Abdominal Trauma Detection Using CNNs With and Without Organ Segmentation

Allison Redfern AMR10211@NYU.EDU
Center for Data Science
New York University
New York, NY 10011, USA

Abstract

Traumatic injuries are a leading cause of death among young people, often resulting in abdominal injury. Rapid and accurate detection of injuries from medical imaging is crucial for patient care. This project employs Convolutional Neural Networks (CNNs) to detect liver injury from 3D abdominal CT scans, investigating the impact of organ segmentation on injury detection. The segmentation model, a 3D U-Net, performs pixel-level classification of liver or non-liver in each 3D scan. Predicted segmentations are used to create cropped 3D liver scans for each patient. Both cropped and uncropped images are utilized to train two classification model types: a 2D ResNet18 & LSTM model and a 3D ResNet18 model. The models trained on cropped images outperform those trained on uncropped images due to the reduction of noise and improved focus on the liver region.

Keywords: Medical Image Segmentation, Medical Image Classification, Convolutional Neural Networks, Injury Detection, CT Scans, 3D CNNs

1 Introduction

Traumatic injuries, a leading cause of death for people under 40, often result from blunt force trauma to the abdomen, such as in vehicle crashes [Colak et al. (2023)]. Prompt and accurate interpretation of imaging is essential for trauma patients' health. This project aims to automate traumatic injury detection from abdominal CT scans, facilitating in rapid life-saving care. Originating as a 2023 Kaggle competition, its goal was to detect internal bleeding and assess injury severity in abdominal organs, including the liver, kidneys, spleen, and bowel [Colak et al. (2023)]. This project deviates from the Kaggle competition by concentrating solely on detecting liver damage severity. A key focus is investigating the impact of organ segmentation on injury detection. This involves training an organ segmentation model to create cropped liver images, then comparing liver injury classification models trained on both cropped CT scans and the original uncropped CT scans.

Medical imaging segmentation and classification have seen extensive research, utilizing various techniques. 2D U-Nets have emerged as popular CNN architectures for organ segmentation in 2D images. However, with the increasing availability of 3D medical imaging data, there's a growing need for volumetric segmentation methods. Çiçek et al. introduced the 3D U-Net architecture, an extension of the 2D U-Net, which enables volumetric segmentation by replacing all 2D operations with their 3D counterparts [Çiçek et al. (2016)]. In terms of classification, two main methods from literature are implemented and compared. First, a 2D CNN structure followed by a long short-term memory (LSTM) layer has been utilized, as demonstrated by Islam et al., who successfully implemented a CNN-LSTM model to diagnose COVID-19 from X-ray images [Islam et al. (2020)]. This approach is similar to the methodologies used by the top performers in the original Kaggle competition [Dhankhar et al. (2023)], [Viel (2023)]. Additionally, 3D CNNs have gained traction in medical image classification [Singh et al. (2020)]. Yang et al. demonstrated the effectiveness of 3D CNNs in classifying Alzheimer's diagnoses and introduced various visual representations for model explainability [Yang et al. (2018)]. Overall, the combination of 3D U-Nets for segmentation and 2D CNN-LSTM models or 3D CNNs for classification presents promising avenues for medical image analysis, and this project investigates the impact of these segmentation models on the classification models' performance.

2 Hypothesis

The main hypothesis of this project is that models will perform better at detecting liver injury when trained on cropped images of the liver rather than entire CT scans. This hypothesis is based on the assumption that focusing the model's attention solely on the region of interest, i.e., the liver, will enable it to learn more discriminative features specific to liver injuries. Cropped images remove irrelevant information outside the liver area, reducing noise and potential distractions for the model during training. By training on cropped images, the model can allocate more capacity to learn features relevant to liver injury detection, potentially leading to improved performance compared to models trained on entire CT scans.

3 Data

The dataset for this project originates from the RSNA 2023 Abdominal Trauma Detection competition [Colak et al. (2023)]. It consists of 3,147 patients, each with respective metadata corresponding to injuries of various abdominal organs. These patients underwent a total of 4,711 abdominal CT scans. Each scan is a 3D image, consisting of 1.5M 2D slices in total. Additionally, there are 206 3D organ segmentation files with pixel-level organ labels for their respective CT scans. There is substantial class imbalance in this dataset. The target feature positive class is that the CT scan shows an injured liver while the negative class is that the scan contains a healthy liver. The positive class makes up for only 10 % of the data. This class imbalance is a common occurrence in medical image data and will be addressed through various methods while modeling.

3.1 Segmentation Data Preprocessing

The 206 organ segmentation files from Kaggle consist of multiple 2D matrices, not full 3D segmentation matrices. To prepare them for analysis, preprocessing is done to compile and concatenate the slices. Additionally, the segmentations' orientation is adjusted to match that of the CT scans by flipping along the Z-axis, transposing, and then flipping along the X and Y axes. To maintain consistency with the preprocessed CT images, every other slice of the segmentation is retained, as discussed in the following section. Since this project focuses solely on liver injury detection, the segmentation labels are converted from multiclass to binary, where the liver is the positive class and all other parts of the abdomen are the negative class. After preprocessing, the segmentation files are split into training, validation, and test sets. The data is divided such that 80% is used for training, 10% for validation, and 10% for testing the organ segmentation model. This split is performed at the patient level to prevent any patient from appearing in more than one set, thus avoiding model leakage.

3.2 Classification Data Preprocessing

The injury metadata used in the classification model originally consists of multiclass labels indicating healthy, low-grade injury, and high-grade injury for multiple organs. For simplicity, the liver is chosen as the target organ in the classification model, and its injury label is converted to a binary value: healthy or injured. Thus, the metadata includes target labels, patient IDs, scan IDs, and file paths to the respective scan images. Next, the metadata is split into training, validation, and testing sets using an 80/10/10 split. This split again avoids leakage by ensuring that each patient appears in only one of the three datasets. Similar to the segmentation preprocessing, the CT scans are also compiled into 3D images by concatenating the 2D slices. Following a methodology similar to the first-place Kaggle winner [Dhankhar et al. (2023)], only every other slice in each CT scan is retained to reduce overall image size and speed up model training time.

4 Materials & Methods

The following subsections discuss the segmentation and classification model architectures and training procedures.

4.1 Segmentation Model Architecture and Training

To segment the 3D abdominal CT scans by liver and non-liver, a 3D U-Net is built and implemented. The architecture of the model follows an encoder-decoder structure, as shown in Appendix Figure A. In the encoder part, each level consists of multiple convolutional layers and batch normalization followed by a max-pooling convolution. This allows the network to capture features at different scales and levels of abstraction. In the decoder part, each level consists of upsampling layers, followed by convolutional layers and additional batch normalization. Skip connections between corresponding encoder and decoder layers help in retaining fine-grained details during upsampling. These skip connections concatenate feature maps from the encoder with those from the decoder, allowing the decoder to refine the segmentation based on both local and global context.

The organ segmentation 3D U-Net is trained on the training segmentations and their respective CT scans. Weighted cross-entropy loss is applied to address the class imbalance between liver pixels and others by heavily penalizing misclassifications of liver pixels. This approach ensures effective learning to distinguish liver from non-liver tissues, even with significant class imbalance, as liver pixels are much fewer than the overall abdomen. Each image is resized to 128x128x128 pixels and normalized to scale the pixel values between 0 and 1 which helps prevent gradients from becoming too large or too small, allowing for more stable and efficient optimization. The learning rate and batch size are tuned, selecting the combination leading to the lowest loss in the validation set. The model performed best with a batch size of 2 and a learning rate of 0.0001. The best model is trained for 60 epochs due to its high generalizability on the validation set. The epoch with the lowest validation loss is chosen as the final model. The final model is evaluated using the dice score, which is calculated as twice the intersection of the predicted liver segmentations and true liver segmentations divided by the sum of the predicted and true liver areas. A higher dice score closer to 1 indicates better overlap between the predicted and true positive regions, thus reflecting the model's accuracy in delineating the liver area.

After the segmentation model is tuned, trained, and evaluated, the model is then used to mask all 4,711 abdominal CT scans to only show the predicted liver, which is referred to in this paper as the *cropped* images. The original, or *uncropped*, CT scans and the cropped CT scans are then used as input for the classification models.

4.2 Classification Model Architectures and Training

The first classification model, shown in Appendix Figure Ba, combines a 2D ResNet18 and a bidirectional LSTM layer. ResNet18, pretrained on ImageNet, serves as a feature extractor for the abdominal slices. ResNet18 consists of a series of residual units, each comprising several convolutional layers followed by a shortcut connection and batch normalization. The ResNet18 block is modified to accept single-channel input, and its final fully connected layer is adjusted to output features for the LSTM layer. After ResNet18, the bidirectional LSTM captures dependencies between CT scan slices. The LSTM output undergoes global max pooling to retain relevant information for liver injury. Batch normalization is applied to stabilize training, and dropout is used prevents overfitting. The output passes through two fully connected layers, with additional dropout, for final injury prediction.

The second classification model, depicted in Appendix Figure Bb, is a 3D ResNet18 model. Originally developed for video data, the 3D ResNet18 replaces the 2D operations in its 2D counterpart with 3D operations. This model utilizes pretrained weights and is modified to accept single-channel input. Batch normalization is applied between each layer to stabilize training, and dropout is used before the final fully-connected layer to reduce overfitting.

Both models utilize a weighted cross-entropy loss function to heavily penalize misclassifications of the minority class (liver injury). A weight decay factor of 0.0001 is included to regularize the model and prevent overfitting. Images are resized to 64x128x128 pixels and normalized. Additionally, a random rotation between -5 and 5 degrees is applied to training CT scans to enhance model generalization. Batch size and learning rate are optimized for both models, selecting combinations yielding the highest F1-score in the validation set. For the 2D ResNet & LSTM model, the best batch size is 8, and the best learning rate is 0.00001. For the 3D ResNet model, the best batch size is 8, and the best learning rate is 0.0001. Each model is trained twice, once with cropped liver images and once with uncropped abdominal images, for 10 epochs, and each model with the lowest validation loss is selected as the final model. After training, each model is evaluated on the test set and compared using area under the receiver operating characteristic curve (AUC), area under the precision recall curve (AUPRC), and F1-score. AUC measures the ability of the model to distinguish between positive and negative classes across all thresholds, while AUPRC and F1-score focuses on the precision-recall trade-off, being particularly useful for this imbalanced dataset.

5 Results

The results of the segmentation and classification models are discussed in the following subsections.

5.1 Segmentation Results

The final segmentation model achieved a dice score of 0.79, meaning that 79% of the predicted segmentation overlaps with the ground truth, indicating a strong performance in accurately delineating object boundaries. An example of the actual and predicted segmentation for one cross-section of the abdomen is shown in Figure 1.

(a) Abdomen Cross-Section (b) Actual Segmentation (c) Predicted Segmentation

Figure 1: Liver segmentation example

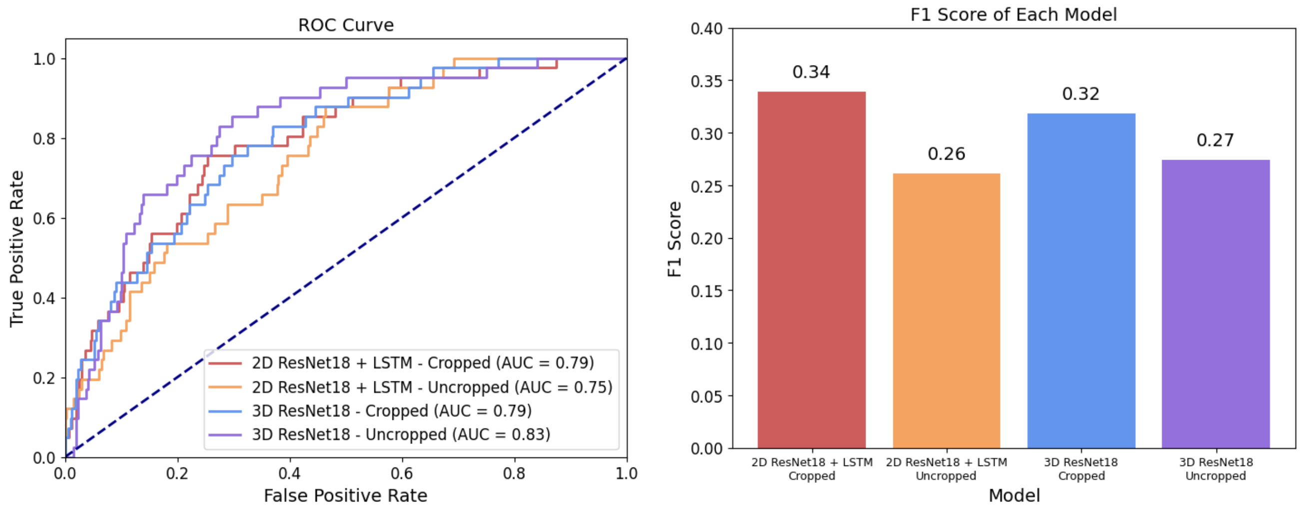
This example is representative of the trends in outputs of the segmentation model. Although the segmentation does consistently include the complete liver, it also includes additionally area surrounding it, which is considered further in the Discussions section.

5.2 Classification Results

The AUC and F1-score results are shown in Figure 2 and the AUPRC results are shown in Appendix Figure C. Although the 3D ResNet18 model using uncropped scans has the highest AUC of the four models, both models using cropped images have higher AUPRC and F1-scores, indicating that they are better at capturing the minority class, or liver injury, instances. The F1-scores could be significantly improved, and potential improvements to increase overall model capacity and generalization ability are discussed further in the Discussions section.

6 Discussion

The hypothesis that models trained on cropped liver images would outperform those on uncropped images is confirmed, as the AUPRC and F1-scores are higher for cropped image models. To further explain the reasoning behind this performance, Appendix Figure D displays false positive rates for each model, showing consistently lower rates for cropped image models. However, when considering images with other organ injuries, although the



(a) ROC curve and AUC of each model

(b) F1-score of each model

Figure 2: Classification performance metrics

uncropped models' false positive rates remain highest, false positive rates increase for both model types. This suggests that both models may be picking up context from other organs rather than specifically classifying liver injury. The predicted liver segmentations consistently included surrounding areas, possibly explaining why false positives increase for cropped image models. Nonetheless, the consistently lower false positive rates of cropped image models likely result from reduced noise when removing context from other organs.

Given the lower F1-scores across all four models, future improvements could be made to enhance liver injury detection. While using a weighted cross-entropy loss function encourages positive class prediction, the models tend to over-predict this class. In the medical context, prioritizing injury detection is crucial, but excessively high false positive rates can strain limited resources and lead to unnecessary interventions. Thus, fine-tuning the weights in the loss function or exploring alternative loss functions for better class balance could improve model performance. Additionally, different model architectures could improve capacity and address overfitting observed in each classification model. While regularization methods like batch normalization, weight decay, and dropout were applied, additional regularization techniques could be beneficial. Once overfitting is mitigated, training the model for more epochs could enhance performance. Moreover, incorporating new model types, like transformers, which leverage attention mechanisms to capture long-range dependencies, may further improve performance. Overall this investigation revealed that incorporating segmentation does enhance the models' ability to accurately detect liver injury. Although improvements can be made to increase overall model capacity, furthering research with a focus on segmentation followed by a classification model provides promising results to improve injury detection and ultimately facilitates prompt and effective clinical interventions for trauma patients.

References

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016. URL `http://arxiv.org/abs/1606.06650`.

Errol Colak, Hui-Ming Lin, Robyn Ball, Melissa Davis, Adam Flanders, Sabeena Jalal, Kirti Magudia, Brett Marinelli, Savvas Nicolaou, Luciano Prevedello, Jeff Rudie, George Shih, Maryam Vazirabad, and John Mongan. Rsna 2023 abdominal trauma detection, 2023. URL `https://kaggle.com/competitions/rsna-2023-abdominal-trauma-detection`.

Nischay Dhankhar, Harshit Sheoran, and Qishen Ha. Rsna 2023 abdominal trauma detection: 1st place solution: Team oxygen, 2023. URL `https://www.kaggle.com/competitions/rsna-2023-abdominal-trauma-detection/discussion/447449`.

Md. Zabirul Islam, Md. Milon Islam, and Amanullah Asraf. A combined deep cnn-lstm network for the detection of novel coronavirus (covid-19) using x-ray images. *Informatics in Medicine Unlocked*, 20:100412, 2020. ISSN 2352-9148. doi: 10.1016/j.imu.2020.100412. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7428728/`.

Satya P. Singh, Lipo Wang, Sukrit Gupta, Haveesh Goli, Parasuraman Padmanabhan, and Balázs Gulyás. 3d deep learning on medical images: A review, 2020.

Theo Viel. Rsna 2023 abdominal trauma detection: 2nd place solution, 2023. URL `https://www.kaggle.com/competitions/rsna-2023-abdominal-trauma-detection/discussion/447453`.

Chengliang Yang, Anand Rangarajan, and Sanjay Ranka. Visual explanations from deep 3d convolutional neural networks for alzheimer's disease classification. *CoRR*, abs/1803.02544, 2018. URL `http://arxiv.org/abs/1803.02544`.

Appendix

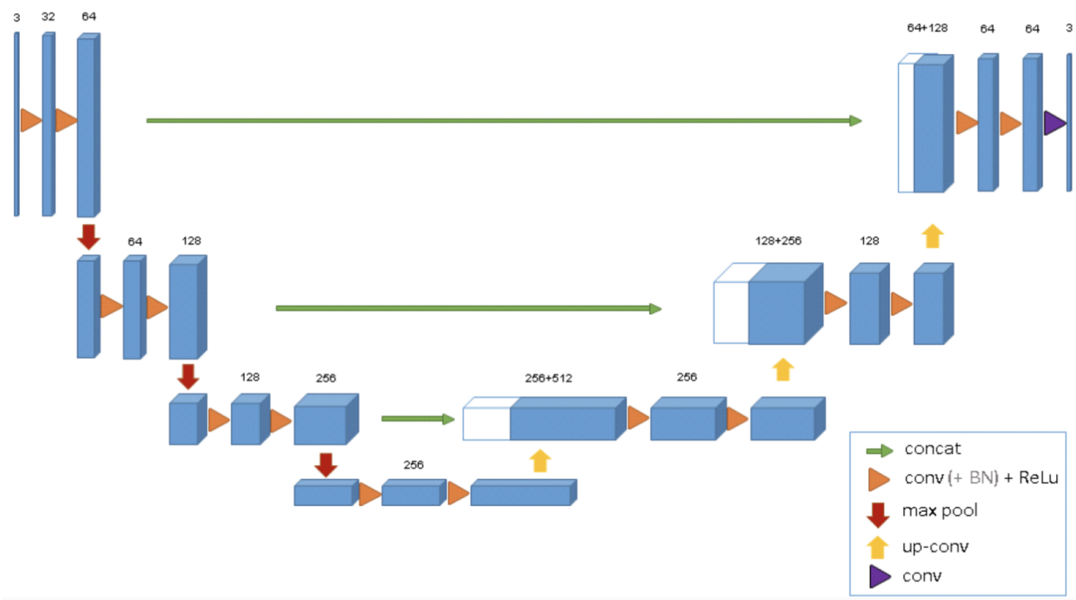
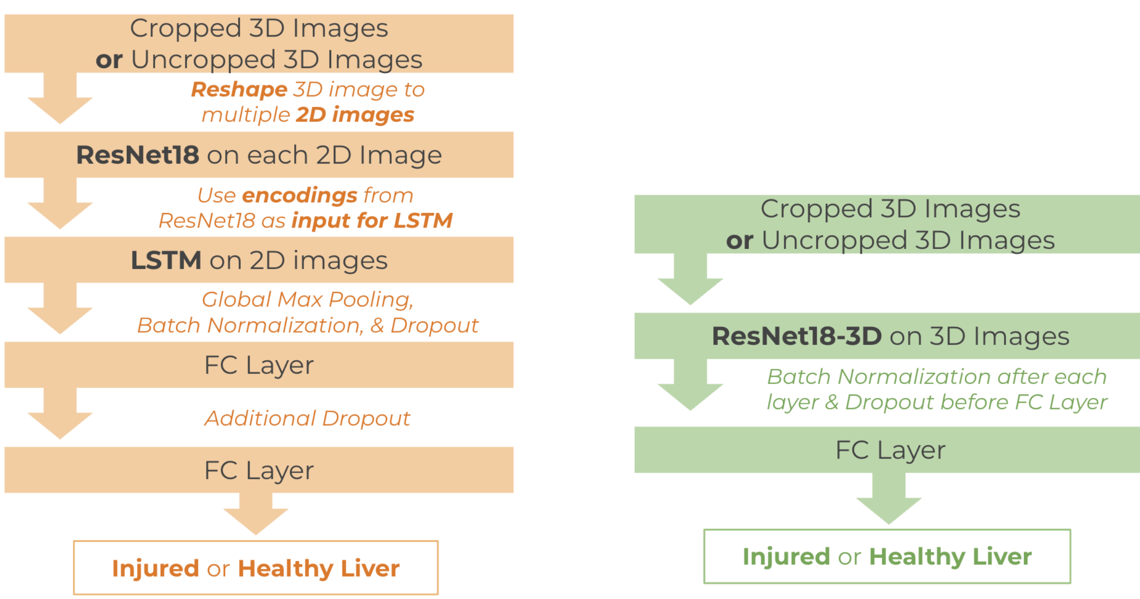


Figure A: Segmentation 3D U-Net model architecture



(a) 2D ResNet and LSTM

(b) 3D ResNet

Figure B: Liver injury classification model architectures

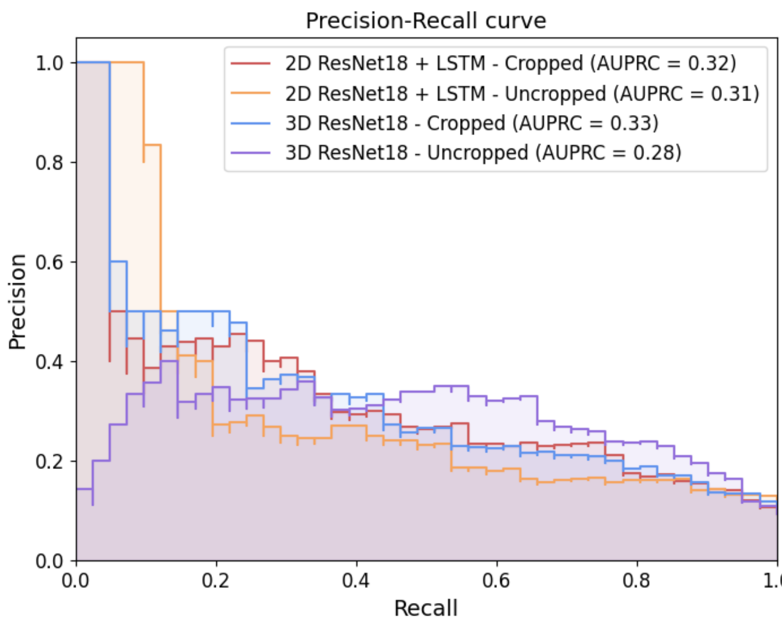
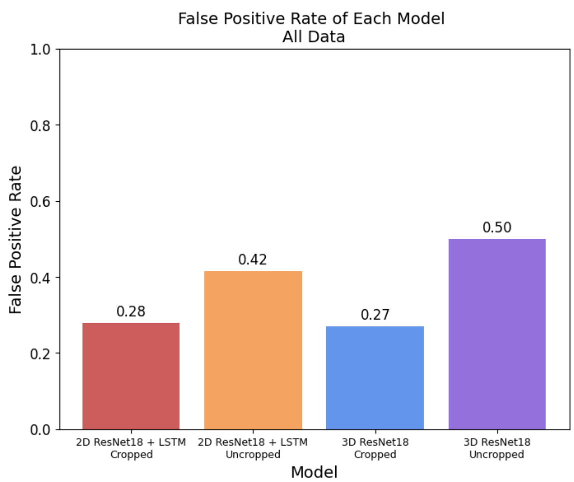
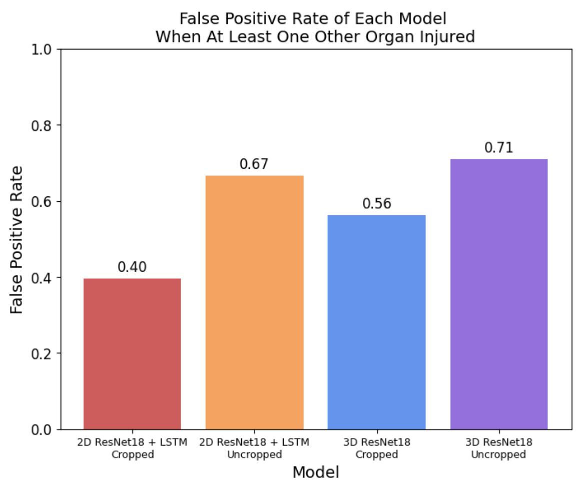


Figure C: Area under Precision-Recall Curve (AUPRC)



(a) FPR using all data



(b) FPR when other organ injured

Figure D: False positive rates (FPR)