the end of the convolution blocks for feature reduction based on the global features. These features, passed through a fully connected layer are finally concatenated with textual features and fed to the final classifying layer with Softmax [90] activation function through subsequent hidden layers.

4.2.1.2 VGG Network

The VGG network [91], developed by the Visual Geometry Group at Oxford University, caters the first idea of using blocks. Blocks or the repeated structures in code with any modern DL framework is very easy to implement and hence has gained immense popularity. The VGG network comprises two different types of blocks, convolutional blocks with max-pooling and fully connected dense blocks.

Instead of using pre-trained VGG network models, we define our customized VGG architecture, where we can tweak the parameters freely. We took the 2D VGG network as an inspiration to build the 3D VGG network in our case that can operate on 3D vision-dose datasets, as shown in Figure 4c. The VGG network, firstly, consists of three VGG blocks with 1, 2, and 4 number of convolutions, respectively, in the blocks in order. Each VGG block consists of a defined number of 3D convolution layers with ReLU activation and a kernel of size 3. 3D max-pooling with strides = 2 is then performed at the end of all the convolutions in each VGG block. Next, the output of the three consecutive VGG blocks is fed to a 3D global average-pooling layer. The features are then passed through a fully connected layer and are integrated intermediately with the textual features and are finally fed to the classifier like in the case of the CNN model. Since, VGGNet performed the best on some combinations of the data as reported later, we further customized the VGGNet model for experimentation. Firstly, we deepened the VGGNet by replacing the convolution layers inside each block with ResNet layers (discussed in the next subsection), as shown in Figure 4e. This particular architecture is inspired by the recently published model in [92]. Secondly, we investigated the performance of initial VGGNet with ReLU activation function in each convolutional layer and ultimately adding a LeakyReLU activation after the max-pooling layer. However, these architectures were not effective in upgrading the best performing results of the prior VGGNet model.

4.2.1.3 ResNet

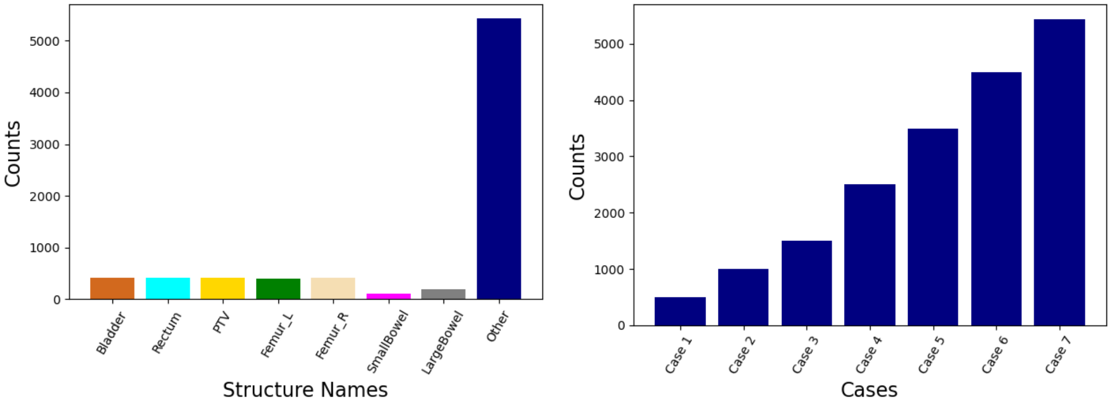
A residual neural network (ResNet) is an artificial DNN that is based on skipping connections or shortcuts to jump over some layers. ResNet [93] models have typically been implemented with double- or triple-layer skips where a ReLU activation and batch normalization layers are used in between. Prior to the invention of ResNet, the CNN architecture continued going deeper and deeper where ImageNet [94], VGG network, and GoogleNet [95] had 5, 19, and 22 layers, respectively. However, deep networks are often hard to train when the network depth is increased by simply stacking layers together. These networks lead to overfitting, as in the case of back-propagation of the gradient to earlier layers, repeated multiplications may potentially make the gradient very small. Although GoogleNet was instrumental in adding an auxillary loss in a middle layer for an added supervision, it was not much effective. Hence, the core idea of ResNet by introducing shortcut connections represented a major breakthrough in this domain.

Similar to the case of VGG, we use our customized 3D ResNet architecture on vision-dose data to tweak the parameters freely as shown in Figure 4d. Inspired by the 2D ResNet, we also developed a 3D ResNet model with 3D convolutions and max-pooling that can work on 3D data. ResNet model with Our ResNet architecture consists of a convolution block, followed by three sequential ResNet blocks. The Convolution block consists of a 3D convolution layer with 32 filters, kernel size of 9 and a stride of 2, followed by batch normalization, ReLU activation and 3D max-pooling. All our ResNet blocks consist of 2 residual blocks. Each residual block consists of a 3D convolution (kernel size: 3) with batch normalization and ReLU activation, followed by another 3D convolution (kernel size: 3) with batch normalization. The output of this is added to the input of each residual block with ReLU activation and passed down to the next layer. We perform a 3D max-pooling, followed by a 20% dropout after the residual blocks. In the first residual blocks of the last two ResNet blocks, the output of the two subsequent 3D convolutions is added to the input of the residual layer after convolution through a kernel of size 1 and hence, ReLU activation is applied. We perform a 3D global average-pooling after the third ResNet block and pass it through a fully connected layer for intermediate concatenation with textual features. Next, classification is performed exactly in a similar way as mentioned in the previous two cases.

4.2.1.4 Sampling the 'Other' Classes

Data imbalance is a very important challenge in training DL architectures. It negatively impacts the performance by biasing it towards the majority class depending upon the level of imbalance although a number of studies have demonstrated that it might not be a vital factor [8]. In the prostate cancer dataset, a high level of imbalance is observed, which contains an extremely high representation of the 'Other' class, compared to PTV and the other six OAR classes, as illustrated in Figure 5a. Our training dataset contains 416, 414, 418, 407, 412, 116, 188, and 5432 samples from ‘Bladder’, ‘Rectum’, ‘PTV’, ‘Femur_L’, ‘Femur_R’, ‘SmallBowel’, ‘LargeBowel’, and ‘Other’ classes, respectively, where the majority class, i.e., ‘Other’ has about thirteen times the number of samples present in the largest minority class, i.e., ‘PTV’. It constitutes about 70% of the overall prostate structure name standardization dataset, which clearly outnumbers the plethora of structures under consideration in the prostate. Since the majority class presents a substantial amount of imbalance, sampling is potentially very useful in this case.

Sampling comes in two types: undersampling and oversampling. In our case, we have only one majority class and it is easier to undersample the majority class to prevent the model from biasing towards the majority class. Plus, oversampling the minority class to some extent will make the models bias towards these classes with high chances of overtraining them. Hence, we undersampled the majority class and randomly selected 500 samples from that majority class in each case as the largest minority only contains 418 samples in the training set. The DNNs performed roughly the same with or without undersampling the ‘Other’ class and hence, used undersampling to compare the various DNNs for data preparation and model selection. Next, we show the performance of the DNNs when the amount of undersampling is varied. Since, we undersampled the majority classes to 500 samples initially, we show the performance of the DNNs when the majority class was undersampled to 500 (about 90.8% undersampling), 1000 (about 81.6% undersampling), 1500 (about 72.4% undersampling) 2500 (about 54% undersampling), 3500 (about 35.6% undersampling), 4500 (about 17.2% undersampling), and samples and not oversampled at all as shown in Figure 5b. Without considering the last case, the percentage of undersampling ranges from 17.2% (4500 samples) to 90.8% (500 samples). Performances of the DNNs show that there is a small trade-off between undersampling and model performance.



(a) (b)

Fig. 5.: Bar plots showing (**a**) the distribution of various data classes in the RT Prostate Structure Naming Dataset, and (**b**) the variation in the number of samples from the 'Other' class in different cases of consideration.

4.2.2 Results

In most cases, our architectures perform strongly on our final model on vision-dose and text data with or without undersampling. Our models, which are built on both vision-dose and text data, consist of two different neural network architectures: one for image or dose or both and another for text. In all these cases, the first mentioned DNN is used on the vision-dose data, and the latter one is used on the text data. So, a '3D CNN and 1D CNN' method means that a 3D CNN is used on images or doses or both, and 1D CNN is applied to the text. We explain our evaluation metrics and analyze our results in the subsections below.

4.2.2.1 Evaluation Metrics

In order to evaluate the models, we have used the following metrics: precision, recall, and F1-Score, as proposed by the earlier works on this data. These metrics can be macro-averaged, i.e., independently calculating the values for each class and then averaging the values across the different classes, or weighted averaged, i.e., independently calculating the values across different classes and then doing a weighted average of the values of different classes. In the case of a highly imbalanced data set, using weighted averaged metrics will potentially skew the values towards the majority class/classes. Hence, we used macro-averaged metrics across the multi classes (eight classes) instead of weighted averaged metrics as we were particularly interested in seeing the models' effectiveness towards the minority classes.

4.2.2.2 Data Preparation and Selection

The performances of deep networks using both vision-dose and text data on prostate cancer patients are shown in Table 5. The table shows the performance of CNNs on textual data and CNNs, ResNet, and VGG network on the vision-dose data, where we vary the nature of the input vision-dose data. In the first case, we input the bitmaps, delineated images, and doses as 3D arrays with three channels to the model along with texts. In the second case, we input the masked images and doses (as mentioned above) as 3D arrays with two channels to the model along with texts. This way, we made a comparison between a time efficient and a not so time efficient case only to highlight that the time-efficient case performs better than the other with respect to the prediction by the deep multimodal network. These evaluated network performances reported in Table 5 substantiate this. This is because the learning of a DNN becomes more challenging with large amounts of input data, and in our second case, we present the same information to the neural network but with less data space when compared to the first case.

Data Modality	Method	Precision (in %)	Recall (in %)	F1-Score (in %)
struc+ image+ dose+ text	3D CNN and 1D CNN	91.93	92.93	92.42
struc+ image+ dose+ text	3D ResNet and 1D CNN	92.72	93.74	93.2
struc+ image+ dose+ text	3D VGG and 1D CNN	93.51	92.99	93.19
masked image+ masked dose+ text	3D CNN and 1D CNN	93.65	93.29	93.4
masked image+ masked dose+ text	3D ResNet and 1D CNN	91.05	94.83	92.76
masked image+ masked dose+ text	3D VGG and 1D CNN	94.66	94.39	94.45

Table 5.: Performance of the CNN-based Models for the Prostate cancer patients during data selection.

4.2.2.3 Model Selection

The performance of the various DNN architectures for the prostate cancer patients when fitted on the masked vision-dose and text data are shown in Table 6. We experimented with 3D CNN, 3D ResNet and 3D VGG Network architectures for the vision-dose data while we used 1D CNN for the textual part. The 1D CNN on texts performs very well with respect to the macro-averaged F1-Scores. It can be seen that the VGG network can learn geometric features from the data very well (shown in the next paragraph) and the scope of performance improvement in this case is much limited compared to other cases. ResNet and simple 3D CNN also exhibit strong performances in some cases.

The performances of the various architectures for prostate cancer patients with varying combinations of data modalities are shown in Table 6. It is evident from the first nine rows of the table that the networks can learn the geometric features to a great extent from the masked vision-dose data together, whereas their performance drops when either masked image or masked dose is only considered. This establishes the fact that the addition of relevant data modalities can lead to a significant leap in performance. With either of these geomteric modalities, the networks report F1-

Scores in the low 70s while, with both of these geometric modalities the F1-Scores elevate to low 90s. The F1-Scores, as reported in the table, point out that the performance of VGG and ResNet eclipses the performance of CNN, consistently to some extent when either masked dose or masked image or both masked dose-image data are considered. On the other hand, 1D CNNs exhibit a solid performance of 93.17% F1-Score when trained on the textual data itself. When the text was added on top of the image or dose data, the performance of VGG only improved over the performances of other models. Thus, it shows that the 3D VGG network and 1D CNN architecture can improve learning by adding data if the data contain information vital for decision-making. With all these data modalities, 3D ResNet and 1D CNN also shows improved performance with a small amount of majority class undersampling which is discussed in the next subsection. This statement is further justified by the performance of our final data model, where the DNNs further advanced their learning capability on training on multimodal vision-dose and text data. The performances of the DNNs on this final data model are reported in Table 5.

4.2.2.4 Model Performance on Varying the Number of Majority Class Samples

The performance of the various DNN architectures for prostate cancer patients when trained on a varying number of samples from 'Other' classes are shown in Table 7. The variation in performances of our architectures (3D CNN and 1D CNN, 3D ResNet and 1D CNN, and 3D VGG network and 1D CNN) with the variation in number of samples from the majority class is shown in Figure 6. The architectures exhibit a strong performance with or without undersampling the 'Other' class, although performances are slightly improved in most cases with varying degrees of undersampling. The table displays the model performances when the 'Other' class

Masked Image	Masked Dose	Text	Method	Precision (in %)	Recall (in %)	F1-Score (in %)
✓	-	-	3D CNN	71.75	74.81	72.99
✓	-	-	3D ResNet	73.94	74.24	73.92
✓	-	-	3D VGG	76.19	74.46	74.82
-	✓	-	3D CNN	74.18	70.79	72.17
-	✓	-	3D ResNet	74.01	62.94	66.39
-	✓	-	3D VGG	81.53	77.98	79.45
✓	✓	-	3D CNN	93.6	91.52	91.93
✓	✓	-	3D ResNet	94.23	93.65	93.82
✓	✓	-	3D VGG	93.0	94.9	93.8
-	-	✓	1D CNN	92.18	94.24	93.17
✓	-	✓	3D CNN and 1D CNN	93.31	92.49	92.83
✓	-	✓	3D ResNet and 1D CNN	91.33	95.18	93.15
✓	-	✓	3D VGG and 1D CNN	92.24	91.35	91.71
-	✓	✓	3D CNN and 1D CNN	91.86	93.66	92.61
-	✓	✓	3D ResNet and 1D CNN	90.13	94.97	92.29
-	✓	✓	3D VGG and 1D CNN	91.82	93.42	92.54
✓	✓	✓	3D CNN and 1D CNN	93.65	93.29	93.4
✓	✓	✓	3D ResNet and 1D CNN	91.05	94.83	92.76
✓	✓	✓	3D VGG and 1D CNN	94.66	94.39	94.45

Table 6.: Model performances for the Prostate cancer patients with varying data modalities.

was undersampled at 500, 1000, 1500, 2500, 3500, 4500 samples, respectively. The F1-Scores provided by 3D CNN and 1D CNN vary between 77.26% and 93.46% at different levels of undersampling, whereas in the case of 3D ResNet and 1D CNN, it varies between 80.98% and 94.35%. F1-Scores for 3D VGG network and 1D CNN varies between 76.33% and 94.45%. Overall, it can be pointed out that the 3D VGG with 1D CNN consistently performs well in all the cases and it either outperforms or performs at par with 3D ResNet and 1D CNN. These models have a slight edge over 3D CNN along with 1D CNN on text. This is because VGG network and ResNet are deeper than just CNNs which has the advantage of using more parameters to

Table 7.: Model performances for the Prostate cancer patients with variation in the majority class samples.

Total Samples from 'Other' Class	Method	Precision (in %)	Recall (in %)	F1-Score (in %)
500	3D CNN and 1D CNN	70.62	89.32	77.26
500	3D ResNet and 1D CNN	76.47	87.21	80.98
500	3D VGG and 1D CNN	71.71	83.61	76.33
1000	3D CNN and 1D CNN	90.99	94.24	92.54
1000	3D ResNet and 1D CNN	89.08	96.91	92.57
1000	3D VGG and 1D CNN	89.82	95.16	92.25
1500	3D CNN and 1D CNN	91.65	95.46	93.46
1500	3D ResNet and 1D CNN	86.63	95.3	90.34
1500	3D VGG and 1D CNN	88.79	96.35	92.05
2500	3D CNN and 1D CNN	87.82	94.7	90.97
2500	3D ResNet and 1D CNN	88.86	96.7	92.38
2500	3D VGG and 1D CNN	92.56	95.76	94.09
3500	3D CNN and 1D CNN	91.74	93.75	92.72
3500	3D ResNet and 1D CNN	93.6	95.47	94.35
3500	3D VGG and 1D CNN	92.48	95.36	93.83
4500	3D CNN and 1D CNN	92.34	92.56	92.38
4500	3D ResNet and 1D CNN	91.54	95.45	93.37
4500	3D VGG and 1D CNN	91.42	92.95	92.12
5432 (No Sampling)	3D CNN and 1D CNN	93.65	93.29	93.4
5432 (No Sampling)	3D ResNet and 1D CNN	91.05	94.83	92.76
5432 (No Sampling)	3D VGG and 1D CNN	94.66	94.39	94.45

learn more from the dataset. Furthermore, the idea of using residuals from a network performs well on its own and overshadows the performance of 3D CNN. On the other hand, our VGG network or ResNet architecture is neither too deep nor too shallow which is useful for effective training and minimizing the chances of probable overtraining. The performances of the various DNNs vary with the degree of undersampling. The performance of VGG network gives state-of-the-art results (F1-Score: 94.45%) without the need for any undersampling although it increases the training time to some extent. The performance of 3D ResNet reaches its crest when the majority class is undersampled at 3500 samples (F1-Score: 94.35%), which requires less training time and at the same time obscures the performance of the other architectures. The third best performance is also reported by 3D VGG and 1D CNN when the majority class is undersampled at 2500 samples (F1-Score: 94.09%). When the majority class samples are undersampled at 1500 samples, 3D CNN and 1D CNN records its best performance with an F1-Score of 93.46%. In most cases, the 3D VGG network and ResNet successfully eclipses the performance of CNN on the vision-dose data along with 1D CNNs on the text, which establishes the superiority of deeper networks in learning the image and dose features over others.

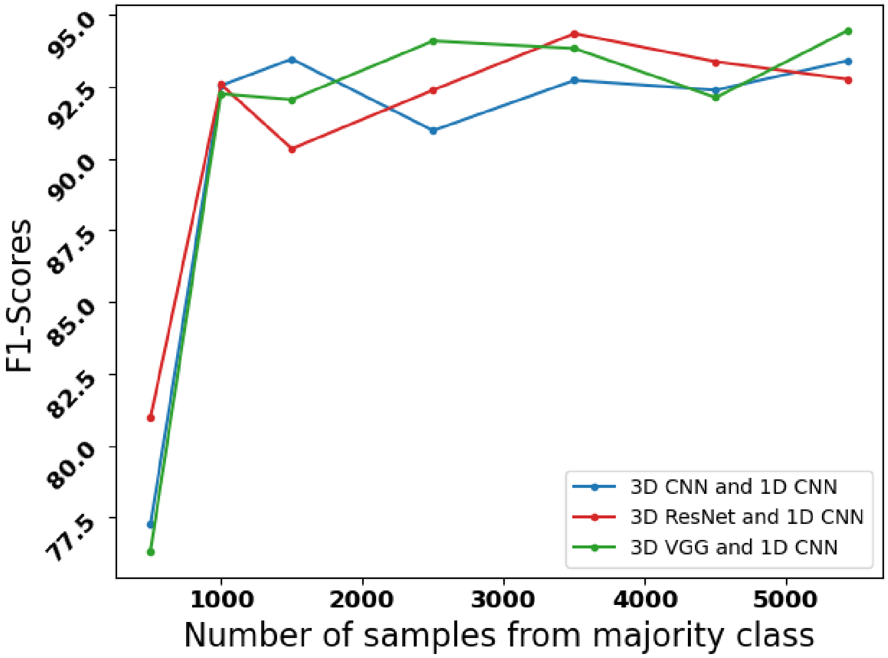


Fig. 6.: Line curve showing the variation in F1-Scores of the models with variation in the number of samples from the majority 'Other' class.

The performances on the further modifications of the initial VGGNet, i.e., 3D VGGNet with nested ResNet and 3D VGGNet with LeakyReLU are reported in Ta-

Table 8.: Model performances with 3D VGG nested ResNet and 3D VGG with Leaky ReLU activation for the Prostate cancer patients while varying the majority class samples.

Total Samples from 'Other' Class	Method	Precision (in %)	Recall (in %)	F1-Score (in %)
500	3D VGG with nested ResNet and 1D CNN	86.75	96.37	90.62
500	3D VGG with LeakyReLU and 1D CNN	85.57	96.28	89.74
1000	3D VGG with nested ResNet and 1D CNN	87.37	95.49	90.92
1000	3D VGG with LeakyReLU and 1D CNN	83.56	97.29	88.55
1500	3D VGG with nested ResNet and 1D CNN	89.67	96.27	92.64
1500	3D VGG with LeakyReLU and 1D CNN	91.66	95.4	93.44
2500	3D VGG with nested ResNet and 1D CNN	90.8	93.38	92.02
2500	3D VGG with LeakyReLU and 1D CNN	87.92	96.26	91.5
3500	3D VGG with nested ResNet and 1D CNN	90.57 4	95.6	92.89
3500	3D VGG with LeakyReLU and 1D CNN	88.56	94.75	91.44
4500	3D VGG with nested ResNet and 1D CNN	90.97	93.33	92.11
4500	3D VGG with LeakyReLU and 1D CNN	92.69	92.35	92.51
5432 (No Sampling)	3D VGG with nested ResNet and 1D CNN	92.19	94.71	93.36
5432 (No Sampling)	3D VGG with LeakyReLU and 1D CNN	91.95	94.29	93.04

ble 8. The F1-Scores do not show improvement over the best performance as reported by 3D VGGNet in Table 7 although the performances are comparable to that of other networks on the vision-dose and text data. For the case where only 500 samples were selected from the majority class, these two architectures show superior performances over the other architectures. In all our architectures, it can be noted that Recall is higher than Precision in almost all cases which shows that our architectures are effective in diminishing the effect of false negatives.

Further analysis of the top three model performances reveal that the architectures with or without majority class undersampling perform decently across most of the classes. However, the models show that it is harder for them to learn the 'PTV' and

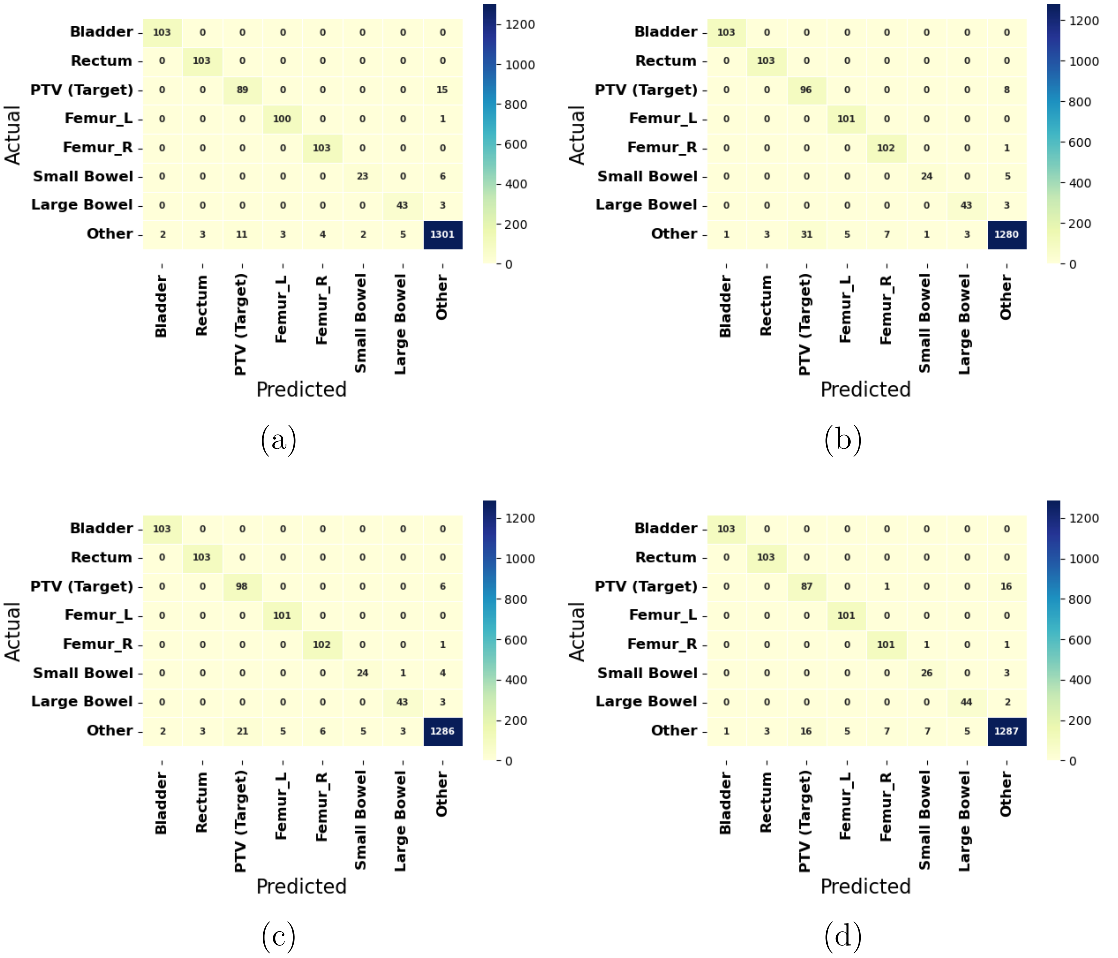
Table 9.: Class-wise performances of the top three Models for the Prostate cancer patients

Class	VGG (with 5432 Majority Class Samples)			ResNet (3500 Majority Class Samples)			VGG (2500 Majority Class Samples)		
	Precision (in %)	Recall (in %)	F1-Score (in %)	Precision (in %)	Recall (in %)	F1-Score (in %)	Precision (in %)	Recall (in %)	F1-Score (in %)
Bladder	98.1	100	99.04	99.04	100	99.52	98.1	100	99.04
Rectum	97.17	100	98.56	97.17	100	98.56	97.17	100	98.56
PTV (Target)	89.0	85.58	87.25	75.59	92.31	83.12	82.35	94.23	87.89
Femur_L	97.09	99.01	98.04	95.28	100	97.58	95.28	100	97.58
Femur_R	96.26	100	98.1	93.58	99.03	96.23	94.44	99.03	96.68
Small Bowel	92.0	79.31	85.19	96.0	82.76	88.89	82.76	82.76	82.76
Large Bowel	89.58	93.48	91.49	93.48	93.48	93.48	91.49	93.48	92.47
'Other'	98.11	97.75	97.93	98.69	96.17	97.41	98.92	96.62	97.76

'Small Bowel' classes compared to the 'Other' classes. One of the reasons behind the comparatively poorer learning of the 'PTV' class is that the range of randomness in the physician-given names is vast compared to the 'Other' classes. As for 'Small Bowel', it has the lowest representation of samples in the dataset. Hence, the models find it tough to learn from the infinitesimally smaller representation of the minority 'Small Bowel' class. For effective learning, data augmentation or oversampling can be considered in the future. Apart from these two classes, the models displayed a superior performance with a consistent F1-Score of more than 90.0%. The class-wise performance of the top three models for the prostate cancer patients are shown in Table 9.

4.3 Discussion

In this section, we discuss the performances of various multimodal models for RT prostate structure name standardization. Since all the data types of the multimodal data in each case are not homogeneous, early integration of the overall data was not performed. Instead, we performed early integration of the multimodal geometric data, i.e., vision and dose. Textual data were first trained in parallel and then immediately



(a)

(b)

(c)

(d)

Fig. 7.: Confusion Matrices of the best three predictions for the Prostate Cancer Patients by the F1-Scores are shown in (**a**) 3D VGG network and 1D CNN without undersampling, (**b**) 3D ResNet and 1D CNN with 3500 majority class samples, and (**c**) 3D VGG network and 1D CNN with 2500 majority class samples. Confusion Matrices of the best predictions of the architecture for the Prostate Cancer Patients by the F1-Scores are shown in (**d**) 3D CNN and 1D CNN with 1500 majority class samples.

integrated with the geometric feature inside the DNN architecture. Undersampling the ‘Other’ structures to a small extent boosted the performance of the DNNs trained on the entire vision-dose and textual data. The degree of undersampling is also essential for tuning the model performance, which establishes that an intermediate amount of undersampling works best in the case of ResNet. In many cases, we observed that though the overall accuracy decreases in the case of the multimodal models compared to the textual single view model, the macro-averaged F1-Score increases, which shows better learning across the different minority classes and less bias. One of the limitations of this research is the absence of more recent deep learning models which will be addressed in our future work. These may include experimenting with advanced models such as DenseNet, Squeeze Net, ENet or some vision transformers and comparing their performances. That apart, we did not apply any data augmentation methods which will be also explored in the future. We also plan to explore how undersampling the majority samples or oversampling the minor samples impacts the performance of these advanced models.

It is established for the first time that an architecture considering the 3D masked image and masked dose with text leads to an overall performance improvement of RT structure name standardization over using the hand-crafted geometric features with text. In addition, we are the first to show that using the masked image and masked dose is more time- and performance-efficient when compared to using bitmaps, delineated images, and doses. Although the performance of the 1D CNN on the textual data is quite good, the performance enhancement by adding the geometric data still shows that the neural network model can perform better with the help of information contained in the data from other modalities. Interestingly, we also observed that using a 3D VGG network or 3D ResNet on the vision-dose data and 1D CNN on the textual data offers a slight edge over other DNNs for the respective modalities. The VGG architecture apparently overshadows the other architectures without any amount of majority undersampling. Hence, we introduced 3D VGGNet with nested ResNet and 3D VGGNet with LeakyReLU activation along with 1D

CNN on textual data to further investigate the scope of performance improvement. While these architectures produced comparable results, they could not shroud the performance of the initial VGGNet with 1D CNN. Hence, the 3D VGG network on the masked vision-dose data and 1D CNNs on text exhibit the best performance without the majority class undersampling and establishes the state-of-the-art with a macro-averaged F1-Score of 94.45% whereas, 3D ResNet and 1D CNN records the second best performance (F1-Score: 94.35%) when the majority class is undersampled at 3500. The confusion matrices of the top three models by their performances along with that of the top model for 3D CNN and 1D CNN are shown in Figure 7. Hence, our unique deep-learning-based methods considering the heterogeneous multimodal data provide state-of-the-art results in automating the prediction of standard prostate RT structure names.

4.4 Summary

We first used a feed forward neural network on the SVD-reduced features (2304 units) of the geometric data and CNN on the textual data. Extensive experimentation with undersampling the majority class and separately training each modality revealed that the network can learn very well from the texts. Addition of SVD-reduced features along with texts led to a deterioration in the performance of our DNN architecture due to information loss in reducing the geometric features. Hence, training the whole 3D vision-dose data along with texts is significant to experiment with. The presence of the contextual information in the 3D vision-dose data is vital for the algorithm to better fathom the data. As a result, we performed extensive experimentation with CNN, ResNet, VGGNet or ResNet-based deep VGGNet on the 3D vision-dose data along with 1D CNNs on the texts where we intermediately stacked the features. Our results show that majority class undersampling sometimes gives better results. The top three results on the prostate structure name standardization dataset is given by 3D VGGNet and 1D CNN without sampling, 3D ResNet and 1D CNN with majority class undersampling at 3500 samples and 3D VGGNet and 1D CNN with majority class undersampling at 2500 samples, achieving an F1-Score of 94.45%, 94.35% and 94.09%, respectively.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this thesis, we reported the performances of various multimodal models for RT prostate structure name standardization.

5.1 Contributions

Since all the data types of the multimodal data in each case are not homogeneous, early integration of the overall data was not performed. Instead, we performed early integration of the multimodal geometric data, i.e., vision and dose. Textual data were first trained in parallel and then immediately integrated with the geometric feature inside the DNN architecture. Undersampling the ‘Other’ structures to a small extent boosted the performance of the DNNs trained on the entire vision-dose and textual data. The degree of undersampling is also essential for tuning the model performance, which establishes that an intermediate amount of undersampling works best in the case of ResNet. In many cases, we observed that though the overall accuracy decreases in the case of the multimodal models compared to the textual single view model, the macro-averaged F1-Score increases, which shows better learning across the different minority classes and less bias.

It is established for the first time that an architecture considering the 3D masked image and masked dose with text leads to an overall performance improvement of RT structure name standardization over using the hand-crafted geometric features with text. In addition, we are the first to show that using the masked image and masked dose is more time- and performance-efficient when compared to using bitmaps, delineated images, and doses. Although the performance of the 1D CNN on the textual data is quite good, the performance enhancement by adding the geometric data still shows that the neural network model can perform better with the help of information contained in the data from other modalities. Interestingly, we also observed that using a 3D VGG network or 3D ResNet on the vision-dose data and 1D CNN on the textual data offers a slight edge over other DNNs for the respective modalities. The VGG architecture apparently overshadows the other architectures without any amount of majority undersampling. Hence, we introduced 3D VGGNet with nested ResNet and 3D VGGNet with LeakyReLU activation along with 1D CNN on textual data to further investigate the scope of performance improvement. While these architectures produced comparable results, they could not shroud the performance of the initial VGGNet with 1D CNN. Hence, the 3D VGG network on the masked vision-dose data and 1D CNNs on text exhibit the best performance without the majority class undersampling and establishes the state-of-the-art with a macro-averaged F1-Score of 94.45% whereas, 3D ResNet and 1D CNN records the second best performance (F1-Score: 94.35%) when the majority class is undersampled at 3500 samples. Hence, our unique deep-learning-based methods considering the heterogeneous multimodal data provide state-of-the-art results in automating the prediction of standard prostate RT structure names.

5.2 Limitations and Future Work

In this section, we discuss the limitations associated with our proposed models. Moreover, we discuss possible extensions of this work that may contribute to the volume of research on radiotherapy structure name standardization.

5.2.1 Limitations

In the clinical domain, ”Black and white” binary medical images may contain ill-defined contrasts between two tissues and a learning process on such images can be too constraining [96]. Also, assigning a single hard label can cause a detrimental approximation which has lead to the emergence of soft prediction on non-binary images [96, 97]. In our case, we have used binary hard image masks where the voxels present on structure edges contain a mixture of tissues, causing a partial volume effect. Soft labels have been used successfully in the medical domain for image segmentation like SoftSeg [96]. Hence, using binary structure sets is a limitation of our pipeline and possible future work would involve the use of soft image labels. The limitations of our work are itemized as follows:

- Using hard binary structure sets is the first limitation of our work.

- Secondly, the absence of more recent deep learning models is another limitation which will be addressed in future work. These may include experimenting with advanced models such as DenseNet, Squeeze Net, ENet or some vision transformers and comparing their performances.

- Recent research has shown that attention based models with deep networks have performed very well on both vision and textual data. In many cases, attention mechanisms with deep networks have eclipsed the performances of the standalone deep networks. Therefore, application of attention mechanisms on top of the deep networks will be an interesting avenue of future work.

- That apart, we did not apply any data augmentation methods which is another limitation of our pipeline. We also plan to explore how undersampling the majority samples or oversampling the minor samples impact the performance of these advanced models.

- The current list of OARs identified for the prostate dataset is per the VA-ROQS project requirement, which has selected these OARs in consensus with an expert team of physicians. Radiation oncologists also delineate other types of OARs for each patient, such as Kidney (left and right) and Liver, in prostate cancer patients. Although these are not critical OARs in prostate cancer treatment, we believe building a system to identify and standardize all structures delineated according to the TG-263 guideline provides the radiation therapy healthcare institutes with an opportunity to produce a robust dataset for downstream analysis projects. Failing to standardize the non-OAR names is another limitation of our work.

5.2.2 Future Work

Segmentation in medical images is already a popular research direction in medical informatics [98, 99, 100]. One such work includes Psi-Net [101] which is a shape and boundary aware joint multi-task deep network, containing one encoder and three decoders for medical image segmentation. Another such development is a high resolution multi-scale encoder-decoder where encoder-decoder is densely connected, containing skip connections and extra deeply supervised high-resolution pathways for blurry medical image segmentation [102]. Similarly, structure delineation or segmentation of the structures from the organ bitmaps can lead to a new direction of research that can be performed on the existing prostate structure standardization dataset.

Besides that, geometric features and dosimetric features have been considered alongside images and texts for standardization of breast radiotherapy structure names [27]. The geometric features that were extracted from the structure images contain nine positional and volumetric features like the coordinates of the structure centroid, magnitude and direction of the vector joining (0,0,0) and the centroid, number of voxels, etc. Instead of considering the dose images, this article has mentioned the usage of a total of 10 dosimetric features for each structure such as minimum dose, median dose, mean dose, maximum dose, V%20.0, V%10.0, etc. Hence, extraction of these geometric and dosimetric features from the prostate structure images and using an advanced learning method with these features along with images and texts is another interesting avenue of future work.

The future works in the domain of improved structure name standardization are itemized as follows:

- Firstly, experimenting with advanced models such as DenseNet, Squeeze Net, ENet or some vision transformers and comparing their performances can be a relevant future work. Also, trying out attention based deep learning models on this standardization task with vision and text data can potentially contribute to this domain in the future.

- Secondly, we did not apply any data augmentation methods which will be also explored in the future. Future work will also include undersampling the majority samples or oversampling the minor samples and compare the performance of the advanced models.

- Thirdly, using an advanced learning algorithm with the dosimetric and geometric features along with images and texts will be a stimulating future work.

- Annotation of the non-critical OARs and standardizing these structure names is one of the vital future works.

- Other future works using the standardized structure sets include dose outlier detection and automated structure delineation. Automated structure delineation or contouring is a new direction of research that can be done by using the structure name standardization dataset. Using popular and efficient models in this domain, prostate structure delineation can be automated with hard binary/soft images which can potentially save the physician's time in manually contouring the images.

- With automated structure delineation, the next vital prospective direction of future work is joint structure delineation and standardization. One possible way of joint structure delineation and standardization is a pipeline-based model where structure delineation is done first followed by structure standardization. The parameters of these two tasks are independent of each other and the tasks can be performed sequentially in the pipeline. Another approach can be a joint optimization based method for automated structure delineation and standardization. In this case, the two tasks share some common parameters and the values can be updated at each iteration of the combined tasks. The two tasks can potentially share a common optimizer and the loss function needs to be computed based on the losses of the two individual tasks. It will be intuitive to compare the performance of the pipeline based automated structure delineation and standardization with the joint optimization based automated structure delineation and standardization in the future.

Appendix A

ABBREVIATIONS

DICOM	Digital Imaging and Communications in Medicine
OAR	Organ at Risk
RT	Radiotherapy
PTV	Planning Target Volume
VHA	Veterans Health Administration
VCU	Virginia Commonwealth University
CT	Computed Tomography
MR	Magnetic Resonance
AAPM	American Association of Physicists in Medicine
ASTRO	American Society for Radiation Oncology
TG	Task Group
NLP	Natural Language Processing
ML	Machine Learning
AI	Artificial Intelligence
IRB	Institutional Review Board
TPS	Treatment Planning System
ROQS	Radiation Oncology Quality Surveillance Program
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
RT	Radiation Therapy
IoU	Intersection over Union

DSC	Dice Similarity Coefficient
VQA	Visual Question Answering
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
MRR	Mean Reciprocal Rank
NDCG	Normalized Discounted Cumulative Gain
WUPS	Wu-Palmer Similarity
BioBERT	Bidirectional Encoder Representations from Transformers for Biomedical Text Mining
DNN	Deep Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
SRU	Simple Recurrent Unit
LSTM	Long Short Term Memory
GNN	Graph Neural Network
ResNet	Residual Network
VGG	Vision Geometry Group
NROP	National Radiation Oncology Program
DMN	Dynamic Memory Network
GRU	Gated Recurrent Unit

REFERENCES

[1] Charles S Mayo et al. “American Association of Physicists in Medicine Task Group 263: standardizing nomenclatures in radiation oncology”. In: *International Journal of Radiation Oncology* Biology* Physics* 100.4 (2018), pp. 1057–1066.

[2] Jean L Wright et al. “Standardizing normal tissue contouring for radiation therapy treatment planning: an ASTRO consensus paper”. In: *Practical radiation oncology* 9.2 (2019), pp. 65–72.

[3] Stanley H Benedict et al. “Overview of the American Society for Radiation Oncology–National Institutes of Health–American Association of Physicists in Medicine Workshop 2015: Exploring opportunities for radiation oncology in the era of big data”. In: *International Journal of Radiation Oncology• Biology• Physics* 95.3 (2016), pp. 873–879.

[4] Rishabh Kapoor et al. “Automated data abstraction for quality surveillance and outcome assessment in radiation oncology”. In: *Journal of Applied Clinical Medical Physics* 22.7 (2021), pp. 177–187. DOI: https://doi.org/10.1002/acm2.13308. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/acm2.13308. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/acm2.13308.

[5] Stuart J Miller et al. “Multi-Modal Classification Using Images and Text”. In: *SMU Data Science Review* 3.3 (2020), p. 6.

[6] Issam El Naqa, Ruijiang Li, and Martin J Murphy. *Machine learning in radiation oncology: theory and applications*. Springer, 2015.

[7] John Kang et al. “Machine learning approaches for predicting radiation therapy outcomes: a clinician’s perspective”. In: *International Journal of Radiation Oncology* Biology* Physics* 93.5 (2015), pp. 1127–1135.

[8] Priyankar Bose et al. “Deep Neural Network Models to Automate Incident Triage in the Radiation Oncology Incident Learning System”. In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. doi: 10.1145/3459930.3469518, Art No.: 51. New York, NY, USA: Association for Computing Machinery, 2021. ISBN: 9781450384506. URL: `https://doi.org/10.1145/3459930.3469518`.

[9] Kory Kreimeyer et al. “Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review”. In: *Journal of Biomedical Informatics* 73 (2017). doi: https://doi.org/10.1016/j.jbi.2017.07.012, pp. 14–29. ISSN: 1532-0464. URL: `https://www.sciencedirect.com/science/article/pii/S1532046417301685`.

[10] Priyankar Bose, Satyaki Roy, and Preetam Ghosh. “A Comparative NLP-Based Study on the Current Trends and Future Directions in COVID-19 Research”. In: *IEEE Access* 9 (2021). doi: 10.1109/ACCESS.2021.3082108, pp. 78341–78355.

[11] Darshini Mahendran and Bridget T. McInnes. *Extracting Adverse Drug Events from Clinical Notes*. 2021. arXiv: `2104.10791 [cs.CL]`.

[12] Priyankar Bose et al. “A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts”. In: *Applied Sciences* 11.18 (2021). doi: 10.3390/app11188319, Art No.: 8319. ISSN: 2076-3417. URL: `https://www.mdpi.com/2076-3417/11/18/8319`.

[13] D Rhee et al. "TG263-Net: A deep learning model for organs-at-risk nomenclature standardization". In: *MEDICAL PHYSICS*. Vol. 46. 6. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA. 2019, E263–E263.

[14] Qiming Yang et al. "A Novel Deep Learning Framework for Standardizing the Label of OARs in CT". In: *Artificial Intelligence in Radiation Therapy*. Ed. by Dan Nguyen, Lei Xing, and Steve Jiang. Cham: Springer International Publishing, 2019, pp. 52–60.

[15] Dan Kalman. "A singularly valuable decomposition: the SVD of a matrix". In: *The college mathematics journal* 27.1 (1996), pp. 2–23.

[16] William C Sleeman IV et al. "A Machine Learning method for relabeling arbitrary DICOM structure sets to TG-263 defined labels". In: *Journal of Biomedical Informatics* 109 (2020), p. 103527.

[17] Khajamoinuddin Syed et al. "Integrated natural language processing and machine learning models for standardizing radiotherapy structure names". In: *Healthcare*. Vol. 8. 2. Multidisciplinary Digital Publishing Institute. 2020, p. 120.

[18] Khajamoinuddin Syed et al. "Multi-View Data Integration Methods for Radiotherapy Structure Name Standardization". In: *Cancers* 13.8 (2021). doi: 10.3390/cancers13081796, Art No.: 1796. issn: 2072-6694. url: `https://www.mdpi.com/2072-6694/13/8/1796`.

[19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[20] P Bose et al. "Integrated Structure Name Mapping with CNN". In: *Medical Physics*. Vol. 48. 6. Wiley 111 River St, Hoboken 07030-5774, NJ USA. 2021.

[21] W Sleeman et al. “Using CNNs to Extract Standard Structure Names While Learning Radiomic Features”. In: *Medical Physics*. Vol. 48. 6. Wiley 111 River St, Hoboken 07030-5774, NJ USA. 2021.

[22] Bingyin Hu, Anqi Lin, and Catherine L. Brinson. “ChemProps: A RESTful API enabled database for composite polymer name standardization”. In: *Journal of Cheminformatics* 13 (1 2021), p. 22. ISSN: 1758-2946. DOI: 10.1186/s13321-021-00502-6. URL: https://doi.org/10.1186/s13321-021-00502-6.

[23] Christian Jamtheim Gustafsson et al. “Deep learning-based classification and structure name standardization for organ at risk and target delineations in prostate cancer radiotherapy”. In: *Journal of Applied Clinical Medical Physics* 22.12 (2021), pp. 51–63. DOI: https://doi.org/10.1002/acm2.13446. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/acm2.13446. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/acm2.13446.

[24] Michael Lempart et al. “Deep learning-based classification of organs at risk and delineation guideline in pelvic cancer radiation therapy”. In: *Journal of Applied Clinical Medical Physics* n/a.n/a (), e14022. DOI: https://doi.org/10.1002/acm2.14022. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/acm2.14022. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/acm2.14022.

[25] Christian Jamtheim Gustafsson et al. “Deep learning-based classification and structure name standardization for organ at risk and target delineations in prostate cancer radiotherapy”. In: *Journal of Applied Clinical Medical Physics* 22 (2021), pp. 51–63.

[26] Michael Lempart et al. "Deep learning-based classification of organs at risk and delineation guideline in pelvic cancer radiation therapy." In: *Journal of applied clinical medical physics* (2023), e14022.

[27] Ali Haidar et al. "Standardising Breast Radiotherapy Structure Naming Conventions: A Machine Learning Approach". In: *Cancers* 15.3 (2023), p. 564. DOI: https://doi.org/10.3390/cancers15030564.

[28] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. *Self-Supervised Multimodal Learning: A Survey.* 2023. arXiv: 2304.01008 [cs.LG].

[29] Shervin Minaee et al. "Image Segmentation Using Deep Learning: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.7 (2022), pp. 3523–3542. DOI: 10.1109/TPAMI.2021.3059968.

[30] Fahad Lateef and Yassine Ruichek. "Survey on semantic segmentation using deep learning techniques". In: *Neurocomputing* 338 (2019). doi: https://doi.org/10.1016/j.neuc pp. 321–348. ISSN: 0925-2312. URL: https://www.sciencedirect.com/science/article/pii/S092523121930181X.

[31] Ali A. Alani, Georgina Cosma, and Aboozar Taherkhani. "Classifying Imbalanced Multi-modal Sensor Data for Human Activity Recognition in a Smart Home using Deep Learning". In: *2020 International Joint Conference on Neural Networks (IJCNN).* 2020, pp. 1–8. DOI: 10.1109/IJCNN48605.2020.9207697.

[32] Tsung-Yi Lin et al. "Microsoft COCO: Common Objects in Context". In: *Computer Vision – ECCV 2014.* Ed. by David Fleet et al. Cham: Springer International Publishing, 2014, pp. 740–755.

[33] Mengye Ren, Ryan Kiros, and Richard Zemel. *Exploring Models and Data for Image Question Answering.* doi: 10.48550/ARXIV.1505.02074. 2015. URL: `https://arxiv.org/abs/1505.02074`.

[34] Licheng Yu et al. “Modeling Context in Referring Expressions”. In: *Computer Vision – ECCV 2016.* Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 69–85. ISBN: 978-3-319-46475-6.

[35] Mateusz Malinowski and Mario Fritz. “A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input”. In: *Advances in Neural Information Processing Systems.* Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014. URL: `https://proceedings.neurips.cc/paper/2014/file/d516b13671a4179d9b7b458a6ebdeb92-Paper.pdf`.

[36] Ranjay Krishna et al. “Visual genome: Connecting language and vision using crowdsourced dense image annotations”. In: *International journal of computer vision* 123.1 (2017). doi: https://doi.org/10.1007/s11263-016-0981-7, pp. 32–73.

[37] Yuke Zhu et al. “Visual7W: Grounded Question Answering in Images”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* June 2016.

[38] Alexei Yavlinsky, Edward Schofield, and Stefan Rüger. “Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation”. In: *Image and Video Retrieval.* Ed. by Wee-Kheng Leow et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 507–517.

[39] Priyankar Bose, Pratip Rana, and Preetam Ghosh. “Attention-based Multimodal Deep Learning on Vision-Language Data: Models, Datasets, Tasks,

Evaluation Metrics and Applications". In: *IEEE Access* (2023), pp. 1–1. DOI: `10.1109/ACCESS.2023.3299877`.

[40] Xiaocui Yang et al. "Image-Text Multimodal Emotion Classification via Multi-View Attentional Network". In: *IEEE Transactions on Multimedia* 23 (2021). doi: 10.1109/TMM.2020.3035277, pp. 4014–4026.

[41] Fan Yang et al. "Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification". In: *Proceedings of the Third Workshop on Abusive Language Online*. doi: 10.18653/v1/W19-3502. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 11–18. URL: `https://aclanthology.org/W19-3502`.

[42] Hengcan Shi et al. "Key-Word-Aware Network for Referring Expression Image Segmentation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018.

[43] Sahar Kazemzadeh et al. "Referitgame: Referring to objects in photographs of natural scenes". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 787–798.

[44] Junhua Mao et al. "Generation and Comprehension of Unambiguous Object Descriptions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[45] Stanislaw Antol et al. "VQA: Visual Question Answering". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015.

[46] Yash Goyal et al. "Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[47] Abhishek Das et al. “Visual Dialog”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[48] Rowan Zellers et al. “From Recognition to Cognition: Visual Commonsense Reasoning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[49] Asma Ben Abacha et al. “VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019.” In: *CLEF (Working Notes)* 2.6 (2019).

[50] Fenglin Liu et al. “DiMBERT: Learning Vision-Language Grounded Representations with Disentangled Multimodal-Attention”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16.1 (2021), pp. 1–19.

[51] Bryan A. Plummer et al. “Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015.

[52] Zhong Ji et al. “SMAN: Stacked Multimodal Attention Network for Cross-Modal Image–Text Retrieval”. In: *IEEE Transactions on Cybernetics* 52.2 (2022). doi: 10.1109/TCYB.2020.2985716, pp. 1086–1097.

[53] Atsushi Shimada et al. “Kitchen Scene Context Based Gesture Recognition: A Contest in ICPR2012”. In: *Advances in Depth Image Analysis and Applications*. Ed. by Xiaoyi Jiang et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 168–185. isbn: 978-3-642-40303-3.

[54] Chunxiao Liu et al. “Focus Your Attention: A Focal Attention for Multimodal Learning”. In: *IEEE Transactions on Multimedia* 24 (2022). doi: 10.1109/TMM.2020.3046855, pp. 103–115.

[55] Yifei Zhang et al. “Deep multimodal fusion for semantic image segmentation: A survey”. In: *Image and Vision Computing* 105 (2021). doi: https://doi.org/10.1016/j.imavis.2 p. 104042. ISSN: 0262-8856. URL: `https://www.sciencedirect.com/science/article/pii/S0262885620301748`.

[56] Tao Zhou et al. “RGB-D salient object detection: A survey”. In: *Computational Visual Media* 7.1 (Jan. 2021). doi: 10.1007/s41095-020-0199-z, pp. 37–69. URL: `https://doi.org/10.1007%5C%2Fs41095-020-0199-z`.

[57] P Bose et al. “Standardizing Radiotherapy Structure Names with Multimodal Data: Deep Learning Approach”. In: *MEDICAL PHYSICS*. Vol. 49. 6. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA. 2022, E654–E654.

[58] Hermann Blum et al. “Modular Sensor Fusion for Semantic Segmentation”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. doi: 10.1109/IROS.2018.8593786. 2018, pp. 3670–3677.

[59] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. doi: 10.48550/ARXIV.2103.00020. 2021. URL: `https://arxiv.org/abs/2103.00020`.

[60] Ankit Kumar et al. “Ask Me Anything: Dynamic Memory Networks for Natural Language Processing”. In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1378–1387. URL: `https://proceedings.mlr.press/v48/kumar16.html`.

[61] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. doi: 10.48550/ARXIV.1810.04805. 2018. URL: `https://arxiv.org/abs/1810.04805`.

[62] Peng Xu, Xiatian Zhu, and David A. Clifton. *Multimodal Learning with Transformers: A Survey*. doi: 10.48550/ARXIV.2206.06488. 2022. URL: https://arxiv.org/abs/2206.06488.

[63] Franco Scarselli et al. "The Graph Neural Network Model". In: *IEEE Transactions on Neural Networks* 20.1 (2009). doi: 10.1109/TNN.2008.2005605, pp. 61–80.

[64] Jiayi Chen and Aidong Zhang. "Hgmf: heterogeneous graph-based fusion for multimodal data with incompleteness". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery. doi: 10.1145/3394486.3403182. New York, New York, USA, 2020, pp. 1295–1305. URL: https://doi.org/10.1145/3394486.3403182.

[65] Vikas Yadav and Steven Bethard. "A survey on recent advances in named entity recognition from deep learning models". In: *arXiv preprint arXiv:1910.11470* (2019).

[66] Nitish Srivastava and Ruslan Salakhutdinov. "Learning representations for multimodal data with deep belief nets". In.

[67] Kuan Liu et al. *Learn to Combine Modalities in Multimodal Deep Learning*. 2018. arXiv: 1805.11730 [stat.ML].

[68] Jun Shi et al. "Multimodal Neuroimaging Feature Learning With Multimodal Stacked Deep Polynomial Networks for Diagnosis of Alzheimer's Disease". In: *IEEE Journal of Biomedical and Health Informatics* 22.1 (2018). doi: 10.1109/JBHI.2017.2655720, pp. 173–183.

[69] Valentin Radu et al. “Multimodal Deep Learning for Activity and Context Recognition”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1.4 (Jan. 2018). doi: 10.1145/3161174, Art. No.: 157. URL: https://doi.org/10.1145/3161174.

[70] Jiawen Yao et al. “Deep Correlational Learning for Survival Prediction from Multi-modality Data”. In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017.* Ed. by Maxime Descoteaux et al. Cham: Springer International Publishing, 2017, pp. 406–414. ISBN: 978-3-319-66185-8.

[71] Danfeng Hong et al. “More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 59.5 (2021). doi: 10.1109/TGRS.2020.3016820, pp. 4340–4354.

[72] Xin Yang et al. “Bi-Modality Medical Image Synthesis Using Semi-Supervised Sequential Generative Adversarial Networks”. In: *IEEE Journal of Biomedical and Health Informatics* 24.3 (2020). doi: 10.1109/JBHI.2019.2922986, pp. 855–865.

[73] Peicheng Wu and Qing Chang. “Brain Tumor Segmentation on Multimodal 3D-MRI using Deep Learning Method”. In: *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI).* doi: 10.1109/CISP-BMEI51763.2020.9263614. 2020, pp. 635–639.

[74] Lyujian Lu et al. “Predicting progressions of cognitive outcomes via high-order multi-modal multi-task feature learning”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018).* doi: 10.1109/ISBI.2018.8363635. 2018, pp. 545–548.

[75] Michael Hagan et al. “VA-Radiation Oncology Quality Surveillance Program”. In: *International Journal of Radiation Oncology* Biology* Physics* 106.3 (2020), pp. 639–647.

[76] Jinhyuk Lee et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[77] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[78] Athanasios Voulodimos et al. “Deep learning for computer vision: A brief review”. In: *Computational intelligence and neuroscience* 2018 (2018).

[79] Shui-Hua Wang and Yu-Dong Zhang. “DenseNet-201-Based Deep Neural Network with Composite Learning Factor and Precomputation for Multiple Sclerosis Classification”. In: *ACM Trans. Multimedia Comput. Commun. Appl.* 16.2s (June 2020). ISSN: 1551-6857. DOI: 10.1145/3341095. URL: https://doi.org/10.1145/3341095.

[80] Brett Koonce. “SqueezeNet”. In: *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization.* Berkeley, CA: Apress, 2021, pp. 73–85. ISBN: 978-1-4842-6168-2. DOI: 10.1007/978-1-4842-6168-2_7. URL: https://doi.org/10.1007/978-1-4842-6168-2_7.

[81] Huiyi Li. “Image semantic segmentation method based on GAN network and ENet model”. In: *The Journal of Engineering* 2021.10 (2021), pp. 594–604. DOI: https://doi.org/10.1049/tje2.12067. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/tje2.12067. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/tje2.12067.

[82] Salman Khan et al. “Transformers in Vision: A Survey”. In: *ACM Comput. Surv.* 54.10s (Sept. 2022). ISSN: 0360-0300. DOI: 10.1145/3505244. URL: https://doi.org/10.1145/3505244.

[83] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[84] Suyash Lakhotia and Xavier Bresson. “An Experimental Comparison of Text Classification Techniques”. In: *2018 International Conference on Cyberworlds (CW)*. IEEE. 2018, pp. 58–65.

[85] Yoon Kim. “Convolutional Neural Networks for Sentence Classification”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. DOI: 10.3115/v1/D14-1181. URL: https://www.aclweb.org/anthology/D14-1181.

[86] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[87] Ahmed Tealab. “Time series forecasting using artificial neural networks methodologies: A systematic review”. In: *Future Computing and Informatics Journal* 3.2 (2018), pp. 334–340. ISSN: 2314-7288. DOI: https://doi.org/10.1016/j.fcij.2018.10.003. URL: https://www.sciencedirect.com/science/article/pii/S2314728817300715.

[88] Jin Xu et al. “Reluplex made more practical: Leaky ReLU”. In: *2020 IEEE Symposium on Computers and Communications (ISCC)*. 2020, pp. 1–7. DOI: 10.1109/ISCC50000.2020.9219587.

[89] Yuanzhi Li and Yang Yuan. “Convergence Analysis of Two-layer Neural Networks with ReLU Activation”. In: NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 597–607. ISBN: 9781510860964.

[90] Steven Gold, Anand Rangarajan, et al. “Softmax to softassign: Neural network algorithms for combinatorial optimization”. In: *Journal of Artificial Neural Networks* 2.4 (1996), pp. 381–399.

[91] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2014. DOI: 10.48550/ARXIV.1409.1556. URL: https://arxiv.org/abs/1409.1556.

[92] Md Foysal Haque, Hye-Youn Lim, and Dae-Seong Kang. “Object Detection Based on VGG with ResNet Network”. In: *2019 International Conference on Electronics, Information, and Communication (ICEIC)*. 2019, pp. 1–3. DOI: 10.23919/ELINFOCOM.2019.8706476.

[93] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. DOI: 10.48550/ARXIV.1512.03385. URL: https://arxiv.org/abs/1512.03385.

[94] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[95] Christian Szegedy et al. “Going Deeper With Convolutions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.

[96] Charley Gros, Andreanne Lemay, and Julien Cohen-Adad. "SoftSeg: Advantages of soft versus binary training for image segmentation". In: *Medical Image Analysis* 71 (2021), p. 102038. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2021.102038. URL: https://www.sciencedirect.com/science/article/pii/S1361841521000840.

[97] João Lourenço Silva and Arlindo L. Oliveira. *Using Soft Labels to Model Uncertainty in Medical Image Segmentation*. 2021. arXiv: 2109.12622 [cs.CV].

[98] Yabo Fu et al. "A review of deep learning based methods for medical image multi-organ segmentation". In: *Physica Medica* 85 (2021), pp. 107–122. ISSN: 1120-1797. DOI: https://doi.org/10.1016/j.ejmp.2021.05.003. URL: https://www.sciencedirect.com/science/article/pii/S1120179721001848.

[99] Getao Du et al. "Medical image segmentation based on u-net: A review." In: *Journal of Imaging Science & Technology* 64.2 (2020).

[100] Phillip Chlap et al. "A review of medical image data augmentation techniques for deep learning applications". In: *Journal of Medical Imaging and Radiation Oncology* 65.5 (2021), pp. 545–563. DOI: https://doi.org/10.1111/1754-9485.13261. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1754-9485.13261. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1754-9485.13261.

[101] Balamurali Murugesan et al. "Psi-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation". In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019, pp. 7223–7226. DOI: 10.1109/EMBC.2019.8857339.

[102] Sihang Zhou et al. “High-Resolution Encoder–Decoder Networks for Low-Contrast Medical Image Segmentation”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 461–475. DOI: `10.1109/TIP.2019.2919937`.

VITA

Priyankar Bose received the B.Tech degree in Electronics and Telecommunication Engineering from KIIT University, India in 2018. He works on a variety of problems in the field of Biological Data Sciences. His research interests include data/text mining, machine learning, NLP and biological data modeling and simulations.