A Study of Spatio-Temporal Learning Approaches Using Echocardiograms for Risk Assessment of Thoracic Aortic Aneurysms

Sandi Alakhras

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

January 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Sandi Alakhras**

Entitled: **A Study of Spatio-Temporal Learning Approaches Using Echocardiograms for Risk Assessment of Thoracic Aortic Aneurysms**

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair and Examiner
Dr. Yiming Xiao

_____ Examiner
Dr. Mirco Ravanelli

_____ Supervisor
Dr. Thomas Fevens

Approved by _____
Lata Narayanan, Chair
Department of Computer Science and Software Engineering

_____ 2023

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

A Study of Spatio-Temporal Learning Approaches Using Echocardiograms for Risk Assessment of Thoracic Aortic Aneurysms

Sandi Alakhras

Aortic dissection and rupture are fatal complications that happen when the aortic tissue's integrity is compromised, leading to fatal consequences. Once an aortic dissection takes place, 41% of patients do not even make it to the hospital. Unfortunately, the diagnostic outlook is not much brighter. It is estimated that 40% of patients presenting with aortic dissection do not meet the current diagnostic criteria.

This thesis aims to assess the risk levels of thoracic aortic aneurysms' dissection and rupture from patients' echocardiograms. To do this, we study the effects of spatial and temporal learning of the heart's movement in the echocardiograms. We investigate the pure visual learning from still 2D images extracted from the echocardiogram's sequence, then assess the temporal learning across frames in the echocardiogram video by incorporating 3D convolutions over the whole sequence, and in terms of aggregating the visually learned content of each frame in the echocardiogram over the sequence length. We also experiment with implementing a visual attention mechanism to filter out the visual context. Finally, we study the effect of adding a tabular data learning stream to our architecture that learns from the patient's tabular data information and incorporates it into the best-performing model. The results of this thesis - although not conclusive- suggest that temporal dependencies are present between echocardiogram frames throughout the video, which points out the diagnostic importance of analyzing the movement of the beating heart tissue through time.

Acknowledgments

I would like to take this opportunity to extend my gratitude to Dr.Thomas Fevens for not only supervising and guiding me through this work but also for offering his advice and wisdom to my personal and career growth. In addition, this work would not have been possible without the collaboration that was offered to us by the Mcgill Richard Leask lab, so I would like to thank all the lab members for their continuous efforts and support throughout this project.

إلى ابي و امي :

إلى من جرعوا الكأس فارغاً ليسقوني، إلى من حصدوا الأشواك عن دربي ليمهدوا لي طريق العلم

إلى سرين ويارا وتالا :

إلى من شاركوني كل بسمة و كل دمعة، إلى أول اصدقائي

To Charbel, I am lucky to have someone by my side who would push me like you did when the pool of motivation and optimism was low. Finally, I would like to thank my extended family here in Canada and all the friends I have made here who have eased my immigration journey and made me feel safe and loved in a place very far away from home.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

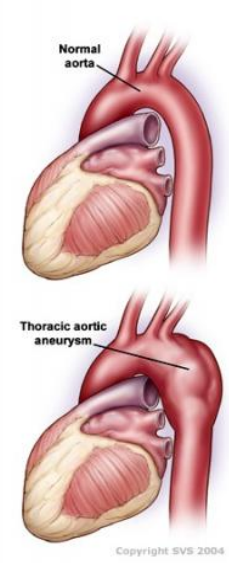
1.1 Introduction and Motivation

An aneurysm is a balloon-like bulge in the wall of blood vessels or arteries. It begins as a weak spot in the blood vessel wall, which balloons out of shape over time by the force of the pumping blood [36]. A representation of an aortic aneurysm can be seen in figure 1.1a. As a result of aortic aneurysms, and due to the constant force exerted by the pumping blood, the mechanical properties of the aortic wall weaken over time thus making the valve prone to dissections and ruptures.

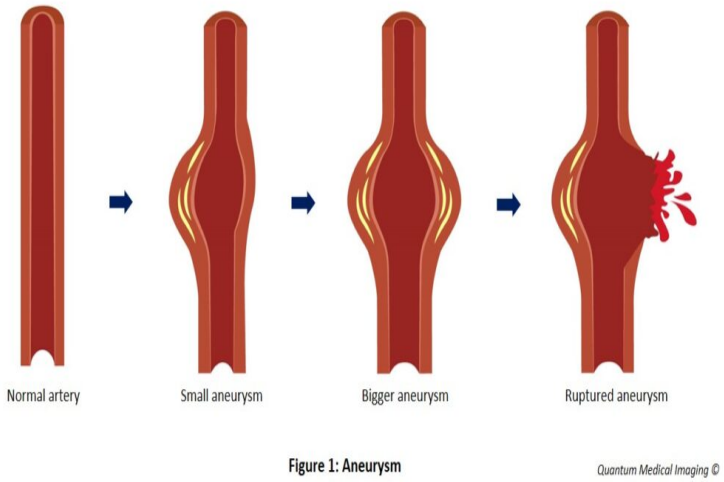
Aortic dissection is a separation between the inner and middle layers of the aortic artery wall[22]. An aortic dissection happens when a tear occurs in the inner layer of the aortic valve and blood floods through the tear, causing the inner and middle layers of the aorta to separate [22]. As a result, normal blood flow to the body might be slowed, hampered, or may result in a complete rupture of the aortic valve, bursting blood through the hole into the surrounding body cavity.

Aortic dissection is a life-threatening condition, as reports suggest that only 41% of patients experiencing a dissection make it to the hospital [27]. For those who actually do make it to the hospital, the mortality rates for the non-elective ascending aortic replacement surgeries are 15%-24% [42] [53]. With acute dissection and rupture being fatal complications, prevention is only possible with surgical intervention by correctly diagnosing and assessing the risk before these deadly complications happen.

Currently, most societies (including the American Heart Association and endorsed by the North



(a) Normal versus aneurysmal aortic tissue [36]



(b) An illustration of an aneurysm rupture[25]

American Society for Cardiovascular Imaging) utilize the maximum aortic diameter as the selective criterion of elective surgical prevention. [22]. Alarmingly, approximately 40% of patients who present with dissection have aortic diameters below diagnostic criteria [39]. We, therefore, perceive the urgent need for novel criteria to identify and diagnose patients at risk of such fatal complications.

With ultrasound imaging being one of the safest, cheapest and most available medical imaging modalities [35], its deep learning applications and research are critical. Echocardiography (ECHO) – is the use of ultrasound to examine the heart. It is a safe and non-invasive technique [35]. In this thesis, we aspire to solve the diagnostic problem mentioned above by assessing the risk level of the occurrence of an aortic dissection from echocardiogram images.

1.2 Visual Sequence Learning

Echocardiography is one of many medical imaging modalities that produce sequential imaging or video-like output. Therefore, since echocardiograms are sequence-like images, we focused on leveraging and exploring different types of learning that would process visual sequence learning.

However, we find that the trends in deep learning in medical applications, when it comes to dealing with these sequences, tend to rely on processing their image frames as separate, independent

dataset entities rather than using the sequence as a whole, and therefore end up losing all temporal context and dependencies between those frames.

While applications of image recognition, classification, segmentation and detection have seen state-of-the-art research and application in the last decade, little attention has been shown to similar video applications. Some reasons might be attributed to the fact that video datasets are more difficult to attain, or due to the fact that separate frames 2D image convolutional models are potent enough to be achieving remarkable close performance when compared to specific video-based models.

However, when it comes to problems relating to heart-related diagnostics, the behaviour and the movement of the beating heart through the full cardiac cycle (from one beat to another) is a medically major visual diagnostic factor.

A common source of inspiration for Machine learning systems comes from humans' natural way of learning or observing, or from types of learning observed in nature. Take, for example, reinforcement, transfer, and multi-task learning, which were all derived from natural systems humans use to learn and perform tasks. The task of video classification follows a similar logic. Yes, it is evident that humans look mainly at the content of individual frames of the video to identify its content, but they also look at the change per frame of the video to process the change of visual context through time.

1.3 My Contributions

Through the course of this research, we can summarize the contributions and objectives of this thesis into two folds. Firstly, research the best sequential data representation to capture both the spatial features of still frames and the temporal data from the movement between frames. Secondly, explore a hybrid model architecture to combine patients' tabular information to the video classification task.

(1) We investigate the best visual learning convolutional architecture in terms of different data input forms and dimensions.

(2) We then follow to present different approaches for capturing the temporal aspect of the echocardiograms in terms of capturing sequence dependencies between the frames through

a third dimension of temporal Convolutions, Recurrent Neural Networks (RNN) or pooling over the time steps.

(3) We finally add a dense stream of multilayer perceptron to our architecture to process the patients' tabular information and sum everything up to a novel end-to-end trainable ensemble deep learning model which processes the frames of the mini-clips separately, then pools over them in time and combines a separate stream of Tabular patient information with the outputs.

1.4 Thesis Outline

The remainder of this research is organized as follows:
In Chapter 2, we present the literature review and background work related to the concepts explored in this thesis including the biological aspect of heart imaging, the bases of visual learning and sequential learning that are going to be used throughout this thesis, and a brief introduction into the common practices in video classification.

In Chapter 3, We start by describing our echocardiogram dataset's input and labels, and its preprocessing steps. We then present and analyze the different sequential visual frameworks along with their respective experiments and results on our echocardiogram dataset.

In Chapter 4, we introduce our complimentary patients' meta-data dataset. We then present our hybrid ensemble model, which processes the patients' echocardiograms as well as their meta-data, and displays its respective experiments and results.

In Chapter 5, we conclude by summarizing and analyzing our experiments' results. We also provide some insight into the possibility of future work.

Chapter 2

Background

In this chapter, we will cover the related background information and literature which lead to the inspiration of this thesis. First, we will start with the general modalities of medical and cardiac imaging. After that, we will dive closer into the visual and sequential aspects of deep learning followed by shedding some attention on visual sequence learning. We finally will cover some characteristics of tabular data processing and its applications in deep learning.

2.1 Cardiac Imaging Modalities

As the underlying technologies progressed, the field of medical imaging took a turn into new dimensions and the dream of imaging a beating heart became one of the routinely performed practices. Cardiac imaging, also known as cardiovascular imaging, has become one of the most essential practices in modern medicine not only for its cardiac-related diagnostic and prognostic purposes but also for its importance in the guidance of other invasive procedures [33]. With the advancement of the field of cardiac imaging in the past decades, different imaging modalities have appeared that not only allowed anatomic evaluation of the heart but also functional evaluation. These imaging modalities include but are not limited to cardiovascular computed tomography (CCT), cardiovascular Magnetic resonance imaging (MRI), and Echocardiography.

Computed tomography

A computed tomography (CT) scan is a medical imaging modality that is composed of a series of X-ray images taken from various angles around the target organ. It uses computer processing (tomographic reconstruction algorithms) to create cross-sectional (tomographic) images or slices of the body. Since its development in the 1970s, CT imaging has proven to be one of the most essential and versatile medical imaging techniques [33].

Cardio Vascular CT, which can be thought of as a special application of CT, was first introduced as an electron-beam CT (EBCT) which, although is not a very practical method for imaging a moving organ, remains one of the most used methods for coronary calcification diagnosis. Another critical application of cardiac CT is coronary CT angiography (CCTA) which is a cardiac imaging test that determines if the coronary arteries have plaque buildup, causing them to narrow and decrease the blood supply to the heart. This high negative predictive value and specificity test have become an essential assessment in accident and emergency departments to rule out significant coronary artery disease in patients with chest pain [33]. An example of CCTA is shown in 2.1.

Figure 2.1: A coronary CT angiography [33]

Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a non-invasive medical imaging modality that is used to generate three-dimensional detailed anatomical images of organs. It is comprised of a sophisticated system of strong magnetic fields, magnetic field gradients and radio waves in a manner that excites

and detects the change in the direction of the rotational axis of protons found in the water that make up living tissues [6]. MRI is commonly used in hospitals around the world for diagnostic and prognostic purposes. In addition, MRI imaging provides better contrast in soft-tissue images when compared to CT imaging.

Cardiovascular Magnetic Resonance (CMR) is a cardiac application of MRI targeted to specifically image the heart organ. It mostly gained popularity in hospitals with the introduction of ECG-gated imaging in 1983 which greatly improved the field of dynamic cardiac imaging. CMR can be used to assess the structural and functional properties of the heart and is used in the diagnosis of several cardiac diseases such as myocardial ischemia and viability, cardiomyopathy, myocarditis, iron overload, vascular diseases, and congenital heart disease [37]. an Example of a CMR can be seen in figure 2.2.

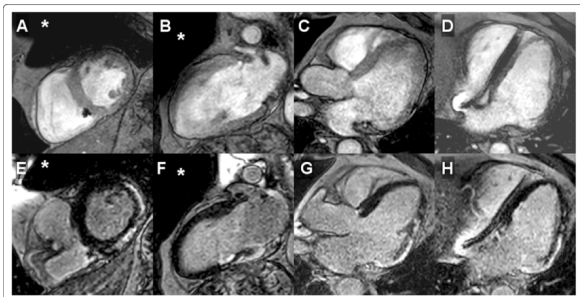
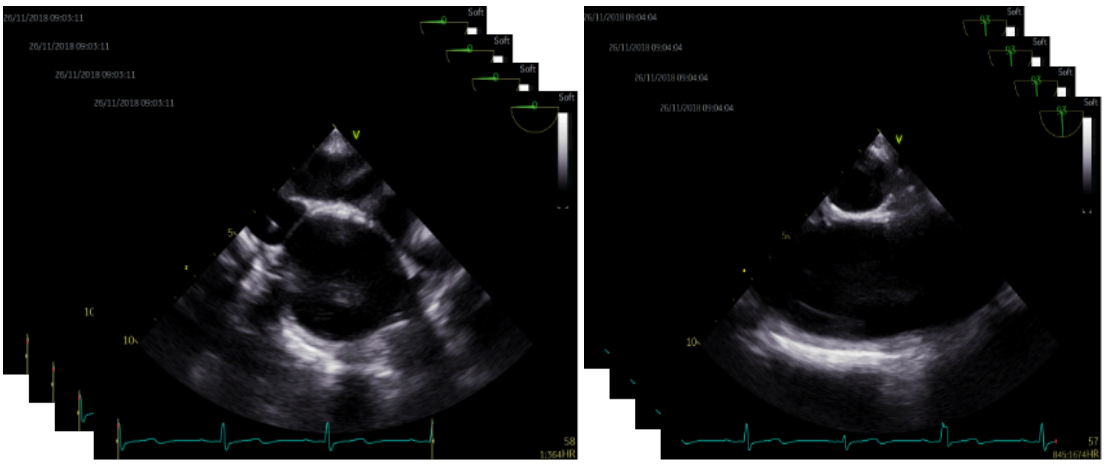


Figure 2.2: CMR imaging in a single-chamber [34]

Echocardiograms

Echocardiography is a non-invasive medical imaging technique and one of the most frequently used cardiovascular diagnostic practices after electrocardiography [35]. An echocardiogram is an application of the Doppler ultrasound tool which works by emitting sound waves that bounce off the heart's chambers and structures, therefore, creating a detailed but foggy image of the vascular cardiac structure. Generally speaking, the two most performed echocardiograms are transthoracic echocardiograms (TTE) and transesophageal echocardiograms (TEE) [35]. TTE remains the cornerstone of echocardiogram diagnosis due to its ease of operation that can take place in many settings including clinics, inpatient rooms and emergency situations and does not usually take more than 30 min to perform. A typical TTE Examination involves externally applying an echo probe perpendicularly to the patient's skin at various angles and positions to get the respective imaging of the heart

structure. While TTE has numerous indications such as Heart failure, Heart murmur, Congenital heart disease, Endocarditis and others it falls inherently limited in its image quality and diagnostic accuracy in certain cases. Its main limitation comes from the fact that the skin and body fat act as barrier mediums that absorb their share of the ultrasound waves before they reach the heart. In Contrast, in a TEE the imaging probe is inserted directly through the left atrium in the esophagus where the ultrasound waves have far fewer structures to penetrate thus offering a far superior imaging quality. An example of long and short-axis TEE images of the ascending thoracic aortic valve can be seen in figure 2.3.



(a) Short Axis ascending aortic valve TEE (b) Long Axis ascending aortic valve TEE

Figure 2.3: Example of input data

2.2 Visual Learning

2.2.1 Image Classification and Convolutional Neural Networks

Image classification is one of the most active research tasks in the world of deep learning and its challenges and applications have shaped the industry and led the innovation of computer vision research to the levels we know today[17]. Interest in the field sparks not only because of its countless applications but also from the perplexing fact that this essential vision task can be almost effortless to humans and other animals but constitutes the most challenging topic in the world of machine learning.

The task of classification in its non-data science terms relies on assigning categorical labels to input entities from a range of predefined labels. Image classification follows the same logic: given an image and a predefined set of classes, decide which class this image belongs to (e.g. classifying digits images, determining if a picture is a cat or dog).

One clear Example of the progression of computer vision and image classification specifically can be seen when examining the winning models of the ImageNet Large Scale Visual Recognition Challenge. Traditionally most machine learning algorithms used in the computer vision field relied on hand-crafted features and engineered filters to determine the relationship between structures in images algorithmically. However, with hardware units advancements, more specifically graphical processing units (GPU), came a drastic shift of improvement pace that established Convolutional neural networks' place in the world of deep learning.

RESNET

As trends in deep learning proved over time that deeper and deeper networks added expressiveness to the model, it also created the notorious problem of vanishing gradients that caused the model's performance to get saturated or even start degrading rapidly.

ResNet [21] was specifically designed to overcome the mentioned problem. ResNet introduced the novel idea of "identity shortcut connection" with its core idea of introducing residual blocks designed to let connections skip one or more layers as in figure 2.4. The authors of [21] hypothesize that if multiple stacked layers can fit a function F(x) then it is equivalent to assume that those layers can fit the residual function F(x) - x and then add the output to the identity mapping x. This not only added ease of computation but also helped the deep models' degradation problem by creating identity mapping shortcuts that prevent vanishing gradients down the deeper layers.

The remarkable results that ResNets and the liberty the authors chose of having the pre-trained weights available for public use have reshaped the industry of deep learning.

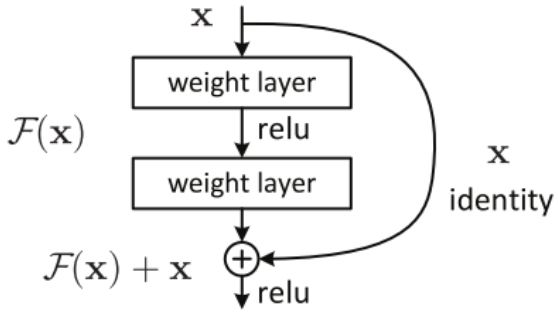


Figure 2.4: Residual Block [21]

2.3 Sequential Learning

2.3.1 Recurrent Neural Networks

Recurrent neural networks (RNNs) are a family of deep learning models that have evolved through the years to claim their standing in the sequential learning field [17] and have particularly gained popularity in the natural language field due to their ability to model variable length inputs. RNNs utilize an important form of parameter sharing in deep neural networks that allows them to extend and generalize the learning process through different positions across different lengths in the input sequence.

RNNs present the concept of ”memory” (referred to as state) that keeps track of past and future time steps in the input sequence by implementing a constant feed-forward operation between the input time steps. To form their memory through the sequence, they are expressed as a recursive function that applies a transformation A at every time step t that incorporates the input x and the state from the previous time-step t-1. The state (or hidden state) at time t can be seen below.

$$s_t = A(s_{t-1}, x_t, \theta) \tag{1}$$

Therefore, by sharing a function A, and parameters θ an RNN can be thought of as a network made up of multiple smaller copies of the same network, each passing its learned information to its successor. This chain representation that RNNs possess is the natural reason that relates them to sequence and text modelling.

As an unrolled network, when it comes to training, RNNs are trained as vanilla feed-forward

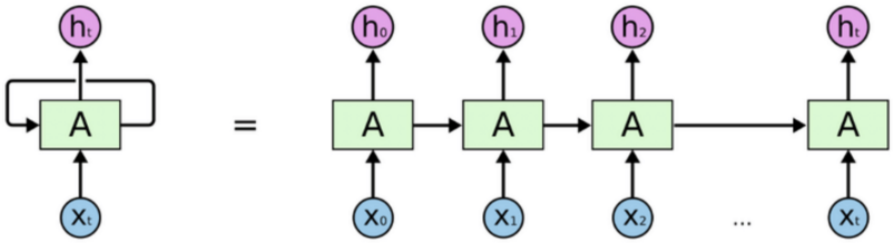


Figure 2.5: Un-Rolling RNNs [38]

neural networks where the gradients would be calculated through back-propagation and the transformation A adjusted per the optimizer of choice.

Long Short Term Memory

As the input sequences grew longer, RNNs displayed difficulty maintaining the learned features [5]. The main reason goes back to a problem labelled as vanishing gradients. As the gradients are backpropagated through the long sequence network time-steps t, the gradients tend to shrink, decreasing in value until they reach near-zero numbers. This causes the weights of the earlier time steps to remain practically the same. Similarly to how ResNet introduces skip connections between the earlier and deeper layers to mitigate the vanishing gradients problem, Long Short Term Memory (LSTM) networks present gated connections to overcome the long-term dependency in RNNs. LSTM design relies on the concept of cells that are characterized by forget, input and output gates. The forget gates choose whether the information coming from the previous time state is to be discarded or remembered. In the input gates, the cell updates its information to learn new patterns and then passes the updated information to the output gate which outputs the cell context to the succeeding time step. A similar approach has been implemented in other variations of gated recurrent networks (GRU Networks [9]) that cuts down on the complexity by only utilizing two gated connections.

2.4 Video Recognition and Classification

Video recognition and classification are one of the main cornerstones in computer vision problems and applications. However, it may be argued that the video recognition domain has not yet seen

the state-of-the-art uprising that the image classification world has seen during the last decade [49]. This domain oversight might be attributed to the perplexing fact that, until recently, image-based 2D CNNs operating on individual video frames could achieve results that are extremely close to those of state-of-the-art video classification models [49]. We note, however, that with the advancement of deep learning and specifically sequence learning, more and more modern research efforts have been made to model videos not just as a collection of images but as a visual sequence representation which produced remarkable results over traditional, individual frames 2D CNNs models.

Therefore, When it comes to the world of action recognition and video classification, the current approaches tend to analyze the visual input with either 2D convolutional filters over individual frames or aggregated frames of the whole video with pooling or recurrent analysis of the resulted feature maps,2 streams (visual and temporal) architecture inputs, 3D convolutional filters or, more recently, transformer architecture. We will be discussing the most popular architectures below.

2.4.1 3D Convolutional Based Models

R3D

Following the successful path that [30] achieved by evaluating different types of visual and temporal dimensions' connections, and the practical implementation presented in [48] of utilizing 3D Convolutional filters and pooling layers in the context of large-scale action recognition and classification, the field of 3D ConvNets started a state of rapid growth.

The authors of [19] proposed the R3D architecture which is based on the ResNets [21] model's structure to ease the training process of 3D Convolutional deep models that hold huge numbers of parameters. To achieve that, they utilized the same shortcut connection concepts presented in ResNets 2.4 to bypass the signal from one layer to another deeper layer down the architecture. A representation illustrating the R3D layers architecture can be seen in figure 3.6.

R(2+1)D

With the success 3D ConvNets achieved, problems regarding the huge number of parameters and complexity inherently followed. With that in mind, the authors of [49] presented R(2+1)D as a

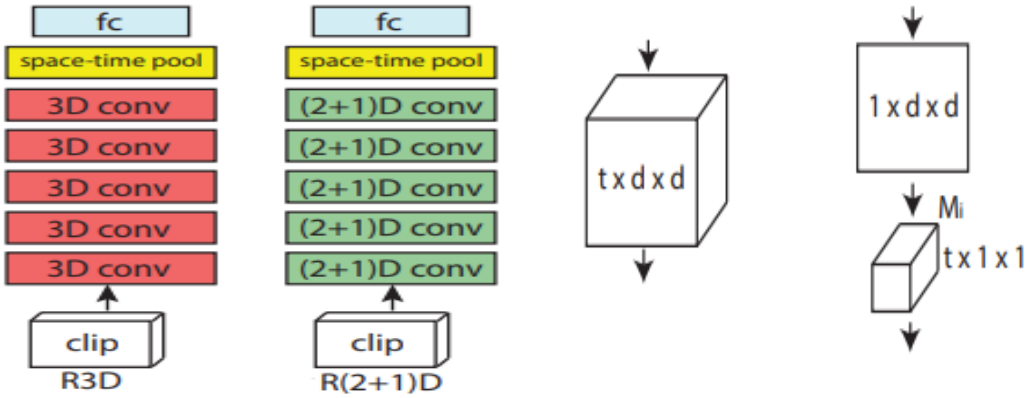
Layer Name	Architecture	
	18-layer	34-layer
conv1	$7 \times 7 \times 7, 64$, stride 1 (T), 2 (XY)	
conv2_x	$3 \times 3 \times 3$ max pool, stride 2 $\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 3$
conv3_x	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 4$
conv4_x	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 6$
conv5_x	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 3$
	average pool, 400-d fc, softmax	

Figure 2.6: R3D Network Architecture [19]

hybrid model which could utilize the depth that 3D convolutions possessed in processing temporal dimensions and the compactness of training that 2D and 1D convolutions present. Its architecture explicitly factorizes or decomposes the 3D convolution into 2 parts pertaining to a 2D spatial convolution which is followed by a 1D temporal convolution. A representation of this convolutional decomposition as well as a structural comparison between R(2+1)D and R3D can be seen in figure 2.7. We note that in part 2.7b, the 3D convolution (on the left) is implemented using a t × d × d filter size, where t denotes the temporal sequence length and d is the visual input's respective width and height. In the factorized version (on the right), the convolutional computation is split into a 2D convolution with a filter size of 1 × d × d followed by a 1D convolution of a t × 1 × 1 filter size. M_i represents the number of 2D filters and is adjusted so the number of parameters of the (2+1)D block matches that of the 3D block

By doing so, the authors achieved an extra non-linear operation (between the separated 2D and 1D convolutions) that renders the model more robust in representing complex non-linear patterns. In addition, according to the authors, the separation of the 2D visual and 1D temporal convolutions also

eases and improves the optimization process yielding considerable improvement over traditional 3D ConvNets(R3D) for interchangeable experimentation between the two.



(a) Similarities in terms of structure between the R3D and R(2+1)D architecture

(b) 3D convolutions(left) compared to decomposed 2D and 1D convolutions(right)

Figure 2.7: R(2+1)D Structure [49]

2.4.2 2D Convolutional based Models

Following the success that 2D CNNs displayed in image classification applications, many efforts went into attempting to leverage that success into the action recognition field. Research in this field aims at incorporating pre-trained 2D CNNs that have already proved their efficacy, and utilize them to classify images at the frame level. These can be mainly split into still frame classifications, aggregated frames visual extraction with temporal pooling or recurrent processing, and two-stream networks.

In the case of 2D still frame processing, the video is simply taken as a set of pictures and dataset entities are presented by an independent frame and its corresponding label. Although one might think that the network would fail to learn essential temporal patterns, [30] shows how powerful 2D CNNs can be even when operating on independent still frames of the video.

On the other hand, feature map aggregation (which will be discussed in detail in chapter 3), keeps the integrity of the video as a series of closely related images. In its applications, CNNs are trained on video frames (2D images) to get the extracted feature maps, and then perform temporal integration through a form higher dimensional feature encoding [16] [56] or recurrent processing of

the feature maps [10] [47] [4].

Another research path when stumbling upon 2D ConvNets that incorporate the temporal domain is the two-stream architectures. In this case, the network processes two different streams of input, the spatial stream input which consists of the series of static RGB coloured frame images and the temporal stream input being dense optical flow images extracted by applying filters to the corresponding image frame to indicate the motion information explicitly and then fusing the two streams together [46]. We note, however, that this method can only explicitly capture motion patterns between consecutive frames and thus fails to capture long-period temporal dependencies and has no considerations for frame order in the frame sequence [14].

2.5 Tabular Data in Deep Learning

Despite tabular data being the most abundant type of data that is easily acquirable in the real world, we find its application yet weakly explored in deep learning literature [28]. With the massive success that deep learning had achieved in complex data domains such as images, audio and language [17] recent interest has been growing to apply this success to tabular data. Such emerging research can be generally architecturally divided into differentiable decision trees [20] [40], Attention-based models [2][3], and modelling of multiplicative interactions [41].

The motivation for tabular data deep learning applications stems not only to seek higher performance in the separate tabular data application but also to incorporate tabular data learning as a pipeline for hybrid multi-modal learning problems. In such multi-modal models, tabular data learning is incorporated as a pipeline with other data input types such as images, audio, text or other input types that deep learning has proven effective for [18]. In these models, tabular data points are stored as heterogeneous scaled feature vectors, and the models are usually trained end-to-end via classic gradient back-propagation and optimization.

Unfortunately, unlike the established benchmarks seen in computer vision (ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[44]) or in natural language processing (General Language Understanding Evaluation (GLUE)[51]) there exists little comparison between the performances of tabular data deep learning models due to the fact that different papers assess their

models on different datasets. In addition, due to the absence of such benchmarks and the ease of access and implementation of traditional current state-of-the-art ”shallow” machine learning networks, models such as gradient boosted decision trees (GBDT)[13], XGBoost [8], and others have become the ”go to” architectures for popular tabular learning tasks while basic Multi-Layer Perceptrons (MLPs) remain the deep learning baseline models [18].

Chapter 3

Visual Data Description and Extraction

We will start this Chapter with a description of the dataset's inputs and labels along with their cleaning and preprocessing steps followed by the proposed frameworks, experiments, and results for visual data processing.

Going through this and the following chapters' experiments, we will adopt a narrative approach that would display the frameworks and architectures designed in the order we chronologically reasoned through them according to our previous experiments' results and established trends and practices in machine learning.

Each suggested framework is initiated by an introduction of the visual sequence learning methodology and its motivation followed by the corresponding experimental setup and its respective results.

3.1 Dataset

As mentioned in the introduction, this project is based on the dataset that was provided to us by McGill University's Leask Laboratory team and if it wasn't for their constant support and follow-ups, this work would have been possible.

3.1.1 Original Dataset Description

In collaboration with the Department of Chemical Engineering(McGill University, Montreal, Quebec, Canada), Department of Surgery of Royal Victoria Hospital(McGill University, Montreal,

Quebec, Canada), Division of Cardiology of Royal Victoria Hospital, (McGill University, Montreal, Quebec, Canada), Research Centre of Montreal Heart Institute(Montreal, Quebec, Canada) and in compliance with the Canadian tri-council policy statement on ethical conduct for research involving humans, informed consent for anonymous data collection was obtained from a cohort of 32 distinct patients undergoing elective aortic valve repair or replacement surgery.

According to the Leask lab's previous work, [12], the Collected information consisted of:

(1) Short axis Transesophageal Echocardiogram (TEE) Images: The TEE imaging probe was inserted through the esophagus to the level of the aorta in order to take an image of the maximum aortic valve diameter. At least 1 non-truncated heart cycle was captured in the imaging video. This process was operated using the GE Vivid 7 echocardiographic unit [15].

(2) Ex Vivo Physical Tensile Analysis: A specimen of the aortic ring was obtained for each patient immediately after resection, clipped for anatomic orientation, and stored in physiologic saline at 4°C until further processing and testing. The maximum aortic diameter was measured for each ring before sectioning four 1.5×1.5 cm^2 testing squares, equally distributed around the circumference of the aorta representing the 1-anterolateral wall, 2-posterolateral wall, and the 3-inner and 4-outer curvature. Five unique thickness measurements were taken for each testing square using a Mitutoyo Litematic VL-50A constant force digital micrometer (Mitutoyo Corp, Kanagawa, Japan). The testing squares were then connected to an EnduraTEC ELF 3200 planar biaxial tensile tester (Bose, Eden Prairie, Minn) using hooked 4-0 silk sutures in a 37°C bath of Ringer's lactate solution. The testing squares were oriented for equibiaxial stretching along their circumferential and longitudinal axes. Each sample was preconditioned for 7 cycles (ie, stretch and relaxation) followed by 3 cycles of data acquisition at a constant displacement rate of 0.1 mm/s in the range of 0% to 60% strain. Circumferential ex vivo energy loss and stiffness were calculated from the circumferential engineering stress-strain relation. Energy loss is the percentage of elastic energy needed to stretch the testing square that is dissipated when the tissue is relaxed. The physiologic interpretation of energy loss is the percent loss of elastic recoil energy in the tissue that is not returned to blood flow (ie, maintaining normal Windkessel function). Its physical definition is the ratio of the area

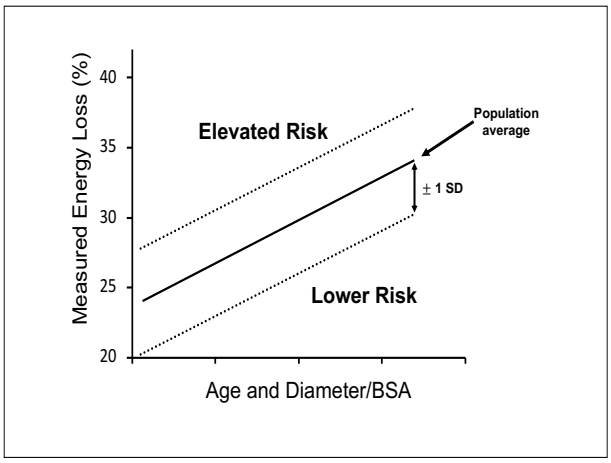


Figure 3.1: Leak's Lab Risk Label Margins[32]

between the loading and unloading curve over the area under the loading curve. Because aortic tissue has a nonlinear stress-strain curve, stiffness is defined as the slope of a line tangent to the stress-strain loading curve.

Labels

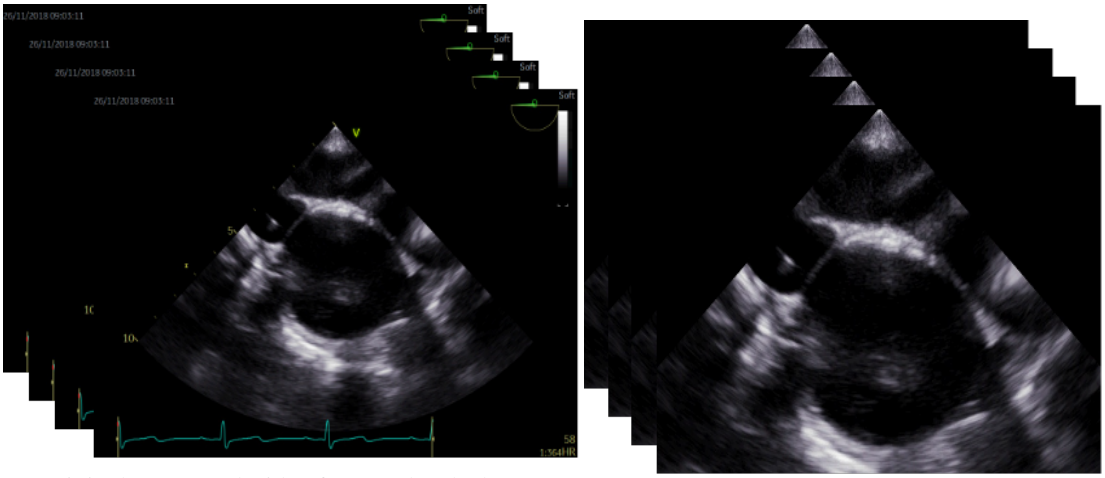
Regarding the ground truth used for labelling the input, categorical risk labels were provided to us by the Leask Lab which consisted of 3 discrete classes pertaining to low risk, average risk, and high risk. Each risk label is assigned to its distinct patient's echocardiogram video with the risk calculation being based on calculations performed by the Leask Lab based on the Ex Vivo Physical Tensile Energy Loss measurement (see number 2 above), age, diameter and BSA [32]. More specifically, the Leask lab plotted the patients' measured ex-vivo tensile energy loss value against their corresponding Age and Diameter/BSA and considered the patients with values that fall between +1 and -1 standard deviation(SD) from the population average to be labelled as average risk. All values below -1 SD were labelled as low risk and all values greater than +1 SD were labelled as high risk. A representation of this calculation can be seen in figure 3.1. Note no duplicate measurements (two different measurements for the same patient) are present in the dataset.

As a summary, for the purpose of this section, each patient is represented by an input of: An echocardiogram in the format of a .avi video and a ground truth categorical risk label.

3.1.2 Data Manipulation and preprocessing

A considerable challenge in this research was the gathering, balancing, organizing, cleaning, and pre-processing of the echocardiogram videos, particularly because they were optimized to be rather practical for medical use rather than consistent for deep learning purposes.

The original videos provided to us included annotations such as the image's date, patient code and imaging axis orientation that are perhaps useful for a medical practitioner to view, but make deep learning applications unreliable and distort the learning process. Therefore, we made the decision to black them out to match the rest of the echocardiogram's background. An example of an input video frame is displayed in figure 3.2.



(a) Original Annotated Video frames plus the heart rate monitoring

(b) cleaned data

Figure 3.2: Example of input data

Cleaning the data involved applying a black mask as a top layer on each frame which covered the upper two corners' annotations. We then proceeded to extract the echocardiogram's top pixel starting point, calculated the semi-circle radius, and blacked out everything beyond that radius, with the top pixel being the circle center point.

Following the radius blacking-out procedure mentioned above, the heart rate readings were occasionally found to still appear on top of the main echocardiogram image. For this reason, it was decided to marginally crop the bottom section of the images' frames. The marginal trim applied to the images did not result in a considerable compromise of the echocardiogram images as they

consist of an arch-shaped lower part.

3.1.3 Data Representation

A scope of research in this project was exploring different ways networks behave when given different data representations and analyzing the temporal learning effect when comparing still images vs full long sequences vs short sequences. Regarding our choice for the short sequence (cropped clips) length, we choose a mini-clip length of 16 time-frames based on previous work trends that would achieve the balance of capturing temporal dependencies while keeping the length short thus allowing for creating more clips and not adding complexity overload. Therefore, throughout the experiments presented in Chapter 3, we will be referring to and experimenting with different data representations including:

(A) Independent 2D frame images

(B) The full-length videos (consisting of 48 consecutive frames)

(C) Cropped 16 consecutive frames clips/sequences augmented from the full videos

3.1.4 Class Balancing

Class imbalance is a major obstacle in machine learning in general and can be specifically abundant in most medical imaging datasets due to the natural distribution of case diagnosis and the natural tendency humans have to get tested when a disease has displayed certain symptoms. The problem underlies in the fact that most machine learning models learn to fit their function based on the presented data distribution and with the case of imbalanced datasets, the model learns to perform well on the majority class due to the abundance of its training entities but fails to correctly learn the patterns identifying the minority classes. However, it is usually, and almost certainly in medical data applications, the case that the minority classes are the most important to classify correctly.

In our case, when studying the risk of aortic dissection and rupture, classifying the high-risk cases accurately is a crucial task, and failure to do so renders the model clinically unreliable. According to our ex vivo physical tensile labelling, our dataset only presents with 4 high and 4 low-risk

patients (compared to 24 average-risk patients) creating a major class imbalance problem especially because of the limited size of the dataset, to begin with. Details regarding data distribution between classes can be seen in table 3.1.

Dataset	Total	Low Risk	Average Risk	High Risk
Full Echocardiogram videos	32	4	24	4

Table 3.1: Details Regarding Echocardiograms Original Data Labels

We, therefore, started this project knowing that any architectural experiment to be implemented needs to be paired with data balancing techniques for any learning to take place. We began exploring different methods including over-sampling, under-sampling, weighted loss and the perspective of synthetic data generation which can be seen in appendix A which we didn't proceed with for unreliability purposes. We tested multiple approaches throughout the experiments which are to be presented below and we got the best results when we used the weighted sampling approach.

3.2 Independent Frame Images

We started this work and series of experiments following the trends in medical imaging where most applications tend to divide the imaging sequence into individual frames and use each frame as an independent dataset entry. This practice significantly augments the dataset size (compared to using the entire videos) which is usually required in practice, particularly when using scarce and expensive medical imaging datasets. However, by treating each individual frame as a separate entity, all temporal dependencies between the frames in the sequence are lost.

When implementing the simple single-image classification, we used a ResNet18 as our backbone model due to its powerful performance yet compact design and parameters which were desired qualifications due to our limited dataset size. Using the principles of transfer learning, we preloaded the PyTorch ImageNet [44] ResNet18 weights and froze all the layers except the final fully connected layer which was finetuned to fit our frame dataset. A representation of this architecture can be seen in figure 3.3.

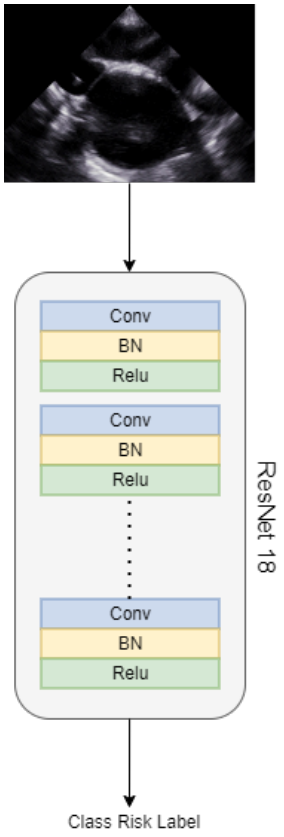


Figure 3.3: single 2D frame input architecture

3.2.1 Experiment Setup

2D Images were extracted by slicing every frame in the echocardiogram video and resized through the PyTorch resize transform to 256×256 and randomly horizontally flipped or mirrored to add randomness. A training/ testing folds split was taken at 80% and one-dimensional test fold accuracy was applied. Details regarding the original(before applying the weighted sampler) training and testing data-class distribution can be seen in table 3.3.

The ResNet18 was pre-trained on ImageNet dataset [44] and its weights were retrieved and loaded into our model through the PyTorch vision library. Cross Entropy Loss and Adam Optimizer were used with a 0.001 learning rate and a scheduled decay of 0.1 with a step size of 15. A summary of these parameters can be found in table 3.2.

Batch Size	Single Input Entity Shape	Loss	Optimizer	Learning Rate
16	$3 \times 256 \times 256$	Cross Entropy	Adam	0.001

Table 3.2: Independent frames as inputs experiment's details.

	Low Risk Frames	Average Risk Frames	High Risk Frames
Train Fold	4822	581	632
Test Fold	848	125	112
Total	5670	706	744

Table 3.3: Details Independent Frames Data Class Distribution

3.2.2 Results

For assessing the success of the model, we treated this problem as a classical multi-class image classification problem and used the single test-fold (as displayed in table 3.3) classification accuracy as the assessment metric. One drawback of dealing with private medical datasets is the lack of different benchmarks or evaluation metrics for comparison. Therefore, the accuracy, confusion matrix 3.4, and visual inspection of the mal-classified entities were utilized as the main assessment criteria.

This experiment resulted in an 80.74 % classification accuracy which meant the model learned some features through the static 2D frames. We note, however, the odd results from the confusion matrix 3.4 where none of the low-risk frames were correctly classified as well as an alarming 34 high-risk videos that were classified as low-risk.

We used this experiment as a base case for static 2D frame entity input which ignores all temporal data between frames and started on a series of coming experiments which utilize the temporal connections between frames.

3.3 3D CNN

Three dimensional convolutional neural networks (3D CNNs) are named as such because they utilize 3D convolutional filters to process the networks' input rather than the conventional 2d convolutional filters which are usually used for processing 2D images. The idea behind 3D convolutions is

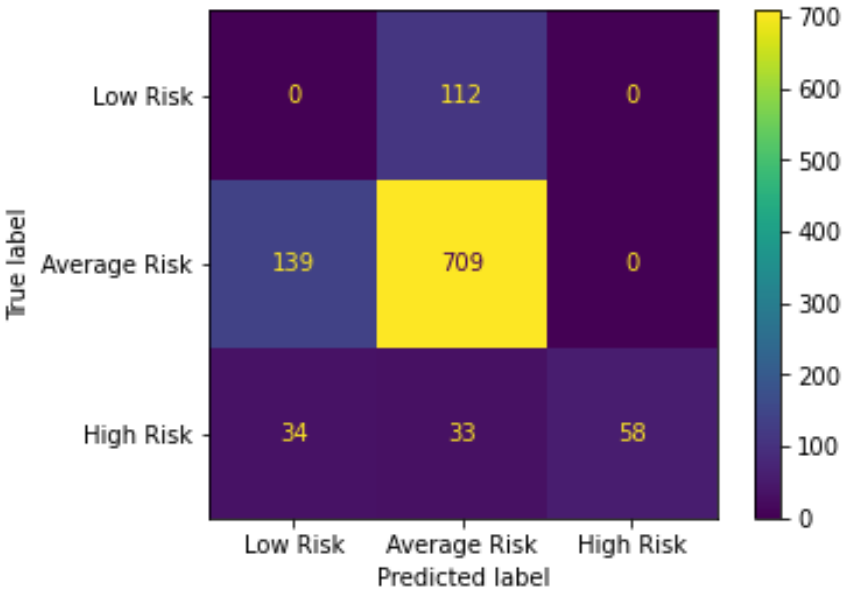


Figure 3.4: Separate 2D Image Frames Confusion Matrix

that a 3-dimensional filter is applied to the input data where the filter moves in the 3 axis directions (x,y,z) and therefore processes the whole sequence input as one 3D entity and outputs 3D volume space (refer to figure 3.5).

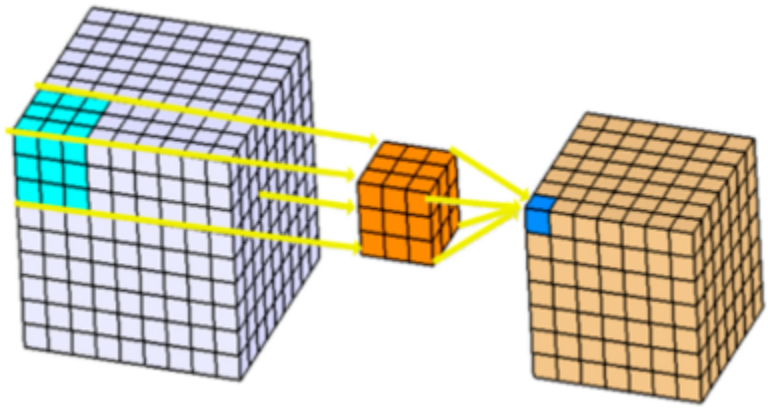


Figure 3.5: 3D Convolutions [29]

Compared to 2D ConvNets, 3D ConvNets - grace to the 3D convolutions and 3D pooling operations - have the capacity to process a third dimension and therefore process the input sequence spatio-temporally, while in 2D ConvNets, only spatial processing takes place. [48] As a result, 3D

ConvNets became rather popular in the field of video classification and action recognition [49] [19]. An Example of how 3D ConvNets processes the input volume can be seen in figure 3.6

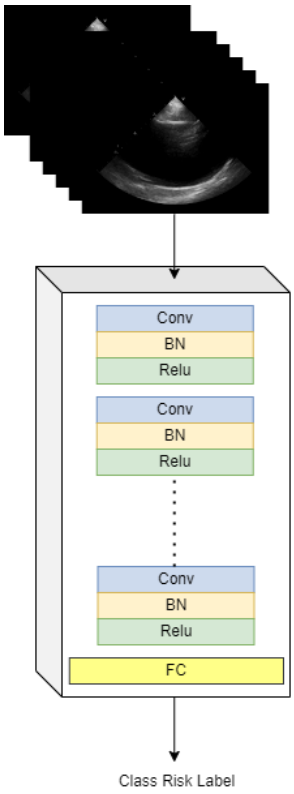


Figure 3.6: 3D ConvNets Processes the image sequence as a whole entity

Therefore, as a start to our video (image sequence) processing series of experiments, we thought of experimenting with 3D ConvNets base models, to compare our findings with still 2d frame experiment and see the effect of how convolving in the third dimension might impact the learning process. We implemented 2 famous 3D convolutional models R3D [19], R(2+1)D [49], (refer to sections 2.4.1 and 2.4.1 for further models' details). We chose these two 3D Conv models specifically because they both utilize ResNet18 as their backbone model, which we want to keep as constant when comparing visual learning methods.

3.3.1 Experiment Setup

In our sequence Experiments, we tested both types of data representation B and C, mentioned in subsection 3.1.3 to see the effect of using the full video echocardiograms and gaining long-term

temporal dependencies on the cost of fewer videos vs augmenting the number of videos (by dividing them to short clips) used on the cost of cropping the full echocardiogram videos and thus losing long term temporal dependencies. Details regarding the original(before applying the weighted sampler) training and testing data-class distribution can be seen in table 3.5. Four Experiments were run in total:

(1) Full length videos on R3D model

(2) Full length videos on R(2+1)D model

(3) Short augmented video clips run on R3D model

(4) Short augmented video clips run on the R(2+1)D model

3D short 16 frames clips and full length videos were extracted with their corresponding frames being cleaned, normalized and resized to o 256×256. The base models were pre-trained on the Kinetics [31] datasets for action recognition. The model's weights were retrieved and loaded through the PyTorch computer vision library. Cross Entropy loss and Adam optimizer were used with a 0.001 learning rate and a scheduled decay of 0.1 with a step size of 15. In addition, due to the higher models' complexity, we could not run the training script with the same batch number we used in the 2D still frames experiment, so we gradually decreased the batch size but could only run our scripts with a batch size of 1. A summary of the parameters can be found in table 3.4.

Experiment	Batch Size	Single Input Entity Shape	Loss	Optimizer	Learning Rate
1	1	$3 \times 40 \times 256 \times 256$ (full video)	Cross Entropy	Adam	0.001
2	1	$3 \times 40 \times 256 \times 256$ (full video)	Cross Entropy	Adam	0.001
3	1	$3 \times 16 \times 256 \times 256$ (cropped clip)	Cross Entropy	Adam	0.001
4	1	$3 \times 16 \times 256 \times 256$ (cropped clip)	Cross Entropy	Adam	0.001

Table 3.4: 3D ConvNets experiments' details.

3.3.2 Results and Analysis

Again for assessing the success of the following architectures and understanding and analyzing the results, we used the single-test fold (as displayed in table 3.5) classification accuracy as

	Full Videos			Cropped Clips		
	Low Risk	Avg Risk	High Risk	Low Risk	Avg Risk	High Risk
Train Fold	3	19	3	38	295	35
Test Fold	1	5	1	7	52	7
Total	4	24	4	45	347	42

Table 3.5: Details of Full vs Clipped Video Class Distribution

the assessment metric. It was generally noted that the mini-cropped-video clips resulted in an enhanced performance compared to the full-uncropped videos. Our rationale behind this result is that providing the model with more entities to learn from, led to the learning of classifications of more meaningful patterns while still retaining temporal information that the mini clips presented. It was additionally noted that the use of 3D ConvNets at experiments 1, 2, and 3 gave a lower classification accuracy compared to the still 2D frames experiment in 3.2.

When analyzing these results 3.6, we wondered whether these results were due to the still 2D frames having enough visual context to capture the function mapping with the temporal aspect playing a smaller role or due to the fact that the 3D ConvNets have enormous numbers of parameters that ended up overfitting our majority class distributions regardless of the efforts we made to balance it out. Another thought on 3D ConvNets is that perhaps convolving in the third dimension with the aim of capturing the temporal data might have ended up being too noisy to capture any motion patterns. In addition, we note in figure 3.7 a similar trend to the one we saw in the previous confusion matrix 3.4, where the correct classification of the average-risk inputs is superior compared to those of the low and high-risk inputs. This might be attributed to the fact that despite our efforts to balance the training class distribution, the model was still learning the abundant class patterns more dominantly.

To ponder more on these ideas we moved on to section 3.4 where we tried to minimize the 3D convolutional noise by separating the video classification process into 2D visual extraction and sequential processing steps.

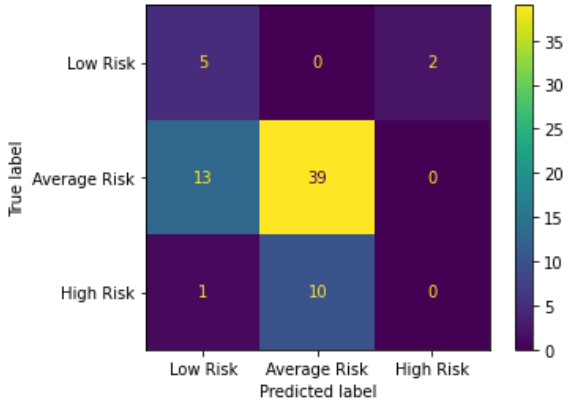


Figure 3.7: Cropped Clips-R3D Model Confusion Matrix (Best Performing 3D Conv Combination)

Experiment	Input Type	Model	Accuracy
1	Full Video	R3D	71.43 %
2	Full Video	R2+1D	71.43%
3	Cropped Clips	R3D	**83.33%**
4	Cropped Clips	R2+1D	78.79%

Table 3.6: 3D ConvNets experiments' Results.

3.4 Aggregated 2D Frames Feature Maps

After Noting the modest results of our 2D still picture frame experiments and the overloaded complexity that the 3D ConvNets introduced, we initiated the investigation of a middle-ground architectural point that could perhaps visually process each of the video's frame images separately and then aggregate over the extracted feature maps output in a sequential manner.

The architectural structure of Aggregating 2D frames in a video consists of a visual backbone network to purely capture the spatial patterns in each frame separately, as each frame is processed in the visual network as a separate entity. Following that, the outputted feature maps are grouped again to be processed as a series of sequence-length time steps. An illustration of this representation can be seen in figure 3.8.

In accordance, we start the following section of this work where we explore hybrid frameworks that utilize the 2D frame aggregation and sequential assessment architectural approach. Throughout the following 2 subsections (3.4.1 and 3.4.2), we will be logically dividing our model into two

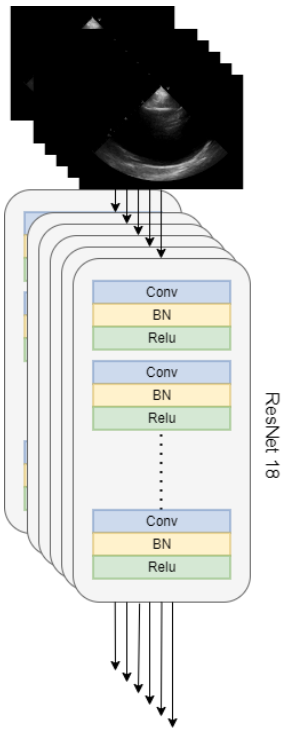


Figure 3.8: 2D convolutional processing of frames sequence

subsections: visual extraction and temporal sequential assessment.

3.4.1 LRCN

As videos in nature are comprised of consecutive visual information over time and with recurrent neural networks(RNN)s being a family of deep learning models specifically associated with modelling temporal dynamics, they consequently were a natural decision to process the outputs of the convolutional feature map.

More specifically, LSTMs, among other RNN variants, have shown to be effective in modelling long-term temporal dependencies and have been applied in various literature in incorporating sequential learning in computer vision ranging from caption generation [57], to unsupervised learning of video representations [47] and action recognition and classification [54] [10].

Specifically, we chose to implement and build our framework similar to the LRCN model presented [10]. Unlike the authors in [10] who utilized a combination of CaffeNet[26] and a unidirectional LSTM, we utilized the ResNet18 as our choice of visual features backbone to extract spatial

patterns in our architecture and fed the resulting feature maps to a bi-directional LSTM structure to capture the temporal dynamics between the frames. We also chose to employ a bi-directional LSTM since the nature of our problem can utilize the bi-directional learning process.

In addition, we compared the above architecture with the baseline of averagely pooling the collected feature maps to establish if the LSTM blocks present any added sequential learning effect. An illustration of the architecture can be seen in figure 3.9.

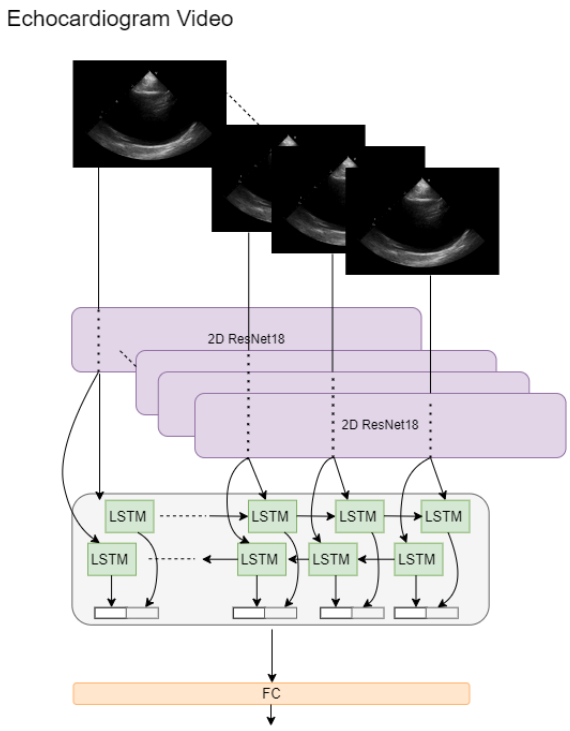


Figure 3.9: LRCN architecture

Experiment Setup

Regarding the inputs of this series of experiments, we also compared learning outcomes between types B and C mentioned in 3.1.3 to formulate a conclusive result on whether the short augment clips are a superior form of input compared to the full videos. We utilized the same training and single-testing folds mentioned in table 3.5. As mentioned above, we also tested the variability of the sequential data processing section in the model in terms of average pooling (which we will be referring to as ResNet18+AvgPooling model) and using LSTM blocks (which we will be referring

to as LRCN) so in total 4 experiments were run which can be summarized as:

(1) Full length videos on LRCN model

(2) Full length videos on ResNet18+AvgPooling model

(3) Short augmented video clips run on LRCN model

(4) Short augmented video clips run on the ResNet18+AvgPooling model

For consistency in terms of spatial feature extraction, PyTorch pre-trained ResNet18 was kept as the visual extraction backbone with its layers frozen, except for the last, fully connected layer. To simulate the process of separately processing each frame and aggregating the sequence later in given a video length T, we reshaped the Data Input from shape $N \times T \times C \times H \times W$ to shape $(N \times T) \times C \times H \times W$. This $N \times T$ frame series was then fed to the ResNet as a separate series of 2D images which results in an aggregated feature map. Finally, we followed to reshape the extracted feature maps to $N \times T \times latentDimension(LD)$ representing the T time steps, with LD = 512 based on previous trends [10], and fed the T time steps feature maps to the LSTM.

We selected similar training hyperparameters to our previously conducted experiments' parameters in terms of Cross Entropy Loss, Adam Optimizer with a learning rate of 0.001 and a scheduled decay of 0.1 with a step size of 15. In addition, unlike the computational overloads in the 3D ConvNets experiments which did not allow for increased batch size, we set our training batch size to 16. A summary of this section's experimental parameters can be seen in table 3.7.

Experiment	Batch Size	Single Input Entity Shape	Loss	Optimizer	Learning Rate
1	16	$40 \times 3 \times 256 \times 256$ (full video)	Cross Entropy	Adam	0.001
2	16	$40 \times 3 \times 256 \times 256$ (full video)	Cross Entropy	Adam	0.001
3	16	$16 \times 3 \times 256 \times 256$ (cropped clip)	Cross Entropy	Adam	0.001
4	16	$16 \times 3 \times 256 \times 256$ (cropped clip)	Cross Entropy	Adam	0.001

Table 3.7: Aggregated 2D LSTM and Pooling experiments' details.

Results and Analysis

We observe in table 3.8 that (similarly to the results achieved in the 3D ConvNets experiments) the mini-cropped-video clips performed better than the full un-cropped videos (a 17.96% increase in the LRCN experiments and a 14.96% increase in the pooling experiments). We reason the logic behind these results in a similar manner in terms of the model having augmented dataset entities to learn from while keeping the sequence long enough to capture temporal dependencies. We thus conclude our cropped clips as the superior data representation format for the following sections.

In terms of the differences in accuracy based on the architectural difference (pooling vs lstm), we interestingly note only a small 2% gap between the results from experiments 3 and 4. We reason that the ResNet is potent enough to capture still visual patterns while the pooling operation, even if it is trivially temporally processing the frames, is sufficient to compare to the LRCN performance. The LRCN superior performance, however, which it offers due to its advantageous LSTMs powerful sequence learning blocks, suggests the presence of temporal dependencies in the input. In addition, we note in figure 3.11 the same trend we saw in the previous experiments of failure to correctly classify low-risk inputs compared to high-risk and average-risk ones. Along with our initial analysis of the model overfitting to the majority class distribution, we also reason this case to the imaging view disparities between our given low-risk echocardiogram videos.

Finally, we also tried looking deeper into the model's interpretability aspect. We implemented a saliency map visualization of where the model's neurons were being activated by visualizing the gradients through the image channels. Saliency maps can be thought of as a small section of attribution methods of interpretability. Attribution methods as described by [37] study what sections of the input (in our case pixels of the image) are responsible for the model to act in a certain way. Therefore, saliency maps can be thought of as visual heatmaps indications that highlight pixels of the input image that mostly influenced the classification decision.

In order to implement the gradient-based saliency maps, we filtered the maximum values' indices of the raw output of our network and took the raw output at these indices which we called the score vector. We then followed to backpropagate through the score vector to obtain the weights values which represent the most influential pixel locations in the input image. Finally, we took the

maximum values across the channels of the weights vector to end up with a single-channel saliency map. A representation of our saliency maps across different frames can be seen in figure 3.10. As we noted in figure 3.10, the model seemed to mainly focus on the content inside the aortic ring, with the exception of the noise present around the image edges.

Due to the recent research which examined the validity of saliency maps [1], we took the interpretation presented in this work with caution as a mere simple interpretation of noise being present in the learning process.

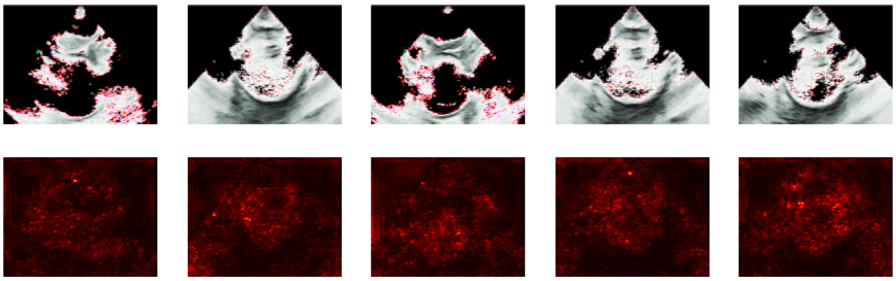


Figure 3.10: Sample echocardiograms frames Saliency Maps

Experiment	Input Type	Model	Accuracy
1	Full Video	LRCN	71.43%
2	Full Video	Average pooling	71.43%
3	Cropped Clips	LRCN	**89.39%**
4	Cropped Clips	Average pooling	86.39%

Table 3.8: 2D aggregated frames experiments' Results.

3.4.2 Attention Mechanism

Studies in visual cognition have shown that humans naturally focus on various specific features in a visual image rather than compressing the whole scene altogether [43]. This is probably related to the fact that while images contain beneficial information for the visual recognition and assessment task, they also contain a lot of clutter and irrelevant noise. With the advancement of deeper networks and the rise in interest of the seek in interpretability, attention-based models are getting more and more popular and are achieving promising results in challenging visual tasks including

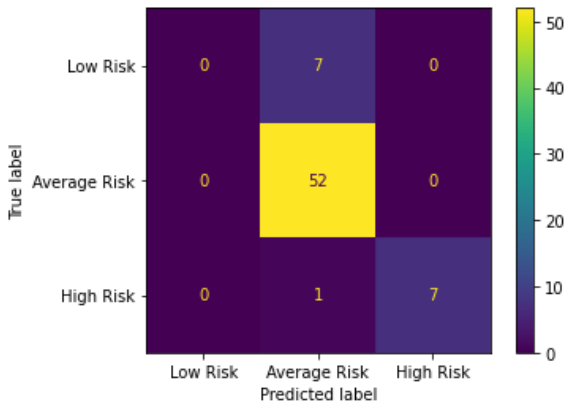


Figure 3.11: cropped clips with LRCN (best performing) model

image captioning [55], object detection [59] and action recognition [45].

When discussing self-attention mechanisms, it is worth mentioning the recently emerged field of transformers that, after being initially introduced in the natural language processing field [50], have proved their robustness and efficacy in the world of computer vision [11] and have revolutionized the indisputable standing that CNNs possessed in the Large Scale Visual Recognition Challenge (ILSVRC). We perceive, however, that with the robustness that vision transformers introduced, came their "data-hungry" training dependency that rendered vision transformers not as successful as CNNs if not given sufficient amounts of data to train on. Due to the mentioned reasons, and given our extremely small dataset, we chose to experiment with incorporating self-attention mechanisms into our model rather than experimenting with a pure-based transformer backbone network.

To do that, we chose to utilize the non-local neural network presented in [52] which is based on the non-local operation originally presented in [7] to capture long-range dependencies and de-noising the visual input to our network. In its essence, the generic non-local operation can be thought of as a method to capture context directly between any two positions in the visual content, regardless of their spatial proximity.

$$y_i = \frac{1}{C(X)} \sum_{\forall j} f(X_i, X_j) g(X_j). \tag{2}$$

To compute a more meaningful, spatially aware signal, the non-local operation is comprised of a combination of A) A pairwise function f to compute the spatial relationships of all position j

compared to position i in the input image X and B) A unary function g to represent an embedding representation of X at j. As presented in equation 2, y_i represents the non-local operation's output signal, which has the same size as the input image X and C(X) points to the normalizing scaling factor.

In our implementation, we went with the standard choice of g being $g(X_j) = W_g X_j$ with W_j as the weights matrix to be learned through a 1×1convolutions. For the decision of the choices for the pairwise function f, we chose the embedded Gaussian function presented in 3. We note that with f being the Gaussian embedding function, y turns into the generic module of self-attention [50]. As for the normalization factor, C was set as per the original paper to $C((X) = \sum_{\forall j} f(X_i, X_j)$. An illustration of the non-local block can be seen in 3.12 which is an adaptation of the figure in [52] where the authors chose to temporally pool the resulting visual feature maps.

$$f(X_i, X_j) = e^{\theta(X_i)^T \phi(X_j)} \tag{3}$$

Where

$$\theta(X_i) = W_\theta X_i \qquad\qquad \phi(X_j) = W_\phi X_j$$

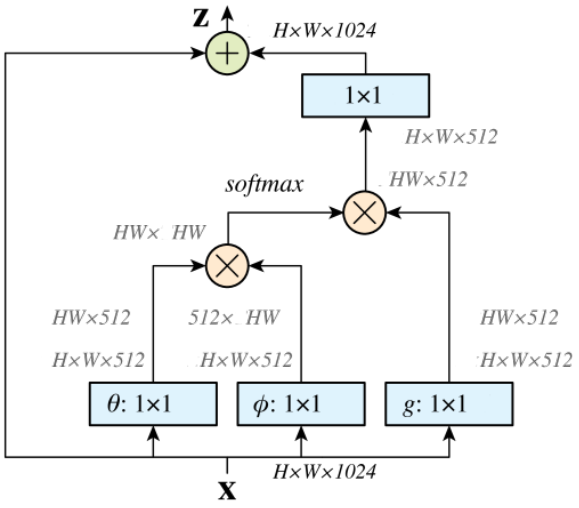


Figure 3.12: spatial non-local block

Experiments Setup

We wrapped up our architecture by inserting the non-local block right after layer 2 in our original ResNet18 model. we chose the location right after the second layer to insert our non-local block based on the results of [52] that showed that the experiments with the attention units inserted in the earlier layers showed better outcomes than the ones with the blocks inserted in the later layers.

Since we chose to implement the non-local blocks as a method of spatial attention only, we also kept our LSTM as the second step to process the feature maps of the ConvNet output. In addition, We also kept the ConvNet base as the ResNet18 and utilized the same cropped clips training and single-testing folds as the previous sections mentioned in table 3.5 under the mini cropped clips category.

Therefore, our model consisted of a visual extraction section which takes as input image sequences of shape $(N \times T) \times C \times H \times W$, feeds it to a ResNet18 with a non-local attention block between its second and third layers and finally output the visual feature maps of shape $T \times 512$ We then proceeded to feed the LSTM the input sequence of T timesteps just like we did in section 3.4.1.

In addition, we also selected similar training hyperparameters to our previously conducted experiments' parameters in terms of Cross Entropy Loss, Adam Optimizer with a learning rate of 0.001 and a scheduled decay of 0.1 with a step size of 15. However, due to the computational overload of the non-local blocks, we had to downgrade the image size to 128×128 and use a batch size of 1.

Batch Size	Single Input Entity Shape	Loss	Optimizer	Learning Rate
1	$16 \times 3 \times 128 \times 128$ (cropped clip)	Cross Entropy	Adam	0.001

Table 3.9: 2D aggregated frames with attention blocks experiment details.

Results and Analysis

As conducted previously, the assessment of success and the interpretation of the result depended primarily on the test single-fold classification accuracy as the assessment metric. Surprisingly, the achieved results 3.10 were lower than those achieved by our LRCN model. While this might be caused by the input images quality being downgraded to 128 × 128 while training on batches of 1,

we also reason that, due to our small dataset, the attention map did not have enough training entities to learn from and ended up learning irrelevant noises which distorted the learning process rather than fine-tuning it.

Input Type	Model	Accuracy
Cropped Clips	Attention-LRCN	78.79 %

Table 3.10: 2D aggregated frames with attention blocks experiments' results.

Chapter 4

Textual Data Fusion

Towards the later stages of writing this dissertation, McGill University's Leask Laboratory team presented to us an additional dataset consisting of tabular or textual information which has the patients cohort we dealt with in Chapter 3 as a subset of its cohort. We were therefore intrigued to compare the echocardiograms learning process against pure statistical machine learning approaches and aimed to incorporate the echocardiograms' learning and the tabular data learning streams. In this chapter, we will start by describing the additional patients' information dataset, followed by introducing our hybrid model which combines the visual, sequence, and tabular data learning streams. We will also be basing some of the feature selection and tabular model decisions on previous work on the subject conducted by the Leask Laboratory's graduate students [32] which has assessed the statistical approach of the project.

4.1 Dataset

4.1.1 Original Dataset Description

Grace to the McGill Leask Labs, and in compliance with the Canadian tri-council policy statement on ethical conduct for research involving humans, informed consent for anonymous data collection was obtained from a cohort of 66 distinct patients undergoing elective aortic valve repair or replacement surgery. The cohort of patients' echocardiogram videos presented in Chapter 3 is a subset of this cohort.

Various patient information categories were collected including: patient demographic information, comorbidities, lifestyle information, aortic geometries, and genetic information. A detail of all the collected information can be viewed in table 4.1.

Variable Name	
Female(0/1)	age (yr)
height (m)	Weight (kg)
bsa	bmi
Systolic Pressure (mmHg)	Diastolic pressure (mmHg)
Bicuspid Aortic Valve	History of smoking
Dyslipidemia	Coronary Artery Disease)
History of Hypertension	Diabetes (I/II)
Aortic diameter/BSA	family history of aortic disease
Energy Loss (mean of both axis)	Non-Marfan Genetic Disorder
Tensile Measured Energy loss (short axis)	
Tensile Measured Energy loss (long axis)	
Severity based on Diameter/BSA and Age	
Presents with NYHA Heart Failure Symptoms	
Aortic valve stenosis (1=mild, 2=mod, 3=severe)	
Aortic Valve Insufficiency (1=mild, 2=mod, 3=severe)	
History of Alcohol Use (7 drinks a week or more)	
Reported Aortic Diameter (Based on CT, MRI, or echo) (mm)	
Ascending diameter (measured from surgical TEE) (mm)	
Sinus ofa Valsalva diameter (measured from surgical TEE) (mm)	
Ascending Aortic Surface Area (measured from surgical TEE) (cm2)	
Genetic Variant of Unknown Significance	
Marfan Syndrome (0=none, FBN VUS =1, positive=2)	

Table 4.1: Patients collected tabular information.

4.1.2 Data Manipulation and Preprocessing

In general, Machine learning models learn whatever we feed into them. Feeding noise will categorically produce noisy predictions and the curse of dimensionality is a complication that has intrigued research in the field of data science over the past decades. In addition, especially given our very small dataset, increasing the dimensionality would exponentially increase the number of parameters which would probably lead our model to over-fit more than it already is.

Therefore, when presented with a 31 features dataset corresponding to patients' information, we went back to the work presented by the Leask Laboratory where the author implemented input

ranking selection of all the patients metadata features.

Figure 4.1 displays the work of the Leask Lab in terms of assessing feature importance and its influence on the classification label in terms of the F-score and Independent Importance methods. Based on this, we decided to only include the top-score features in our tabular dataset. We also eliminated the ascending diameter/BSA feature for redundancy and correlation purposes as their significance is already present in the Ascending Diameter and BSA separate features. A figure of feature correlation before and after eliminating the diameter/BSA feature can be seen in figures 4.2 and 4.3. We also noticed negative correlation between the female feature and the BSA and History of Hypertension features which make sense since females tend to have smaller body surface areas compared to men as well as having a lower statistical chance of developing hypertension. We choose to keep these features due to their medical diagnostic importance.

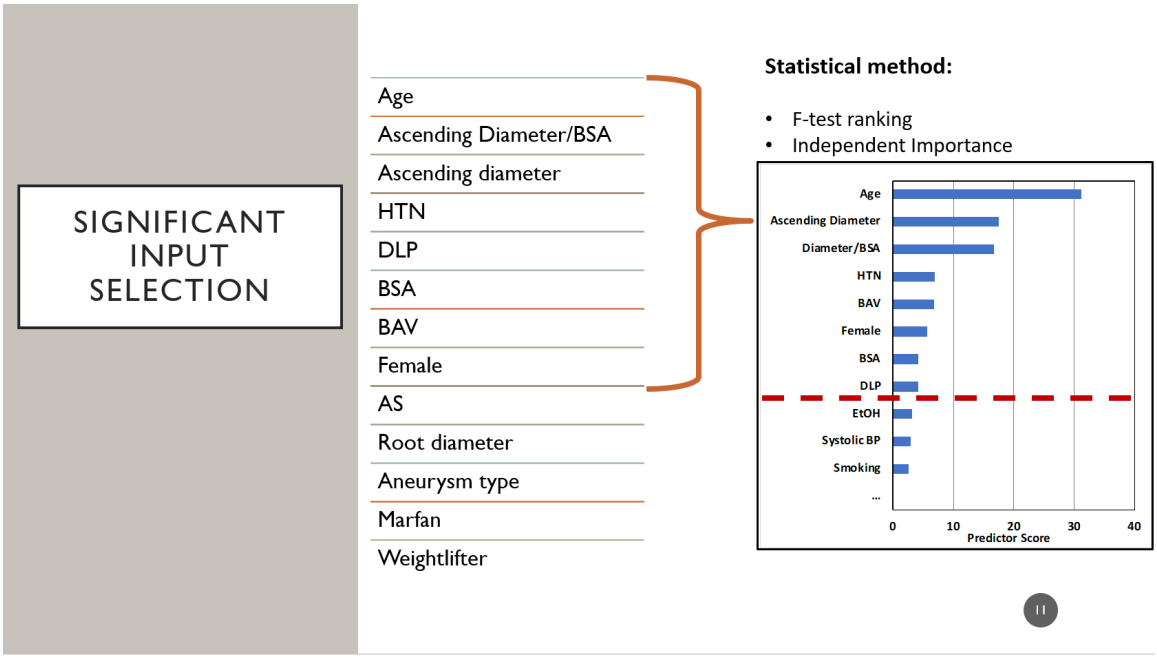


Figure 4.1: Feature Importance [32]

4.1.3 Data Representation

Seeing the nature of our tabular data, and since we took the decision to incorporate the patients' meta data learning as a dense stream in our deep learning model, we had to encode the discrete data as feature vectors. For the female, Bicuspid Aortic Valve and History of Hypertension features

we chose a one-hot binary encoding to represent True/False representation. As for the rest of the numerical features, we went with the representation of a single numerical float value for each. We also applied the min - max scaling approach to the numerical value entities to decrease the noise and stabilize the learning process.

4.2 LRCN-Tab

Echocardiograms and their deduced measurements play a vital role for pathologists to distinguish high-risk aortic aneurysms. However, even when provided with near-conclusive imaging results, clinicians and pathologists traditionally rely on patient-specific data to support their diagnostic decision. Clinical data such as the patient's age, gender, the existence of congenital defects, and history of the disease in the family are all factors that are usually attributed great weight within the clinical diagnostic process, especially for diagnoses relating to cardiovascular diseases. For example, the pathologist risk assessment will greatly differ given a case of a 4cm ascending aortic valve measurement of a 5 yr old compared to that of an 80-year-old patient.

We, therefore, decided, given our specific cardiac risk assessment task, that incorporating such information would add significant clinical value to our diagnostic risk assessment model. To get the best performance possible, we sought to incorporate the tabular data learning stream with our so far best-performing model from section 3.4.2.

Since we wanted to create an ensemble end-to-end trainable model, we decided on processing the tabular information in a dense separate stream through consecutive fully connected layers that will accomplish the job of extracting the meaningful patterns in the tabular data. The challenge, however, was deciding on a fusion method to combine the two streams. We were inspired by [23] techniques regarding different streams fusion models and decided on experimenting with two approaches:

(1) Simple concatenation approach: concatenating the two feature vectors from the visual-sequential and the tabular dense learning streams right before the last fully connected layer.

(2) Squeeze-and-Excitation (SE) [24]: The tabular learning stream is used to scale the visual-sequential learning feature vector in an SE approach followed by a fully connected layer.

An overview of the general model architecture, which does not specify a specific stream fusion method, can be viewed in figure 4.4.

The simple concatenation technique, with a data loader of batch size B, would follow the logic of concatenating the visual sequence stream coming from the bidirectional LSTM last hidden states' outputs(lstm_out) of shape $B \times (2 \times lstm\_out)$ with the tabular learning stream output(meta-out) of shape $B \times meta\_out$. Therefore, simply concatenating the two outputs on the first dimension would results in the final stream of shape $B \times (2 \times lstm\_out + meta\_out)$

On the other hand, the squeeze and excitation fusion method is slightly more complex. SE blocks were developed as complementary blocks to CNNs to improve channel interdependencies at a trivial computational cost. SE blocks work by adding weighted content or relevancy to each channel. The squeeze operation is achieved by using global average pooling to generate channel-wise statistics. Given an input image U of dimensions $H \times W \times C$, per channel c, the output z of the global averaging layer is generated by shrinking U through its spatial dimensions H × W, such that the c-th element is calculated by:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{4}$$

resulting in a vector z of shape $1 \times 1 \times C$ representing the context of each channel c which is then fed into two consecutive fully connected layers along with their activation functions to add non-linearity representations. Finally, the resulting channel-wise feature maps are multiplied with the original image U as a scaling factor to present an output capable of exploiting channel dependencies.

In our case, since we are not squeezing and exciting the same input but rather two different streams, we will use the tabular data stream output to act as the scaling excitation stream z in equation 4 and give context to the sequential visual output stream coming from the LSTM which is represented as the image U in the original description.

4.2.1 Experiments Setup

As discussed we will be taking our so-far best-performing model and incorporating the tabular learning stream into it. To incorporate the two streams, we experimented with two different tabular

data fusion techniques, as discussed in the section above, to test the effect of the addition of our patient information on our imaging deep learning model.

We kept the same visual sequence learning base from our best performing model 3.9 which consisted of the same pre-trained 2D ResNet18 visual processing backbone, a bidirectional LSTM which took the visual feature maps outputs and processed them as a T time steps sequence (with T corresponding to the video sequence length of 16 frames) and a final fully connected network. We also utilized the same cropped clips training and single-testing folds as the previous sections mentioned in table 3.5 under the mini cropped clips category.

Therefore to distinguish the two fusion techniques, 2 experiments were run:

(1) The LRCN-Tab model with the concatenation fusion method

(2) The LRCN-Tab model with the SE fusion method

The same remaining training parameters used for the previous experiments were utilized again including: Cross Entropy Loss and Adam Optimizer with a 0.001 learning rate and a scheduled decay of 0.1 with a step size of 15. The inputs were loaded through the PyTorch data loaders and consisted of two parts, the visual extraction input of shape $(N \times T) \times C \times H \times W$ and the tabular data input of shape $N \times meta\_categ$ corresponding to the number of features in the patient's tabular features.

Experiment	Batch Size	Single Input Entity Shape	Fusion Method
1	16	$16 \times 3 \times 128 \times 128$, 16×6	Concatenation
2	16	$16 \times 3 \times 128 \times 128$, 16×6	SE

Table 4.2: LRCN Tab params

4.2.2 Results and Analysis

We continued to use the test single-fold classification accuracy as an assessment of the success of our model and analyze the results of its experiments. As seen in the results table 3.8, to our surprise, the simple concatenation fusion method model achieved the exact results our LRCN model did, and the SE fusion method model achieved much lower results. We reason that the tabular data input somehow distorted the learning process rather than enhancing it and with the SE fusion method

reinforcing the weight of the tabular learning stream, it actually contributed more to degrade the performance.

Interestingly, we note the same confusion matrix in our experiment(seen in figure 4.5) as the one in Chapter 3 with the LRCN model's confusion matrix 3.11. This is probably due to the fact that the two models have the same corresponding test fold classification accuracy. This also might suggest that the LRCN-Tab model applies more weight to its visual learning stream rather than the tabular learning stream in its classification decision process.

In light of the results above and with the performance of the LRCN and LRCN-Tab models being extremely similar, we tried re-running their respective experiments with a cross-validation approach with k = 4 folds. Running cross-validation in a video classification application where we are considering cropped clips was a challenging task that had to be dealt with attentively. The reason goes to the fact that although each cropped clip (video segment) is a separate training entity with its corresponding label, it is still semantically a part of the original video pertaining to the same patient. Therefore, grouping these clips in one pool and performing traditional k-fold cross-validation will very likely place clips pertaining to the same patient in both the training and testing folds. We thus, made this train/test folds separation at the higher patient (full videos) level. However, given that in our dataset some full videos are much shorter than the others (up to 7s difference and with the videos having 30 frames per second (fps)) and having only four low-risk and four high-risk patients meant that having a short full video in the training set will subsequently produce very little cropped clips and therefore diminish the already small training dataset. As a result, we noticed the great discrepancies between the results of each fold as presented in table 4.4 as the sub-samples from the original videos changed.

Experiment	Fusion Method	Accuracy
1	Concatenation	89.39%
2	SE	74.24%

Table 4.3: LRCN Tab params

	LRCN	LRCN-Tab
Fold 1	89.39%	89.39%
Fold 2	87.88%	86.36%
Fold 3	65.63%	68.75%
Fold 4	72.5%	73.15%

Table 4.4: LRCN and LRCN-Tab cross-validation experiments

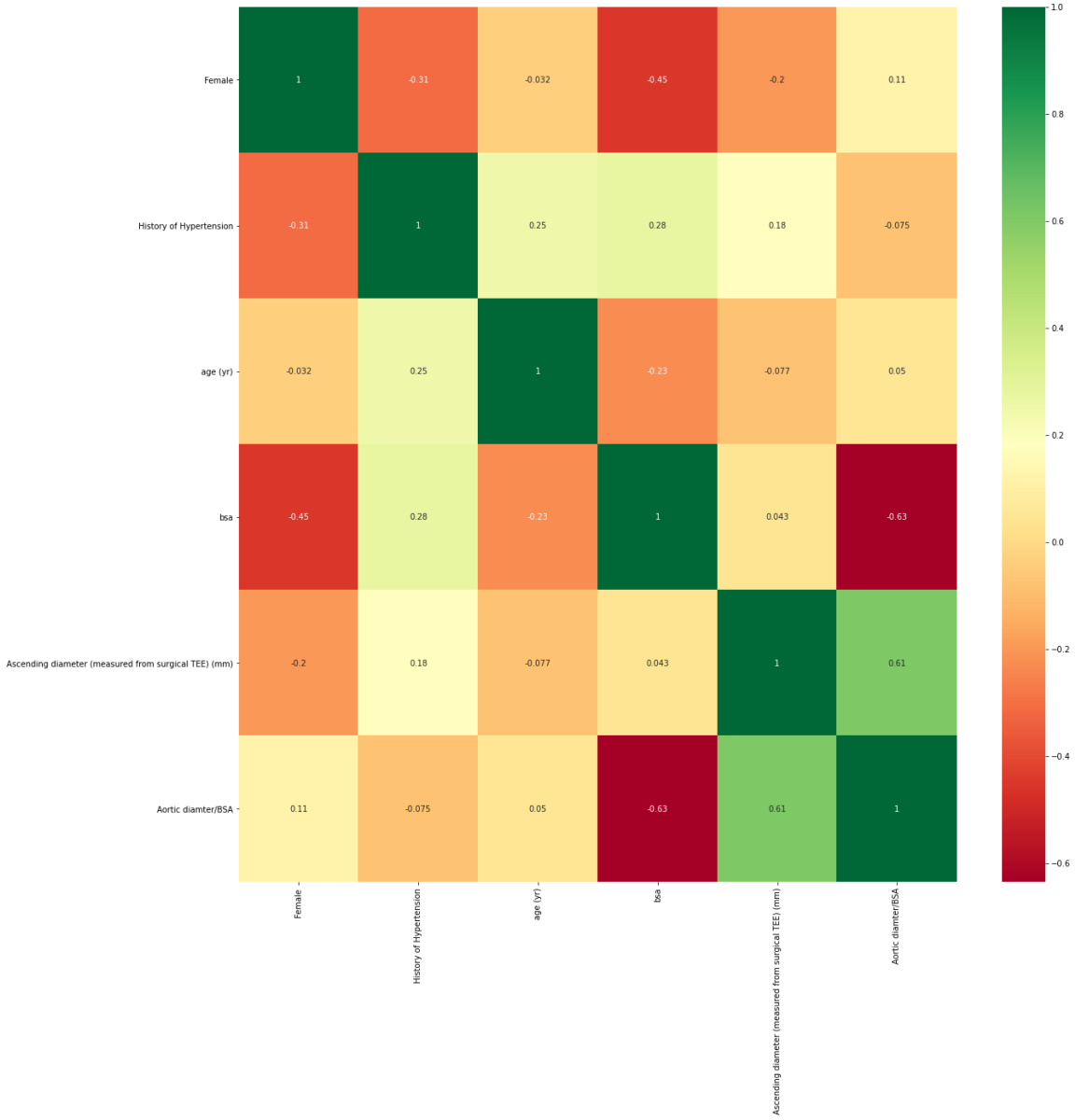


Figure 4.2: Feature correlation before eliminating diameter/Bsa

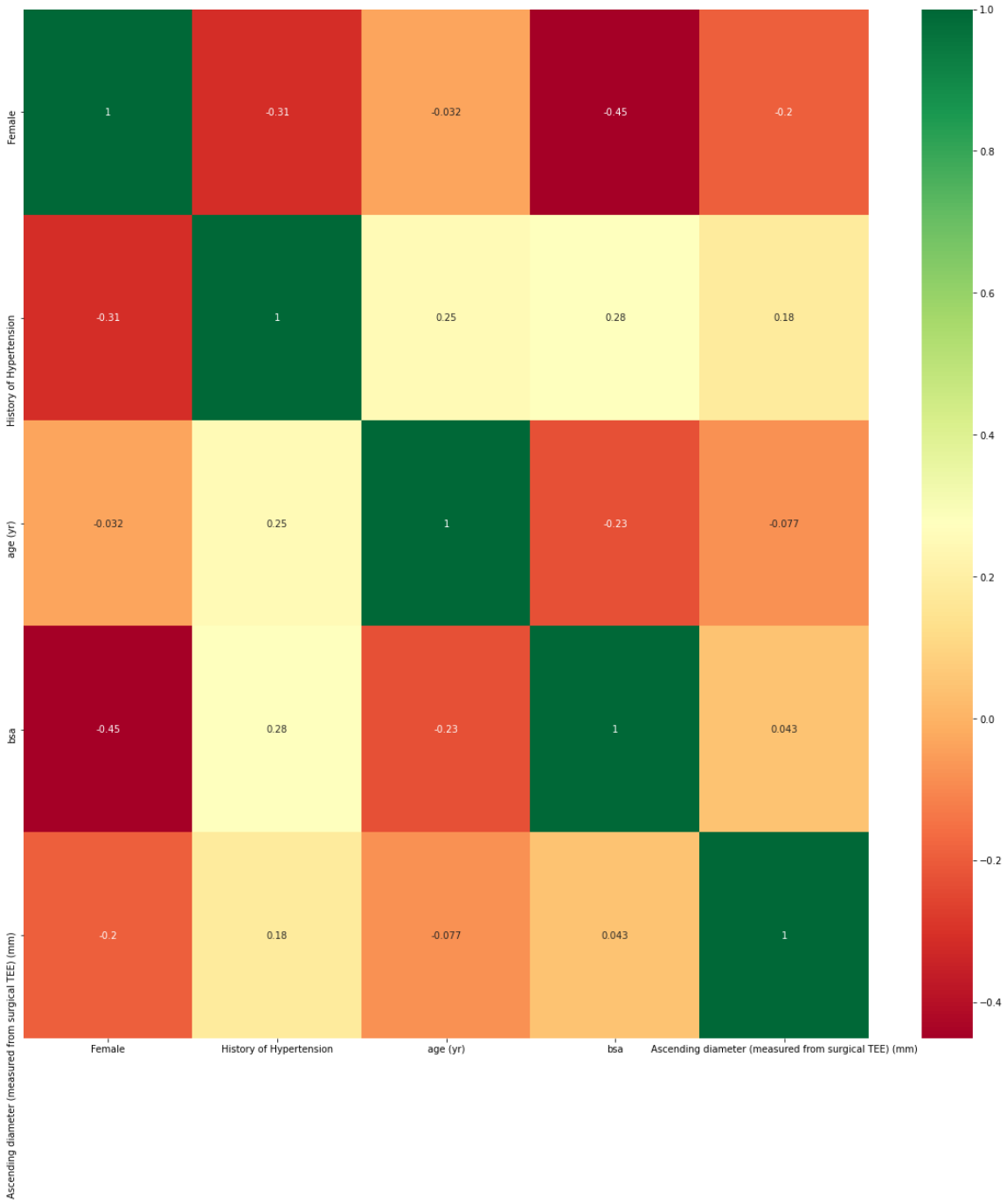


Figure 4.3: Feature correlation after eliminating diameter/Bsa

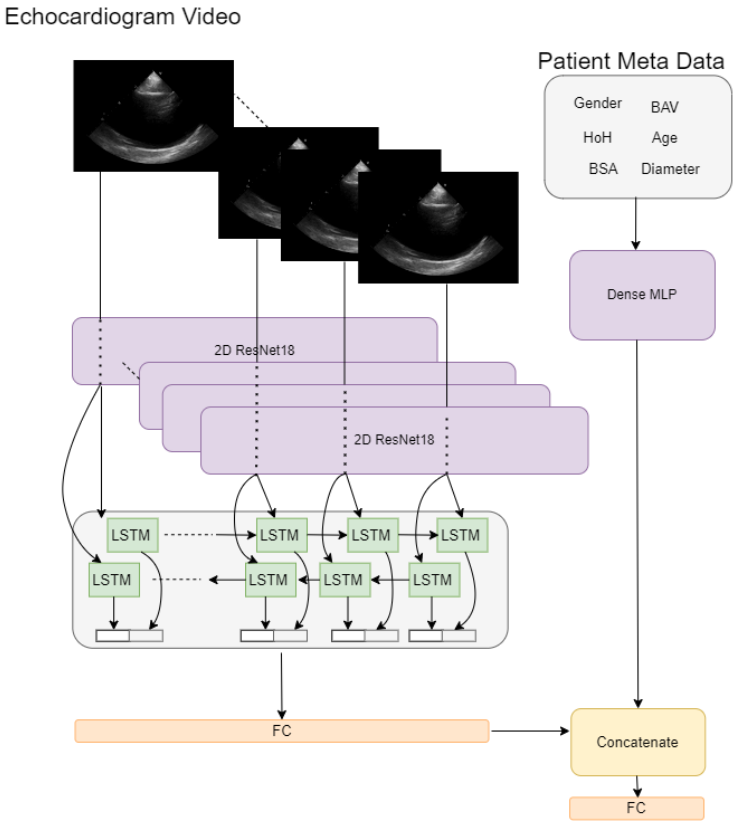


Figure 4.4: LRCN-Tab

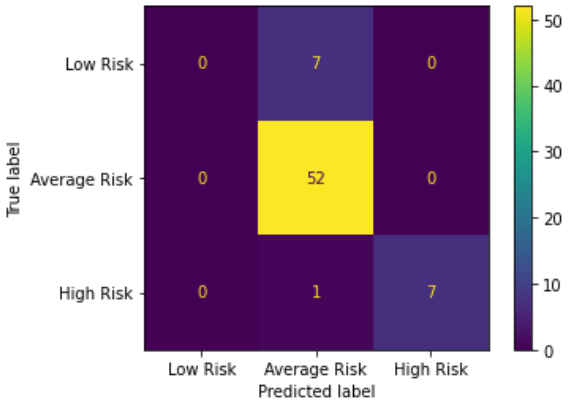


Figure 4.5: Cropped clips with LRCN-Tab concatenation model

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, we investigated different spatial, temporal, and tabular deep-learning approaches to assess aneurysms' risk in thoracic aortic tissue. We split this work into two logical sections pertaining to the datasets dealt with throughout this thesis. In the first section, we explored the optimal data representation in terms of the full echocardiogram videos as opposed to the clipped augmented mini-clips and the independent still frames. While incorporating different data formats, we also compared different sequential visual convolutional-base architectures that could learn spatial as well as temporal features through the data inputs. In our second section, We followed to incorporate a complementary tabular dataset pertaining to the patient's tabular information by a separate tabular data learning stream and then fused the visual and tabular streams together to create an ensemble end-to-end trainable model.

Our results can be split per our experimental hypothesis: We first present the clipped augmented video clips as a superior input data format, compared to inputting the full videos or the still frames. We reason this is due to the fact that the augmented clipped videos(although less in training entities than the still frames and shorter in sequence length than the full videos) achieved the balance of creating enough training entities for the network to learn meaningful spatial patterns while attaining the sequential length for temporal learning between the frames.

In terms of the optimal visual learning approach, we observed the efficiency that the aggregated

2D ConvNet grouped with a bidirectional LSTM presented in terms of the 2D ConvNet not presenting the complexity and the overfitting that the 3D ConvNets presented and still being potent enough to capture the spatial patterns and the sequential learning properties that the LSTM leveraged to learn dependencies between the frames' feature maps. The superiority of a sequential learning architecture also suggests the presence of temporal dependencies and patterns between the frames in the echocardiograms.

5.2 Limitations

As mentioned in this work, the dataset itself represented a challenge which limited this work in terms of: The size of the data: While we did achieve promising risk classification results, the extremely small size of our dataset had left us wondering about the validity of our results and analysis of every architecture applied.

The validity of the input's labels: The discrete class labelling of low, average, and high risk in itself is a debatable subject due to the lack of deterministic diagnostic properties. We note this is especially applicable regarding the low-risk inputs since eventually all patients going through aortic valve replacement surgery (the source of the data collection) will have some risk associated.

The Timeline of acquiring the dataset: The acquisition of this dataset came in a number of steps in time, which was a major driving factor for the amount of work spent in each section in this thesis.

5.3 Future Work

Hopefully, the work presented in this thesis can be considered as a modest stepping stone in the world of exploring Spatio-temporal learning approaches in ultrasound imaging. Below are some future work directions on this topic which might be worth investigating.

Large scale Dataset Applications Application of this work on large-scale ultrasound datasets (though an extension on the current dataset or perhaps in a multi-task learning approach on another closely related task) would verify our results and analysis which would make it generalizable on other sequential medical imaging modalities.

Generation of new data as a source of augmentation Another interesting approach to the

case of small datasets is the generation of new input entities through generative models which have gained vast popularity through the last decade. This topic was explored in Appendix A but was stopped due to the need for medical verification of the generated (inputs/labels).

Different tabular learning and fusion streams While we have explored the simple MLP tabular data learning and fusion through classic and squeeze and excitation methods, it would be interesting for example to consider other gradient-differentiable tabular learning schemes and fusion methods such as differentiable trees and weighted fusion techniques. Another point of view is even treating the two problems as separate but closely related tasks and exploring multi-task learning approaches.

Hand crafted and deep learning input image enhancement techniques Since echocardiograms are by nature fuzzy images characterized by dark-themed colours and shadows, an interesting field to explore would be the possibility to enhance these inputs beyond vanilla normalization and simple transformations.

Hard attention mechanism and temporal attention While we have investigated the use of soft attention mechanism in terms of the visual non-local blocks in the Convolutional network, applying hard attention mechanism (or an indirect form of it) might yield even better results due to the fuzzy nature of echocardiograms distorting the soft attention mechanism calculations. In addition, it might be worth exploring temporal attention as well to study the inter-frame effect and the weight each frame presents to the final decision.

Variability in terms of model's choice, loss functions, sequence length, attention blocks location.. Throughout the course of this work, choices regarding the model's architecture, loss function, optimal sequence length, and feature maps' latent dimensions were taken based on the current trends and practices in the field. Exploring these variables more closely (especially the main model's architecture) might be of high value.

Appendix A

Echocardiogram Cycle generative adversarial networks (GANs)

This Appendix section will acknowledge a small side project that was implemented towards the earlier stages of this thesis journey but was later discontinued for data unreliability purposes as well as the Leask Labs providing us with additional data towards the end of this project.

When the Leask Lab team presented us with the original raw echocardiograms dataset mentioned in Chapter 3, they also provided its speckled video output version from General Electric (GE) software. The way the GE software works is that it reads the original echocardiogram probe output, then applies edge-detecting algorithms to detect the edge of the aorta and highlight them by drawing colored speckles along the aortic wall.

An example of raw vs speckled video frames can be seen in figure A.1.

When we acquired the mentioned dataset, we had some patient references with just the raw echo videos, some with just the speckled GE output videos, and some with both. We, therefore, sought a method where we might generate the missing raw echocardiogram videos and augment the size of our dataset. Cycle GANs for unpaired image-to-image translation [58] were an attractive option that exactly fitted our needs.

Cycle GANs learn to translate unpaired images from source X to target Y and vice versa. Therefore, they not only learn the mapping G: X $\rightarrow$ Y but also its corresponding inverse mapping F: Y $\rightarrow$

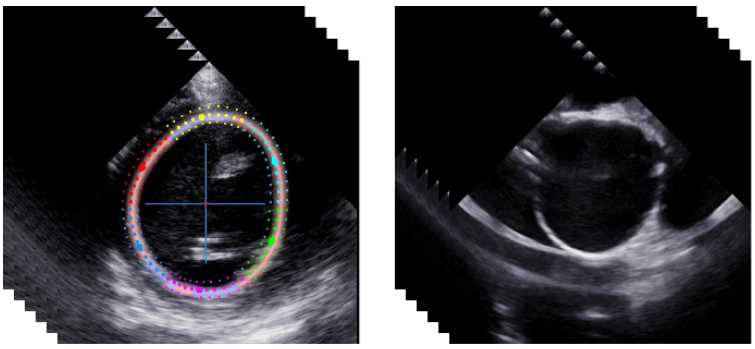


Figure A.1: Speckled GE software output videos(left) compared to raw echocardiogram videos(right).

X. With this implementation, we could utilize the unspeckled raw videos and the speckled videos and exploit the fact that we mostly didn't have paired data entities between the two domains.

According to [58], the Architecture revolves around two generative mappings G: X $\rightarrow$ Y and F: Y $\rightarrow$ X, two adversarial discriminators D_X and D_Y. The full cycle GANs objective, as presented in equation 5, can be summarised as minimizing the adversarial loss of the G: X $\rightarrow$ Y and F: Y $\rightarrow$ X mappings as well as minimizing the cycle consistency losses between the two to prevent the learned mappings from contradicting each other with λ being a flexible factor that controls the relative importance of the two objectives.

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F) \qquad (5)$$

For the implementation, we mostly relied on the authors' PyTorch repository for the implementation details and hyperparameters. We started by loading the generative models' pre-trained weights from the official paper's experiment on horse-zebra and zebra-horse image translation tasks. We normalized and resized the input image frames to 256×256 to be compatible with the model's parameters. We set both the adversarial losses, L_{GAN}, to the mean squared error loss while using the L1 loss for the cycle consistency loss. In addition, λ is set to 10, Adam optimizer is used with a learning rate of 0.0002 and the model is set to train on 200 epochs.

As one would expect with GAN models, the images generated varied along the epochs. A selection of the artificially generated images along epochs 12+ can be seen in figure A.2 as compared to the original images of the corresponding speckled/unspeckled domains.

We visually assessed the speckled imaging results to be of acceptable quality to the naked eye with few unreliable images that could be manually filtered out. However, when it came to the generated unspeckled images, almost half of them had unexpected colour variation filters applied which was probably caused by the colored original speckled image input. Therefore, a high-quality image search meant visually cherry-picking between the enormous project output. Moreover, we noticed a trend of thick white edges around the artificially generated unspeckled images (probably due to the GAN not exactly generating the original aortic wall because of the speckles position obscuring it in the speckled image input). This thick aortic wall might cause the imaging input along with its predefined risk label to be medically unreliable since a few millimeters in the aortic diameter might critically alter the diagnostic risk decision.

In addition, towards the end of this project, we were provided with some of the missing videos mentioned above, which caused the efforts and time to be further invested unnecessary.

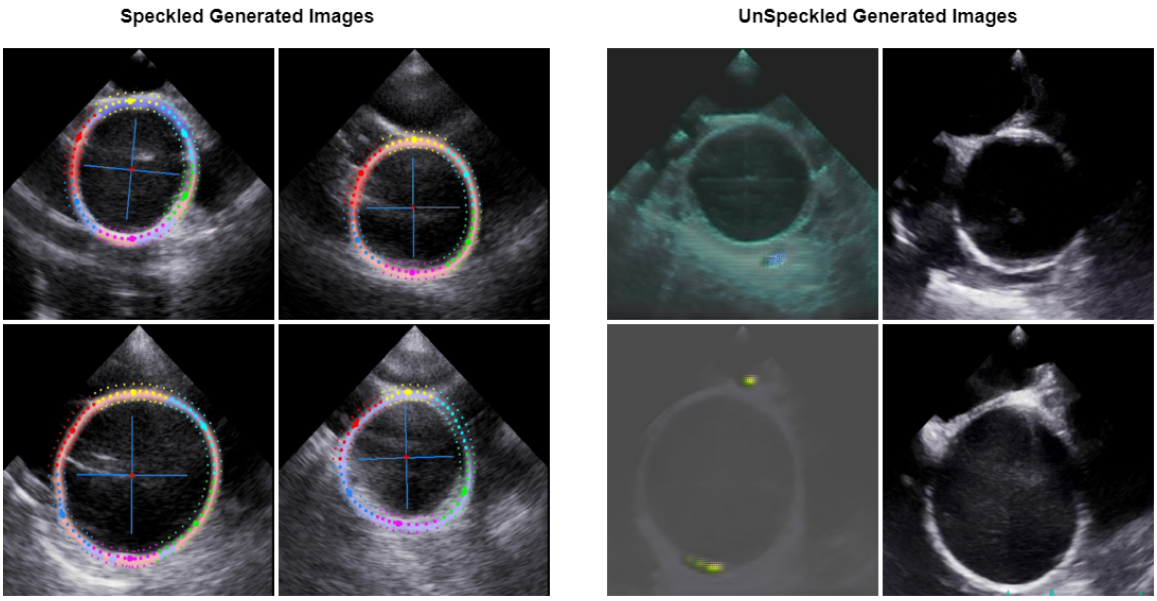


Figure A.2: Artificially Generated speckled images(left) Artificially Generated unspeckled images(right).

Bibliography

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Neural Information Processing Systems*, 2018.

[2] Sercan Ö. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *ArXiv*, abs/1908.07442, 2021.

[3] Sarkhan Badirli, Xuanqing Liu, Zhengming Xing, Avradeep Bhowmik, and S. Sathiya Keerthi. Gradient boosting neural networks: Grownet. *ArXiv*, abs/2002.07971, 2020.

[4] Nicolas Ballas, L. Yao, Christopher Joseph Pal, and Aaron C. Courville. Delving deeper into convolutional networks for learning video representations. *CoRR*, abs/1511.06432, 2016.

[5] Yoshua Bengio, Patrice Y. Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5 2:157–66, 1994.

[6] Abi Berger. Magnetic resonance imaging. *BMJ*, 324(7328):35, January 2002.

[7] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A non-local algorithm for image denoising. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2:60–65 vol. 2, 2005.

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[9] Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv*, abs/1412.3555, 2014.

[10] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634, 2015.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021.

[12] Alexander Emmott, Haitham Alzahrani, Mohammed Alreshidan, Judith Therrien, Richard L. Leask, and Kevin Lachapelle. Transesophageal echocardiographic strain imaging predicts aortic biomechanics: Beyond diameter. *The Journal of Thoracic and Cardiovascular Surgery*, 156:503–512.e1, 2018.

[13] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

[14] Hongwei Ge, Zehang Yan, Wenhao Yu, and Liang Sun. An attention mechanism based convolutional lstm network for video action recognition. *Multimedia Tools and Applications*, pages 1–24, 2019.

[15] GE HealthCare. Ge vivid 7.

[16] Rohit Girdhar, Deva Ramanan, Abhinav Kumar Gupta, Josef Sivic, and Bryan C. Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3165–3174, 2017.

[17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[18] Yu. V. Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *ArXiv*, abs/2106.11959, 2021.

[19] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 3154–3160, 2017.

[20] Hussein Hazimeh, Natalia Ponomareva, Petros Mol, Zhenyu Tan, and Rahul Mazumder. The tree ensemble layer: Differentiability meets conditional computation. *ArXiv*, abs/2002.07772, 2020.

[21] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[22] Loren F. Hiratzka, George L. Bakris, Joshua A. Beckman, Robert M. Bersin, V F Carr, Donald E Casey, Kim Eagle, Luke K. Hermann, Eric M. Isselbacher, Ella A. Kazerooni, Nicholas T. Kouchoukos, Bruce Whitney Lytle, Dianna M. Milewicz, David L. Reich, Souvik Sen, J A Shinn, Lars G. Svensson, and David M. Williams. 2010 accf/aha/aats/acr/asa/sca/scai/sir/sts/svm guidelines for the diagnosis and management of patients with thoracic aortic disease: a report of the american college of cardiology foundation/american heart association task force on practice guidelines, american association for thoracic surgery, ame. *Circulation*, 121 13:e266–369, 2010.

[23] Julia Höhn, Eva Krieghoff-Henning, Tanja Jutzi, Christof von Kalle, Jochen Sven Utikal, Friedegund Meier, Frank Friedrich Gellrich, Sarah Hobelsberger, Axel Hauschild, Justin Gabriel Schlager, Lars Einar French, Lucie Heinzerling, Max Schlaak, Kamran Ghoreschi, Franz J. Hilke, Gabriela Poch, Heinz Kutzner, Markus Vincent Heppt, Sebastian Haferkamp, Wiebke Sondermann, Dirk Schadendorf, Bastian Schilling, Matthias Goebeler, Achim Hekler, Stefan Fröhling, Daniel B. Lipka, Jakob Nikolas Kather, Dieter Krahl, Gerardo Ferrara, Sarah Haggenmüller, and Titus Josef Brinker. Combining cnn-based histologic whole slide image analysis and patient data to improve skin cancer classification. *European journal of cancer*, 149:94–101, 2021.

[24] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:2011–2023, 2020.

[25] Quantom Medical Imaging. Abdominal aortic aneurysm – taking control of the time bomb, Jun 2021.

[26] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *Proceedings of the 22nd ACM international conference on Multimedia*, 2014.

[27] Gunnar Johansson, Ulf Markström, and Jesper Swedenborg. Ruptured thoracic aortic aneurysms: a study of incidence and mortality rates. *Journal of vascular surgery*, 21 6:985–8, 1995.

[28] Manu Joseph. Pytorch tabular: A framework for deep learning with tabular data. *ArXiv*, abs/2104.13638, 2021.

[29] Kaggle.com. 3d convolutions : understanding + use case, 2022.

[30] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[31] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *ArXiv*, abs/1705.06950, 2017.

[32] Lauren Kennedy. A machine learning approach to an integrated risk assessment for thoracic aortic aneurysms. Master's thesis, McGill University, 2022.

[33] Wing Lam and Dudley John Pennell. Imaging of the heart: historical perspective and recent advances. *Postgraduate Medical Journal*, 92:104 – 99, 2015.

[34] Frank Lindemann, Sabrina Oebel, Ingo Paetsch, Arash Arya, Nikolaos Dagres, Sergio Richter, Borislav Dinov, Sebastian Hilbert, Susanne Loebe, Clara Stegmann, Michael Doering, Andreas Bollmann, Gerhard Hindricks, and Cosima Jahnke. Clinical utility of cardiovascular magnetic resonance imaging in patients with implantable cardioverter defibrillators presenting with electrical instability or worsening heart failure symptoms. *Journal of Cardiovascular Magnetic Resonance*, 22, 2020.

[35] Majid Maleki and Maryam Esmaeilzadeh. The evolutionary development of echocardiography. *Iran. J. Med. Sci.*, 37(4):222–232, December 2012.

[36] Vascular Institute of Michigan. Thoracic aneurysm.

[37] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization.

[38] Christopher Olah. Understanding LSTM networks. https://colah.github.io/posts/2015-08-Understanding-LSTMs/, 2015. Accessed: 2022-11-14.

[39] Linda A. Pape, Thomas Tsai, Eric M. Isselbacher, Jae Keun Oh, Patrick T O'Gara, Artur Evangelista, Rossella Fattori, Gabriel Meinhardt, Santi Trimarchi, Eduardo Bossone, Toru Suzuki, Jeanna V. Cooper, James B. Froehlich, Christoph A. Nienaber, and Kim Eagle. Aortic diameter ≥ 5.5 cm is not a good predictor of type a aortic dissection: Observations from the international registry of acute aortic dissection (irad). *Circulation*, 116:1120–1127, 2007.

[40] Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for deep learning on tabular data. *ArXiv*, abs/1909.06312, 2020.

[41] Zhen Qin, Le Yan, Honglei Zhuang, Yi Tay, Rama Kumar Pasumarthi, Xuanhui Wang, Michael Bendersky, and Marc-Alexander Najork. Are neural rankers still outperformed by gradient boosted decision trees? In *ICLR*, 2021.

[42] Vincenzo Rampoldi, Santi Trimarchi, Kim Eagle, Christoph A. Nienaber, Jae Keun Oh, Eduardo Bossone, Truls Myrmel, Giuseppe Massimo Sangiorgi, Carlo de Vincentiis, Jeanna V. Cooper, Jianming Fang, Dean G. Smith, Thomas Tsai, Arun Raghupathy, Rossella Fattori,

Udo Sechtem, Michael G. Deeb, Thoralf M. Sundt, and Eric M. Isselbacher. Simple risk models to predict surgical mortality in acute type a aortic dissection: the international registry of acute aortic dissection score. *The Annals of thoracic surgery*, 83 1:55–61, 2007.

[43] Ronald A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7:17 – 42, 2000.

[44] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[45] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *ArXiv*, abs/1511.04119, 2015.

[46] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[47] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015.

[48] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[49] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[50] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[51] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP@EMNLP*, 2018.

[52] X. Wang, Ross B. Girshick, Abhinav Kumar Gupta, and Kaiming He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[53] Judson B. Williams, Eric D. Peterson, Yue Zhao, Sean M. O'Brien, Nicholas D. Andersen, D Craig Miller, Edward P. Chen, and G. Chad Hughes. Contemporary results for proximal aortic replacement in north america. *Journal of the American College of Cardiology*, 60 13:1156–62, 2012.

[54] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and X. Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. *Proceedings of the 23rd ACM international conference on Multimedia*, 2015.

[55] Ke Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[56] Zhongwen Xu, Yi Yang, and Alexander Hauptmann. A discriminative cnn video representation for event detection. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1798–1807, 2015.

[57] L. Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Joseph Pal, H. Larochelle, and Aaron C. Courville. Describing videos by exploiting temporal structure. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4507–4515, 2015.

[58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[59] Yousong Zhu, Chaoyang Zhao, Haiyun Guo, Jinqiao Wang, Xu Zhao, and Hanqing Lu. Attention couplenet: Fully convolutional attention coupling network for object detection. *IEEE Transactions on Image Processing*, 28:113–126, 2019.