A TECHNICAL SEMINAR REPORT ON

**DEEP LEARNING FOR CLASSIFICATION AND LOCALIZATION OF COVID-19 MARKERS IN POINT-OF-CARE LUNG ULTRASOUND**

A dissertation submitted in partial fulfilment of the Requirements for the award of the degree of

## BACHELOR OF TECHNOLOGY

in

**INFORMATION TECHNOLOGY**

***Submitted by***

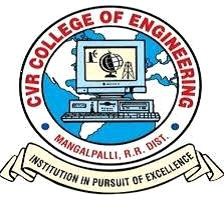
# B Neeraj Kumar (18B81A1225)

***Under the esteemed guidance of***

**A. Srichandana**

Senior Assistant Professor, IT Department

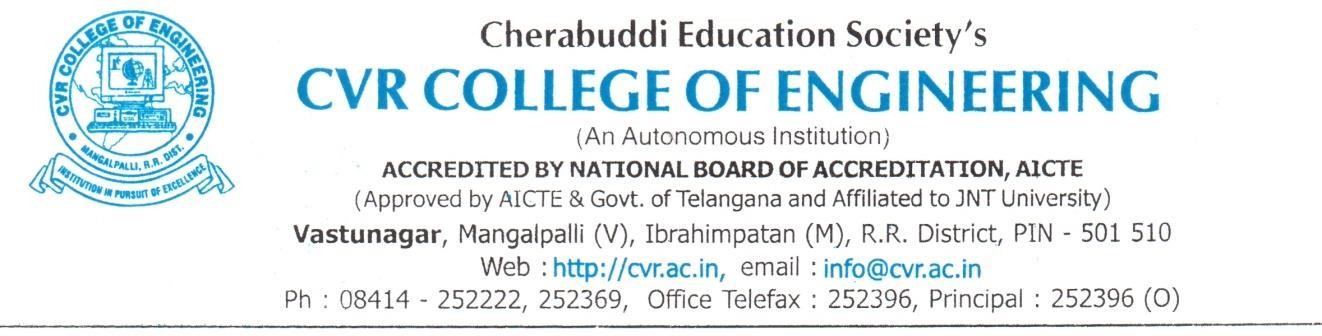
CVR College of Engineering



## DEPARTMENT OF INFORMATION TECHNOLOGY CVR COLLEGE OF ENGINEERING

ACCREDITED BY NBA, AICTE & Affiliated to JNTU-H Vastunagar, Mangalpally (V), Ibrahimpatnam (M), R.R. District, PIN-501 510

2021-2022



## DEPARTMENT OF INFORMATION TECHNOLOGY CERTIFICATE

This is to certify that the Technical Seminar Report entitled **“Deep Learning for Classification and Localization of COVID-19 Markers in Point-of-Care Lung Ultrasound”** is a bonafide work

done and submitted by **B Neeraj Kumar (18B81A1225)** during the academic year 2021-2022, in partial fulfilment of requirement for the award of Bachelor of Technology degree in Information Technology from Jawaharlal Nehru Technological University Hyderabad, is a bonafide record of work carried out by him under my guidance and supervision.

Certified further that to my best of the knowledge, the work in this dissertation has not been submitted to any other institution for the award of any degree or diploma.

**INTERNAL GUIDE HEAD OF THE DEPARTMENT**

**A. Srichandana Dr. Bipin Bihari Jayasingh**

Senior Assistant Professor, IT Department Professor, IT Department

**TECHNICAL SEMINAR COORDINATOR**

**C.V.S. Satyamurthy**

Associate Professor, IT Department



## ABSTRACT

Deep learning (DL) has proved successful in medical imaging, and, in the wake of the recent COVID-19 pandemic, some works have started to investigate DL-based solutions for the assisted diagnosis of lung diseases. While existing works focus on CT scans, this paper studies the application of DL techniques for the analysis of lung ultrasonography (LUS) images. Specifically, it presents a novel fully annotated dataset of LUS images collected from several Italian hospitals, with labels indicating the degree of disease severity at a frame-level, video-level, and pixel-level (segmentation masks). Leveraging these data, and introduces several deep models that address relevant tasks for the automatic analysis of LUS images. In particular, it present a novel deep network, derived from Spatial Transformer Networks, which simultaneously predicts the disease severity score associated to a input frame and provides localization of pathological artefacts in a weakly-supervised way. Furthermore, it introduces a new method based on norms for effective frame score aggregation at a video-level. Finally, the benchmark state of the art deep models for estimating pixel-level segmentations of COVID-19 imaging biomarkers. Experiments on the proposed dataset demonstrate satisfactory results on all the considered tasks, paving the way to future research on DL for the assisted diagnosis of COVID-19 from LUS data.

## ACKNOWLEDGEMENT

The satisfaction of completing this report would be incomplete without mentioning my gratitude towards all the people who have supported me. Constant guidance and encouragement have been instrumental in the completion of this technical seminar.

First and Foremost, I would like to thank the Chairman, Principal, Vice Principal for availing infrastructural facilities to complete this technical seminar report in time.

I offer my sincere gratitude to my internal guide **A. Srichandana,** Senior Assistant Professor, IT Department, CVR College of Engineering for his immense support, timely co-operation and valuable advice throughout the course of the report work.

I am also thankful to **C.V.S. Satyamurthy,** Technical Seminar Coordinator, Associate Professor, IT Department, CVR College of Engineering for his supportive guidelines and for having provided the necessary help for carrying forward this technical seminar without any obstacles and hindrances.

I also thank the **Seminar Review Committee Members** for their valuable suggestions.

## DECLARATION

I hereby declare that the technical report entitled “**Deep Learning for Classification and Localization of Covid-19 Markers in POINT-OF-CARE Lung Ultrasound**” is an original work done and submitted to IT Department, CVR College of Engineering, affiliated to Jawaharlal Nehru Technological University Hyderabad, Hyderabad in partial fulfilment of the requirement for the award of Bachelor of Technology in **Information Technology** and it is a record of bonafide work carried out by me under the esteemed guidance of **A. Srichandana,** Senior Assistant Professor, IT Department**.**

I further declare that the work reported in this technical seminar report has not been submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other Institute or University.

Signature of the Student (B Neeraj Kumar)

(18B81A1225)

## LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **Fig. No.** | **Title** | **Pg. No.** |
| **1** | **Overview of the different tasks considered in this work. Given a LUS image sequence, we propose approaches for: (orange) prediction of the disease severity score for each input frame and weakly supervised localization of pathological patterns; (pink) aggregation of frame-level scores for producing predictions on videos; (green) estimation of segmentation masks indicating pathological artifacts.** | **2** |
| **2** | **The distribution of the probes and the scores of frames grouped by hospital and overall statistics** | **5** |
| **3** | **Illustration of the architecture for frame-based score prediction. An STN modeled by Φstn predicts two transformations θ1 and θ2 which are applied to the input image producing two transformed versions x1 and x2 that localize pathological artifacts. The feature extractor Φcnn is applied to x1 to generate the final prediction.** | **7** |
| **4** | **Four examples of B-mode input image frames (first column), their annotations (second column) including COVID-19 biomarkers (moderate/score 2: orange, severe/score 3: red), and signs of healthy lung (blue). The corresponding semantic segmentations and contours of COVID-19 markers by deep learning are given in the third and fourth column, respectively** | **15** |
| **5** | **Two examples (A, B) of class uncertainty in the segmentations, showing B-mode input image frames (first column), annotations, including COVID-19 biomarkers, the corresponding semantic segmentations by deep learning , and pixel-level COVID-19 class uncertainty by MC-dropout .** | **16** |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table. No** | **Title** | **Pg. No** |
| **1** | **F1 SCORES (%) FOR THE FRAME-BASED CLASSIFICATION UNDER DIFFERENT EVALUATION SETTINGS. SETTING 1 REPRESENTS EVALUATION ON THE FULL TEST SET, SETTING 2 REPRESENTS THE ANALYSIS ON THE TEST SET WITH DROPPED TRANSITION FRAMES AND SETTING 3 REPRESENTS THE ANALYSIS ACCOUNTING FOR INTER-DOCTOR AGREEMENT. THE BASELINE FOR THIS SETTING IS PROVIDED BY THE EVALUATION ON THE SET OF TEST SEQUENCES WITH VIDEO-LEVEL ANNOTATIONS (VIDEO ANN.). BEST AND SECOND BEST F1 SCORES (%) ARE IN BOLD AND UNDERLINES, RESPECTIVELY** | **12** |
| **2** | **MEAN AND STANDARD DEVIATION OF WEIGHTED F1 SCORE, PRECISION AND RECALL COMPUTED OVER THE FIVE CROSS VALIDATION FOLDS, FOR THE PROPOSED VIDEO-BASED CLASSIFICATION METHOD AND BASELINES** | **14** |
| **3** | **CONFUSION MATRICES (%) FOR THE PROPOSED VIDEO-BASED CLASSIFICATION METHOD AND BASELINES** | **14** |

**TABLE OF CONTENTS**

S. No Topic Pg. No 1 Introduction…………………………………………………….………………………. 1

2 Motivation…………………………………….………………………………………… 2

3 Literature Survey……………………………………………………………………... 3

4 Methodology……………….………………………………………………………….... 4

5 Results and Analysis…………………………….......................................................... 8

[Conclusion……………………………………………………......................................... 11](#_TOC_250001)

[References………………………………………………….............................................. 14](#_TOC_250000)

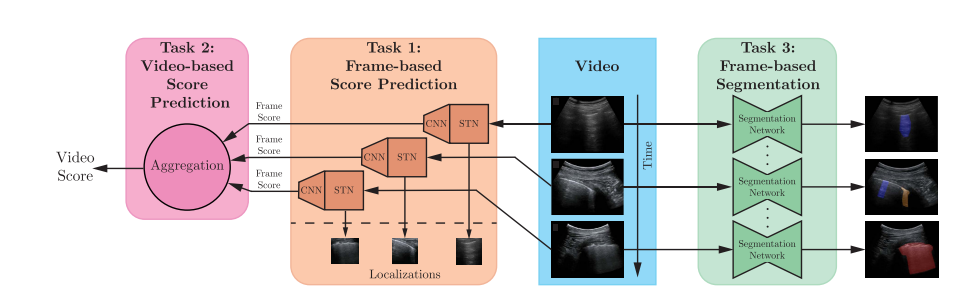
# Chapter 1 INTRODUCTION

The rapid global SARS-CoV-2 outbreak resulted in a scarcity of medical equipment. The threat posed by COVID-19 led the World Health Organization (WHO) to declare the COVID-19 pandemic by March 2020.

Coronaviruses are a group of highly diverse, enveloped, positive-sense, single-stranded RNA viruses and are widely spread in birds and mammals. Sometimes these viruses infect humans, causing mild to moderate respiratory diseases. Before SARS-CoV-2, two coronaviruses were known to cause severe human disease: SARS-CoV, which causes Severe Acute Respiratory Syndrome (SARS); and MERS-CoV, which causes Middle East Respiratory Syndrome (MERS). However, in contrast to SARS and MERS, the symptom onset for COVID-19 is significantly larger, or it may appear in a mild form, allowing infection spread by asymptomatic patients, which in turn has led to the current pandemic.

Although the WHO has emphasized the need for massive testing and contact tracing to better tackle the pandemic, not all countries have the required laboratory infrastructure and reagents to effectively address this task. Additionally, getting results from some of these tests may take a couple of days, leading to non-confirmed COVID-19 patients with mild or no symptoms to further spread the disease while waiting for the test results.

With the rise of deep learning techniques, medical imagery has increasingly claimed attention for the computed assisted analysis of pulmonary conditions. Automated analysis of Computed Tomography (CT) scans, has enabled the identification of malignant nodules. Radiographic analysis, in turn, has also obtained fair results in the detection of tuberculosis sign, as well as other multiple cardiothoracic abnormalities.



*Fig 1: Overview of the different tasks considered in this work. Given a LUS image sequence, we propose approaches for: (orange) prediction of the disease severity score for each input frame and weakly supervised localization of pathological patterns; (pink) aggregation of frame-level scores for producing predictions on videos; (green) estimation of segmentation masks indicating pathological artifacts.*

Lung ultrasound (LUS) is a portable, easy to disinfect, low cost and non-invasive medical imaging tool that can be used to identify lung diseases as an alternative to CT and X-ray. Computer-assisted analysis of lung ultrasound imagery is a relatively recent approach that has shown great potential for diagnosing pulmonary conditions. In the context of the current COVID-19 pandemic, by the time of writing this paper, three works have proposed the use of Deep Learning (DL) models for computer-assisted analysis of COVID-19 LUS imaging. In this sense, in this work it evaluates and compare the performance of several deep-learning techniques for the identification of COVID-19 infections from lung ultrasound imagery.

The insufficient testing capacity in most countries has therefore spurred the need and search for alternative methods that enable diagnosis of COVID-19. In addition, the accuracy of the current lab test, reverse transcription polymerase chain reaction (RT-PCR) arrays, remains highly dependent on swab technique and location.

COVID-19 pneumonia can rapidly progress into a very critical condition. Examination of radiological images of over 1,000 COVID-19 patients showed many acute respiratory distress syndrome (ARDS)-like characteristics, such as bilateral, and multi-lobar glass ground opacifications (mainly posteriorly and/or peripherally distributed),. As such, chest computed tomography (CT) has been coined as a potential alternative for diagnosing COVID-19 patients. While RT-PCR may take up to 24 hours and requires multiple tests for definitive results, diagnosis using CT can be much quicker. However, use of chest CT comes with significant drawbacks: it is costly, exposes patients to radiation, requires extensive cleaning after scans, and relies on radiologist interpretability.

Recently, automatic image analysis by machine and deep learning (DL) methods have already shown promise for reconstruction, classification, regression and segmentation of tissues using ultrasound image. It describes the use of DL to assist clinicians in detecting COVID-19 associated imaging patterns on point-of-care LUS. In particular, it tackles three different tasks on LUS imaging (Fig. 1): frame-based classification, video-level grading and pathological artifact segmentation. The first task consists of classifying each single frame of a LUS image sequence into one of the four levels of disease severity, defined by the scoring system in. Video-level grading aims to predict a score for the entire frame sequence based on the same scoring scale. Segmentation instead comprises pixel-level classification of the pathological artifacts within each frame.

This paper advances the state of the art in the automatic analysis of LUS images for supporting medical personnel in the diagnosis of COVID-19 related pathologies in many directions. (1) Proposed an extended and fully-annotated version of the ICLUS-DB database. The dataset contains labels on the 4-level scale proposed in, both at frame and video-level. Furthermore, it includes a subset of pixel-level annotated LUS images useful for developing and assessing semantic segmentation methods. (2) Introduced a novel deep architecture which permits to predict the score associated to a single LUS image, as well as to identify regions containing pathological artifacts in a weakly supervised manner. The network leverages Spatial Transformers Network (STN) and consistency losses to achieve disease pattern localization and from a soft ordinal regression loss for robust score estimation.

# Chapter 2 MOTIVATION

In the last few years, Lung Ultrasound (LUS) imaging has been proposed as an alternative to the use of CT or X-ray for screening and follow-up of lung diseases. For instance, it has been suggested that lung visualization through ultrasound imaging effectively replaces physical auscultation with stethoscopes. Moreover, when used correctly, LUS imaging could even help to reduce infections between patients and medical staff.

Recent medical correspondence has pointed out the advantages of using LUS imaging as a tool for early diagnosis and follow-up of COVID-19 patients. Some works highlight the benefits of using LUS in the context of the COVID-19 pandemic, especially considering its portability, accessibility, no radiation, ease of disinfection (e.g., using disposable caps of the ultrasound probes), and low cost.

This In the last few years, Lung Ultrasound (LUS) imaging has been proposed as an alternative to the use of CT or X-ray for screening and follow-up of lung diseases. For instance, it has been suggested that lung visualization through ultrasound imaging effectively replaces physical auscultation with stethoscope. Moreover, when used correctly, LUS imaging could even help to reduce infections between patients and medical staff.

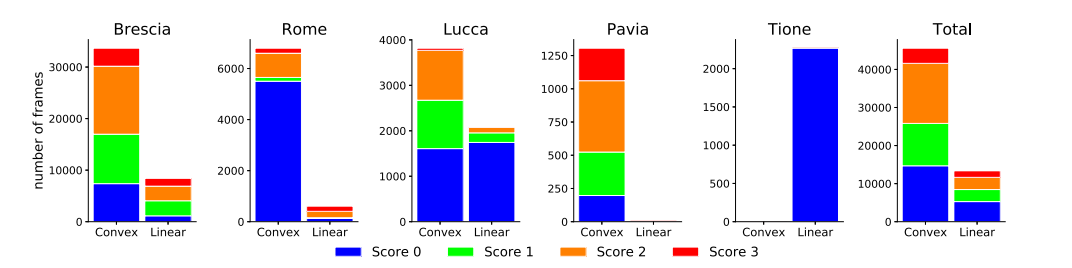
Recent medical correspondence has pointed out the advantages of using LUS imaging as a tool for early diagnosis and follow-up of COVID-19 patient. Some works highlight the benefits of using LUS in the context of the COVID-19 pandemic, especially considering its portability, accessibility, no radiation, ease of disinfection (e.g., using disposable caps of the ultrasound probes), and low cost.

# Chapter 3 LITERATURE SURVEY

Earliest studies involving computer-assisted LUS imaging analysis were based on image processing techniques. These works were aimed at segmentation (identification) of artifacts in LUS images. Most of the identified artifacts are the so-called B-lines, which are associated with the disease presence. Subsequent works applied the typical training-testing approach of classical Machine Learning algorithms (ML) and Deep Learning (DL) networks. Taking into account the strong correlation between LUS and CT imaging findings in COVID-19 patients and the advantages of LUS equipment, it makes sense to develop tools for COVID-19 screening and diagnosis by means of computer-assisted analysis of LUS imaging.

DL has proven to be successful in a multitude of computer vision tasks ranging from object recognition and detection to semantic segmentation. Motivated by these successes, more recently, DL has been increasingly used in medical applications, e.g, for biomedical image segmentation or pneumonia detection from chest X-ray. These seminal works indicate that, with the availability of data, DL can lead to the assistance and automation of preliminary diagnoses which are of tremendous significance in the medical community.

In the wake of the current pandemic, recent works have focused on the detection of COVID-19 from chest CT. A U-Net type network is used to regress a bounding box for each suspicious COVID-19 pneumonia region on consecutive CT scans, and a quadrant-based filtering is exploited to reduce possible false positive detections.



*Fig 2: The distribution of the probes and the scores of frames grouped by hospital and overall statistics*

Differently, A threshold-based region proposal is first used to retrieve the region of interests (RoIs) in the input scan and the Inception network is exploited to classify each proposed RoI. Similarly, a VNET-IR-RPN model pre-trained for pulmonary tuberculosis detection is used to propose RoIs in the input CT and a 3D version of Resnet-18 is employed to classify each RoI. However, very few works using DL on LUS images can be found in the literature. A classification and weakly-supervised localization method for lung pathology is described in. Based on the same idea, in a frame-based classification and weakly-supervised segmentation method is applied on LUS images for COVID-19 related pattern detection.

Here, Efficientnet is trained to recognize COVID-19 in LUS images, after which class activation maps (CAMs) are exploited to produce a weakly-supervised segmentation map of the input image.

First, while in CAMs are used for localization, in this work it exploits STN to learn a weakly-supervised localization policy from the data (i.e. not exploiting explicit labelled locations but inferring it from simple frame-based classification labels).

Second, while in a classification problem is solved, it focuses on ordinal regression, predicting not only the presence of COVID-19 related artifacts, but also a score connected to the disease severity.

Third, move a step forward compared to all previous methods by proposing a video-level prediction model built on top of the frame-based method. Finally, propose a simple yet effective method to predict segmentation masks using an ensemble of multiple state-of-the-art convolutional network architectures for image segmentation.

Additionally, the model’s predictions are accompanied with uncertainty estimates to facilitate interpretation of the results First, while in CAMs are used for localization, in this work to exploit STN to learn a weakly-supervised localization policy from the data (i.e. not exploiting explicit labelled locations but inferring it from simple frame-based classification labels). Second, while in a classification problem is solved, it focus on ordinal regression, predicting not only the presence of COVID-19 related artifacts, but also a score connected to the disease severity. Move a step forward compared to all previous methods by proposing a video-level prediction model built on top of the frame-based method. Finally, propose a simple yet effective method to predict segmentation masks using an ensemble of multiple state-of-the-art convolutional network architectures for image segmentation.

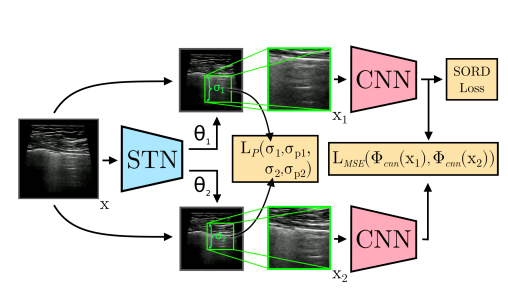
Additionally, the model’s predictions are accompanied with uncertainty estimates to facilitate interpretation of the results.

# Chapter 4 METHODOLOGY

It tackles several challenges towards the development of automatic approaches for supporting medical personnel in the diagnosis of COVID-19 related pathologies (see Fig. 1). In particular, following the COVID-19 LUS scoring system at present a novel deep architecture which automatically predicts the pathological scores associated to all frames of a LUS image sequence (Section IV-A) and optimally fuse them to produce a disease severity score at video-level (Section IV-B). It also shows that the proposed model automatically identifies regions in an image which are associated to pathological artifacts without requiring pixel-level annotation. Finally, to further improve the accuracy in the automatic detection of disease-related patterns, and consider a scenario where frames are provided with pixel-level annotations and to propose a segmentation model derived from a state of the art convolutional network architecture (Section IV-C). In the following, describe the proposed deep learning models.

1. **Frame-Based Score Prediction:**

With the purpose of supporting medical personnel in the analysis of LUS images, It introduces an approach for predicting the presence or the absence of a pathological artifact in each frame of a LUS image sequence and for automatically assessing the severity score of the disease related to such patterns according to the COVID-19 LUS scoring system. And are also interested in the spatial localisation of a pathological artifact in the frame without assuming any annotation about such artifact positions in a frame.



*Fig 3: Illustration of the architecture for frame-based score prediction. An STN modeled by Φstn predicts two transformations θ1 and θ2 which are applied to the input image producing two transformed versions x1 and x2 that localize pathological artifacts. The feature extractor Φcnn is applied to x1 to generate the final prediction.*

Fig. 3 shows an overview of the proposed deep architecture. In the context of deep learning the generalization capability of a network is of critical importance. To this end, data augmentation has shown to be very effective in improving the performance of a network. Previous works showed that augmenting a dataset composed of LUS images can drastically improve the ability of the network to discriminate healthy and ill patients. Another way to achieve robust predictions is to enforce some consistency between two perturbed versions (colour jitter, dropout, etc.) of the same image. This makes the network produce smoothed predictions by attending to the salient features in an image. Inspired by this idea, and propose to use STN to produce two different crops from a single image and enforce the predictions of the network to be similar. And named the approach Regularised Spatial Transformer Networks (Reg-STN).

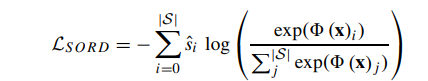
The weak localization is achieved through the use of Spatial Transformer Networks (STN. The use of STN stems from the fact that most of the pathological artifacts are concentrated in a relatively small area of the image, and, hence the entire image should not be considered by the network to make predictions. The consistency loss is defined on the network representations as following:



Unfortunately, LMSE coupled with learnable affine transformations produces degenerate solutions in which the localization network of the STN learns to output identical parameter for the affine transformations. In fact, it is enough to impose θ1 = θ2 to minimize LMSE. To prevent this pathological behaviour of the network, they enforce a prior on the parameters of the transformations. In particular, they stimulate the localization network to produce reasonably scaled patches by minimizing |σ − σp|, where σp is a fixed prior. Now, in order to enable the STN into yielding different parameters θ1 ≠ θ2, they simply choose σp1 ≠ σp2. Hence, a loss is defined as follows:



While ordinal regression can be implemented resorting on the traditional approach of decomposing the problem assuming a |S|-rank formulation, following it introduces a lightweight approach for Soft ORDinal regression (SORD).



Finally, the proposed Reg-STN model is trained end-to-end minimizing the following joint loss function:



1. **Video-Level Score Aggregation**

The identification of potentially pathological artifacts in LUS images is a crucial step towards diagnosis support. However, frame-based predictions should be turned into a single video-based score prediction in order to assess the pathological state of a patient. The video-based score aggregation problem can be formalized as follows. Let v = {xi}Mi=1, be a video, V be the set of videos of any length, and S the set of scores. The goal of video-level score prediction is learning a mapping ♆: V → S.

In principle the mapping ♆could be obtained by taking the maximum score assigned to any frame of the current video because the identification of an artifact of score s in a frame implies that the patient has a severity level of at least s.

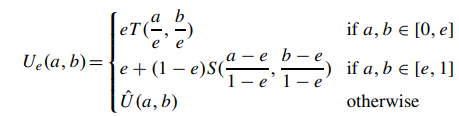
The video-based score aggregation problem can be formalized as follows. Let v = {xi}M i=1, be a video, V be the set of videos of any length, and S the set of scores. The goal of video-level score prediction is learning a mapping: V → S. The architecture is trained using the SORD loss, computed over the video-level prediction. The frame-based predictor outputs prediction scores with a distribution that differs between the training and the test set. In order not to overfit the video-based predictor on the training scores distribution, to completely separate the training sets of the frame-based and video-based predictor. However, this solution is not effective given the vast disproportion in the amount of supervision at the video and frame levels currently available in ICLUS-DB.

In designing the model ♆, consider the fact that it needs to operate in a low-data regime, where few videos are provided with annotations as in the current version of the ICLUS-DB. Inspired by the hard rule previously mentioned, and propose a simple strategy that combines frame-level predictions using a parameterized aggregation layer, i.e.:

♆(v) = ♆U (Φ(x1), . . . , Φ(xM ))

Here Φ is the frame-level mapping and ♆U is an aggregation function based on uninorm, which are a principled way to soften the hard rule. A uninorm U is a monotonic increasing, commutative and associative mapping from [0, 1]×[0, 1] to [0, 1] with neutral element e ∈ [0, 1]. This means that U(a, e) = U(e, a) = a for all a ∈ [0, 1]. If e = 1, U is fully non-compensatory (like taking the minimum between a and b), while it is fully compensatory if e = 0 (like the taking maximum).

Choosing e ∈ (0, 1) allows the uninorm to have a hybrid behaviour. Note that being associative, uninorms can be applied to an arbitrary number of inputs (e.g., U(a, b, c) = U(U(a, b), c)). The appropriate value for the neutral element e from data. The aggregation layer takes as input the sequence of frame-based prediction scores Φ (x), aggregates them along each dimension/score using a uninorm U and returns the softmax of the resulting aggregation as a video-based prediction. The layer has only four parameters, which are the neutral elements for each candidate score {0, 1, 2, 3}, and it is thus amenable to training with little supervision.

Any uninorm with neutral element e can be written as: 

for a certain choice of T , S and Û (a, b) such that min(a, b) ≤ Û (a, b) ≤ max(a, b). The functions T and S are called t-norm and t-conorm respectively, and model the non-compensatory and compensatory behaviour. Different choices for these functions lead to different uninorms. They found the product t-norm T (a, b) = ab (and corresponding t-conorm S(a, b) = a + b − ab) to be the most effective choice as it allows the gradient to flow the most. Concerning the function Û (a, b), common choices are min(a, b) and max(a, b), producing the so-called min-uninorms and max-uninorms respectively. They found min-uninorms to be the best choice in the setting (with respect to max(a, b) but also mean(a, b)), likely because of their fully non-compensatory behaviour in the area of highest discrepancy between frame-based predictions.

The frame-based predictor outputs prediction scores with a distribution that differs between the training and the test set. In order not to overfit the video-based predictor on the training scores distribution, They completely separate the training sets of the frame-based and video-based predictor.

The model was trained using an Adam optimizer with learning rate 10−2 without weight decay and with no learning rate scheduling. For each epoch, They computed the loss for each train video sequence and accumulate its gradients, performing a single optimization step at the end of each epoch. The model was trained for a maximum of 30 epochs and use the loss on the training set to define an early stopping strategy.

Note that the entire architecture including the frame-level component could be trained entirely end-to-end. However, this solution is not effective given the vast disproportion in the amount of supervision at the video and frame levels currently available in ICLUS-DB. Thus the model was trained through aggregation layer after freezing the weights of the frame-based architecture. Full end-to-end training combining frame-based and video-based supervision will be investigated in future work.

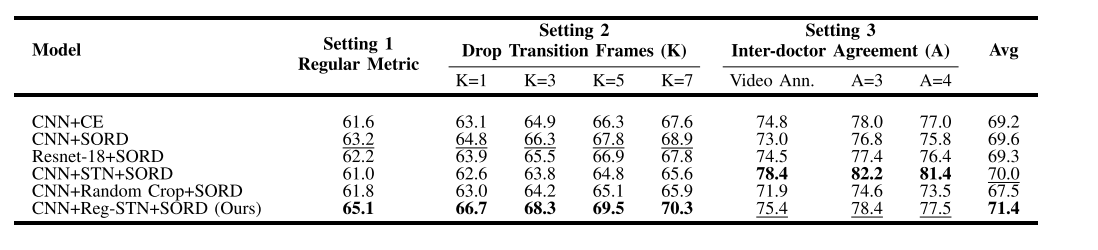
1. **Semantic Segmentation**

Let X = Ri×j and Y denote the input (i.e, the image space) and output (i.e. the segmentation masks) space, respectively. In the earlier presented frameworks for image- and video-based classification, the score set was defined as S = {0, 1, 2, 3}. For semantic segmentation, however distinguish five different scores, i.e. the four scores in S, complemented by the background (BG) score, assigned to pixels that were not annotated for showing markers associated with any of the classes in S. As such, Y = {0, 1, 2, 3, BG}i×j .

To model Ω, they compared several network architectures for end-to-end image segmentation, such as the vanilla U-Net, and the more recently proposed U-Net++, and Deeplabv3+. The baseline U-Net model has three encoding layer blocks, each comprising two convolutional layers with ReLU activations and one maxpool layer (pooling across 2, 2, and 5 pixels in both dimensions, respectively), a latent layer, and a mirrored decoder (where pooling is replaced by nearest neighbour upsampling).

They use skip connections between each layer block of the encoder and decoder. To mitigate overfitting, they apply dropout (p = 0.5) during training at the latent bottleneck of the model. The Unet++ variant leverages the first four encoder blocks of the ResNet50 model, to construct a latent space. The latent space is upsampled in the decoder stage by means of transpose 2D convolutional layers. The decoder contains residual blocks, and also exploits skip connections between (same-sized) hidden layer outputs in the ResNet50 encoder and the decoder. The Deeplabv3+ model similarly employs an encoder-decoder structure, where features are extracted using spatial pyramid pooling (i.e. pooling at different grid scales) and atrous convolutions, resulting in decoded segmentation maps with detailed object boundaries.

They adopt a pixel-wise categorical cross-entropy loss between he segmentation masks g(yn) and the model predictions Ŷ = Ω (h(xn)). Functions g(·), and h(·) are pre-processing transformations applied prior to training. Function h(·) comprises the resizing of all acquired B-mode images to 260 × 200 pixels, preserving the original aspect ratio of the scans by appropriate zero padding, and subsequent normalization between -1 and 1.



*Table 1: F1 SCORES (%) FOR THE FRAME-BASED CLASSIFICATION UNDER DIFFERENT EVALUATION SETTINGS. SETTING 1 REPRESENTS EVALUATION ON THE FULL TEST SET, SETTING 2 REPRESENTS THE ANALYSIS ON THE TEST SET WITH DROPPED TRANSITION FRAMES AND SETTING 3 REPRESENTS THE ANALYSIS ACCOUNTING FOR INTER-DOCTOR AGREEMENT. THE BASELINE FOR THIS SETTING IS PROVIDED BY THE EVALUATION ON THE SET OF TEST SEQUENCES WITH VIDEO-LEVEL ANNOTATIONS (VIDEO ANN.). BEST AND SECOND BEST F1 SCORES (%) ARE IN BOLD AND UNDERLINES, RESPECTIVELY*

During training, they are given a training set of N image-label pairs T = {(xn, yn)}Nn=1 where xn ∈ X and yn ∈ Y. The model parameters are learned by back-propagating the earlier defined categorical cross-entropy using the Adam optimizer (default settings), with a learning rate of 10−5. Training was stopped upon convergence of the training loss.

Each training batch consists of 32 B-mode images and their corresponding segmentation masks, which are balanced across patients and scores to avoid biases resulting from the length of the ultrasound scan (number of frames in a single video) and population-level distribution of scores. While these biases generally aid the overall accuracy, they hamper patient-level decision making across demographics.

To promote invariance to common LUS image transformations and thereby improve generalization at inference, each image-label pair is heavily manipulated on-line during training by a set of augmentation functions that were each activated on the image-label pair with a probability of 0.33. The set of augmentation functions, each applied with a randomly sampled strength bounded by a set maximum, consists of: affine transformations (translation (max. ±15%), rotation (max. ±15◦), scaling (max. ±45%), and shearing (max. ±4.5◦)), multiplication with a constant (max. ±45%), Gaussian blurring (σmax = 3 4 ), contrast distortion (max. ±45%), horizontal flipping (p = 0.5), and additive white Gaussian noise (σmax = 0.015). To further boost robustness and performance, they apply model ensembling and calculate the unweighted average over predicted softmax logits of the U-net, U-net++, and Deeplabv3+ models (all trained with data augmentation).

To allow for qualitatively assessment of the uncertainty of the predictions, they produce pixel-level estimates of model uncertainty by using Monte-Carlo (MC) dropout. During inference, they stochastically apply dropout in the latent space, yielding multiple point estimates of class predictions. The amount of variation in the resulting predictions, ultimately provides an indication of uncertainty for every pixel.

# Chapter 5

**RESULTS AND ANALYSIS**

1. Frame-Based Score Prediction:

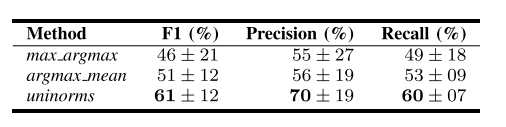
To evaluate the performance of the proposed frame-based scoring method and its constituent components, consider the following baselines: i) CNN trained with Cross Entropy loss (CE), ii) CNN trained with SORD, iii) Resnet-18 trained with SORD, iv) STN based CNN trained with SORD; v) CNN + Random Crop + SORD, a CNN trained on SORD with random crops rather than bounding boxes extracted by STN and vi) The proposed Reg-STN model. In Table I, they evaluate the performance of the method in terms of F1-score. The proposed Reg-STN trained with SORD beat the baseline models in most of the settings and is the second best in the remaining.

In Table I, they evaluate the performance of method in terms of F1-score. Since, the annotations in LUS images are quite subjective (see later) they also report results for two additional metrics, which are then defined as Setting 2 and Setting 3, respectively. The metrics are: i) Setting 1 considers the F1 score computed on the entire test set, ii) Setting 2 considers the F1 score computed on a modified version of the test set obtained by dropping, for each video, the K frames before and after each transition between two different ground truth scores, potentially removing ambiguous frames that present characteristics at the boundary between two classes, thereby allowing us to identify the impact of noisy labeling on the performance of the model; and iii) Setting 3, they drop the most challenging videos by using the inter-doctor agreement between the 5 independent video-level annotations. In practice, they only keep in the test set the videos with at least A doctors agreeing on the video-level annotations. For completeness, they report under Setting 3 also the scores obtained on the complete portion of the test set containing video-level annotations (Video Ann.).

As shown in Table I, the proposed Reg-STN trained with SORD beat the baseline models in most of the settings and is the second best in the remaining. On average, Reg-STN performs the best amongst all baselines. This proves the effectiveness of the proposed method for doing frame-based prediction for pathology detection in LUS images. The experiments were run on a RTX-2080 NVIDIA GPU. As for computational complexity, it takes ∼11 hours to train a CNN + Reg-STN + SORD model on this hardware.

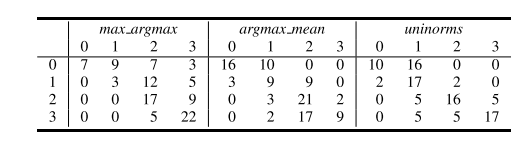
1. Video-Based Score Prediction

They evaluated video-based score prediction in terms of weighted F1 score, Precision and Recall. These are obtained by first computing the metric for each score (zero to three), and then computing the weighted average over scores, where the weight is the fraction of instances having that score. Note that weighted recall corresponds to (multiscore) accuracy, i.e., the fraction of correctly predicted scores over the total number of predictions.

****

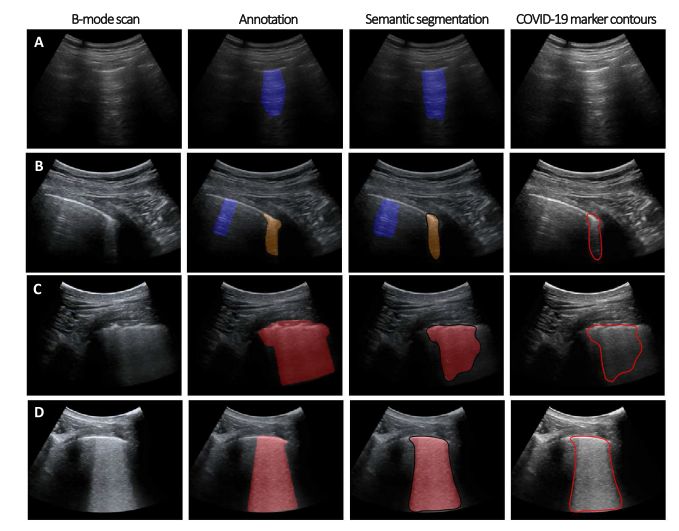
*Table 2: MEAN AND STANDARD DEVIATION OF WEIGHTED F1 SCORE, PRECISION AND RECALL COMPUTED OVER THE FIVE CROSS VALIDATION FOLDS, FOR THE PROPOSED VIDEO-BASED CLASSIFICATION METHOD AND BASELINES*

Table II reports averages and standard deviations of these metrics over the five folds of the cross validation procedure. They compared the video-level predictor with two standard aggregation methods, max\_argmax and argmax\_mean. The former implements the hard rule described in Section IV-B. It labels each frame with the most probable score according to the frame-level predictor, and takes the maximal score along the video. The latter averages frame-level predictions over the video and returns the score with the maximal average. The proposed method outperforms both baselines in terms of F1-score, precision and recall.



*Table 3: CONFUSION MATRICES (%) FOR THE PROPOSED VIDEO-BASED CLASSIFICATION METHOD AND BASELINES*

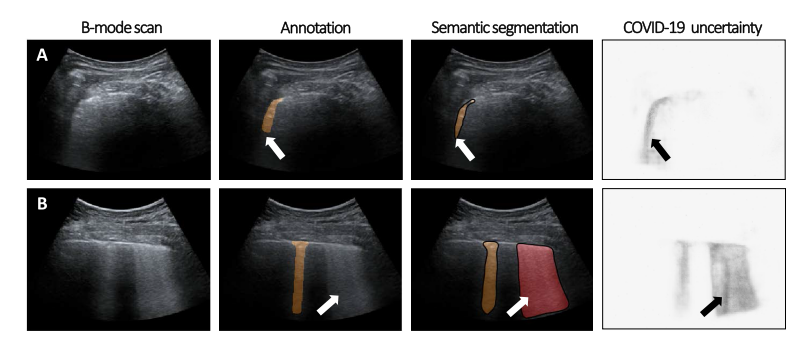
Table III shows confusion matrices for the three methods, obtained by concatenating the predictions for all folds. As expected, the max\_argmax hard rule is strongly biased towards predicting the highest score, resulting in bad performance on all other scores. On the other hand, the argmax\_mean baseline has the best performance in predicting score zero, but performs poorly on the other scores (under-predicting scores one and three and over-predicting score two). The uninorm-based aggregation is more balanced, outperforming each of the baselines on three out of four scores.

**

*Fig 4: Four examples of B-mode input image frames (first column), their annotations (second column) including COVID-19 biomarkers (moderate/score 2: orange, severe/score 3: red), and signs of healthy lung (blue). The corresponding semantic segmentations and contours of COVID-19 markers by deep learning are given in the third and fourth column, respectively*

1. Semantic Segmentation

Fig. 4 shows several illustrative examples of semantic segmentation results of the ensemble network, along with their ground-truth annotations. A quantitative assessment and comparison of segmentation performance for the U-Net, U-Net++, Deeplabv3+, and ensemble models are provided in Table IV. They observed that using on-line augmentation of images and annotations in combination with model ensembling yields a strong performance gain over a baseline U-Net, increasing the Dice coefficient from 0.64 to 0.75 for the union of COVID-19 markers. The ensemble model yields a categorical Dice score of 0.65 (mean across the segmentations for score 0, 2 and 3). This metric was 0.47 for baseline U-net.



*Fig 5:* *Two examples (A, B) of class uncertainty in the segmentations, showing B-mode input image frames (first column), annotations, including COVID-19 biomarkers, the corresponding semantic segmentations by deep learning , and pixel-level COVID-19 class uncertainty by MC-dropout .*

In Fig. 5 They provided a visualization of uncertainty in the predicted segmentations for two example images by plotting the pixel-wise standard deviation yielded by MC dropout across 40 samples. Arrows in (A) indicate a region displaying COVID-19 markers for which ambiguity in the exact shape and extent are well reflected in the pixel-level uncertainty. Arrows in (B) indicate a seemingly false-positive region which was assessed as a high-grade COVID-19 marker by the deep network, and not annotated as such. Interestingly, retrospectively the network output was judged as a true positive by the annotators, showing an area of hyperechogenic lung below the pleural surface, which characterizes a high permeability and advanced disease state.

## CONCLUSION

Labels in the ICLUS-DB turned out to be noisy. Furthermore, for frame-based classification and segmentation tasks the inter-operator agreement was not available. The noise can be indirectly observed in Table I, where using only a selection of training samples, performance improves by almost 5%. Extending the database to obtain frame-level labels from multiple annotators would surely lead to more robust models. Finally, the included LUS videos with score 0 are all of healthy patients, and therefore by no means to claim and to distinguish between COVID-19 patients and those with different pathologies.

## REFERENCES

[1] WHO. (2020). Laboratory Testing Strategy Recommendations for COVID-19: Interim Guidance. [Online]. Available:

https://apps.who.int/iris/bitstream/handle/10665/331509/WHOCOVID-19-lab\_testing-2020.1-eng.pdf

[2] R. Niehus, P. M. de Salazar, A. Taylor, and M. Lipsitch, “Quantifying

bias of COVID-19 prevalence and severity estimates in Wuhan, China

that depend on reported cases in international travelers,” medRxiv

2020.02.13.20022707, Feb. 2020.

[3] Y. Yang et al., “Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis and monitoring the viral shedding of

2019-nCoV infections,” medRxiv 2020.02.11.20021493, Feb. 2020.

[4] S. Salehi, A. Abedi, S. Balakrishnan, and A. Gholamrezanezhad,

“Coronavirus disease 2019 (COVID-19): A systematic review of

imaging findings in 919 patients,” Amer. J. Roentgenology, pp. 1–7,

Mar. 2020.

[5] A. Bernheim et al., “Chest CT findings in coronavirus

disease-19 (COVID-19): Relationship to duration of infection,”

Radiology, Feb. 2020, Art. no. 200463. [Online]. Available:

http://pubs.rsna.org/doi/10.1148/radiol.2020200463

[6] F. Mojoli, B. Bouhemad, S. Mongodi, and D. Lichtenstein, “Lung

ultrasound for critically ill patients,” Amer. J. Respiratory Crit. Care

Med., vol. 199, pp. 701–714, Mar. 2019.

[7] R. Raheja, M. Brahmavar, D. Joshi, and D. Raman, “Application of lung

ultrasound in critical care setting: A review,” Cureus, vol. 11, no. 7,

p. e5233, Jul. 2019.

[8] Y. Amatya, J. Rupp, F. M. Russell, J. Saunders, B. Bales, and

D. R. House, “Diagnostic use of lung ultrasound compared to chest

radiograph for suspected pneumonia in a resource-limited setting,” Int.

J. Emergency Med., vol. 11, no. 1, p. 8, Dec. 2018.

[9] E. Poggiali et al., “Can lung US help critical care clinicians in the early

diagnosis of novel coronavirus (COVID-19) pneumonia?” Radiology,

Mar. 2020, Art. no. 200847.

[10] Q.-Y. Peng, Chinese Critical Care Ultrasound Study Group, X.-T. Wang,

and L.-N. Zhang, “Findings of lung ultrasonography of novel corona

virus pneumonia during the 2019-2020 epidemic,” Intensive Care Med.,

vol. 46, no. 5, pp. 849–850, May 2020.

[11] G. Soldati et al., “Is there a role for lung ultrasound during the

COVID-19 pandemic?” J. Ultrasound Med., Apr. 2020.

[12] G. Soldati et al., “Proposal for international standardization of the use

of lung ultrasound for patients with COVID-19: A simple, quantitative,

reproducible method,” J. Ultrasound Med., Apr. 2020.

[13] K. Stefanidis et al., “Lung sonography and recruitment in patients with

early acute respiratory distress syndrome: A pilot study,” Crit. Care,

vol. 15, no. 4, p. R185, 2011.

[14] K. A. Stewart et al., “Trends in ultrasound use in low and middle

income countries: A systematic review,” Int. J. MCH AIDS, vol. 9, no. 1,

pp. 103–120, 2020.

[15] L. Tutino, G. Cianchi, F. Barbani, S. Batacchi, R. Cammelli, and

A. Peris, “Time needed to achieve completeness and accuracy in bedside

lung ultrasound reporting in intensive care unit,” Scandin. J. Trauma,

Resuscitation Emergency Med., vol. 18, no. 1, p. 44, 2010.

[16] R. J. van Sloun, R. Cohen, and Y. C. Eldar, “Deep learning in ultrasound

imaging,” Proc. IEEE, vol. 108, no. 1, pp. 11–29, Jul. 2019. [Online].

Available: http://arxiv.org/abs/1907.02994

[17] R. J. G. van Sloun and L. Demi, “Localizing B-Lines in lung ultrasonography by weakly supervised deep learning, in-vivo results,” IEEE

J. Biomed. Health Informat., vol. 24, no. 4, pp. 957–964, Apr. 2020.