January 2019

Medically Applied Artificial Intelligence:from Bench To Bedside

Nicholas Chedid

Medically Applied Artificial Intelligence: From Bench to Bedside

A Thesis Submitted to the Yale School of Medicine in Partial Fulfillment of the Requirements

for the Degree of Doctor of Medicine

by

Nicholas Chedid

2019

"Before I came here, I was confused about this subject. Having listened to your lecture, I am still confused – but on a higher level."

Enrico Fermi

"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I got a good job in journalism."

Dave Barry

YALE SCHOOL OF MEDICINE

Abstract

Dr. Richard Andrew Taylor

Doctor of Medicine

Medically Applied Artificial Intelligence: From Bench to Bedside

by Nicholas CHEDID

The intent of this thesis was to develop several medically applied artificial intelligence programs, which can be considered either clinical decision support tools or programs which make the development of such tools more feasible. The first two projects are more basic or "bench" in focus, while the final project is more translational. The first program involves the creation of a residual neural network to automatically detect the presence of pericardial effusions in point-of-care echocardiography and currently has an accuracy of 71%. The second program involves the development of a sub-type of generative adverserial network to create synthetic x-rays of fractures for several purposes including data augmentation for the training of a neural network to automatically detect fractures. We have already generated high quality synthetic x-rays. We are currently using structural similarity index measurements and Visual Turing tests with three radiologists in order to further evaluate image quality. The final project involves the development of neural networks for audio and visual analysis of 30 seconds of video to diagnose and monitor treatment of depression. Our current root mean square error (RMSE) is 9.53 for video analysis and 11.6 for audio analysis, which are currently second best in the literature and still improving. Clinical pilot studies for this final project are underway. The gathered clinical data will be first-in-class and orders of magnitude greater than other related datasets and should allow our accuracy to be best in the literature. We are currently applying for a translational NIH grant based on this work.

Acknowledgements

I would like to thank my advisor Dr. Andrew Taylor, and my colleagues and friends Michael Day, Alexander Fabbri, Maxwell Farina, Anusha Raja, Praneeth Sadda, Tejas Sathe, and Matthew Swallow without whom this thesis would not have been possible.

This work was supported by the National Institutes of Health under grant number T35HL007649 (National Heart, Lung, and Blood Institute) and by the Yale School of Medicine Medical Student Research Fellowship.

I would also like to thank the Sannella Family for their generous support of my medical education through the Dr. Salvatore Sannella and Dr. Lee Sannella Endowment Fellowship Fund.

Contents

List of Figures

List of Tables

Dedicated to my parents whose sacrifices and courage in coming to this country have given me a life of opportunity…

Chapter 1

Deep Learning for the Detection of Pericardial Effusions in the Emergent Setting

1.1 Introduction

1.1.1 Ultrasound for Pericardial Effusion

The first ultrasound was introduced in the 1950s but would not become widely utilized in clinical practice until the 1970s [1]. Real-time ultrasound was developed in the 1980s, which allowed for adoption in emergent settings [1]. Since then, Point-of-Care Ultrasound (POCUS) has become an increasingly important diagnostic tool utilized in the emergency department, and there has been significant research towards improving ultrasound techniques for the evaluation of a wide variety of clinical conditions [1, 2, 3].

One such condition for which ultrasound has been utilized is pericardial effusion. Ultrasound is the preferred diagnostic tool for pericardial effusion given it is fast, accurate, widely available, and non-invasive [4].

However, while some physicians have specific extended training using ultrasonography, there is concern regarding diagnostic variability between those who have

this training compared to those who do not. This raises the concern for potential for error in the diagnosis of pericardial effusion in the emergency room. In a study published in Academic Emergency Medicine, residents and faculty from an emergency medicine training program at a Level 1 trauma center were asked to view ultrasound clips in patients with chest trauma [5]. The overall sensitivity was 73% and specificity was 44%. Given the possible emergent nature of pericardial effusion and the importance of rapid and accurate diagnosis, a diagnostic or clinical decision support tool to reduce error would likely be of significant benefit. A form of artificial intelligence (AI) increasingly used for imaging purposes called a convolutional neural net (CNN) may be able to serve as such a tool.

1.1.2 Use of Neural Networks in Medical Imaging

Medical imaging can be broken down into two basic components: image acquisition and image interpretation. Image acquisition has improved greatly over the past decades with significantly increased acquisition speed and accuracy; however, improvements to image interpretation have been much slower to manifest. This is particularly due to the fact that the image interpretation process has primarily been a human not a technologically driven process with most interpretations performed by physicians. This comes with many of the limitations associated with a human-driven process such as subjectivity, human error, fatigue, limited interpretation speed, and significant variability among providers. Technological aids to the image interpretation process have only recently begun to be developed.

One such aid is machine learning (ML). ML is an application of AI that allows systems to automatically learn and improve from experience without being explicitly programmed. Machine learning has been increasingly used for medical imaging tasks particularly in the fields of radiology and pathology [6]. A specific machine learning technique called a deep convolutional neural network (CNN) has become the new gold-standard machine learning technique in medical imaging research [7].

Neural networks are inspired by the structure and function of a biological nervous system. A neural network is composed of neuronal layers just as a nervous system is composed of layers of neurons. Each neuron in a neural network is connected to neurons in the prior and subsequent neuronal layer but not to neurons within the same layer. Each of these connections is associated with a certain weight value. Each neuron can be thought of as a logistic regression function. Each time the model runs forward it ends with a final error value. The model then runs backward in order to attach new weights to each of the parameters based on the error. This process is repeated until the error stabilizes at a minimum value. Once a neural network has been optimally trained on a set of images to have maximal accuracy in identifying them correctly, it is then tested on a completely novel set of images to see if its predictive capabilities can generalize to fresh images.

Neural networks have been used for a wide variety of medical applications including classification of skin cancer from pathology images [8], detection of pneumonia on chest X-rays [9], and detection of polyps during colonoscopy [10].

The use of neural networks in ultrasounds is much less developed due to several difficulties associated with ultrasound. Ultrasound can be more complex than other imaging modalities, which often contain a single still frame, because it consists of video containing many frames, with very little labeled information. Ultrasound also has decreased resolution compared to other imaging modalities such as CT and MRI. Additionally, for echocardiograms in particular, measurements and the visible anatomy can vary significantly with the beating of the heart. Preliminary work using neural networks for echocardiography has been performed showing an ability to detect hypertrophic cardiomyopathy and cardiac amyloidosis with C-statistics of 0.93 and 0.84 [11]. However there has been very little work conducted on ultrasound acquired in the point-of-care setting.

1.1.3 Need for Data: a Call for Multicenter Collaboration

Perhaps the most important variable in creating a high performing neural network is the sheer quantity of labeled data needed, for example, ultrasounds labeled as effusion present or absent. A larger dataset provides more material for the neural net to learn from enabling greater final accuracy.

Given the vast amount of data necessary to train high performing machine learning algorithms, the quantity of data needed often quickly outstrips that available at a single institution; this has led to some in the field calling for increased multicenter collaborations [12, 13].

In this paper, we aim to demonstrate a proof-of-concept neural network for a clinical decision support tool for pericardial effusion in the emergent setting while highlighting the need for increased multicenter collaboration for the development of high performing neural networks.

1.2 Methods

1.2.1 Image Acquisition and Classification

Image acquisition and classification was done primarily by Nicholas Chedid.

Echocardiograms in the DICOM format were manually gathered using the Emergency Department's picture archiving and communication system (QPath). Ultrasounds were chosen sequentially from all adult patients ($\geq$ 18 years) who had an ED echocardiogram performed within the period March 2013 to May 2017. These ultrasounds were interpreted and labeled by the resident or attending physician who acquired them. Only echocardiograms taken in the parasternal long axis view were included (for optimal visualization of a wider range of cardiac pathology). Additionally only echocardiograms with at least two documented readings by physicians (including at

least one by an attending physician) were included. All echocardiograms and interpretations were also reviewed by me for inclusion. These DICOMs selected for inclusion were saved in a Yale Secure Box folder. Additionally, an Excel spreadsheet was created to organize information relevant to each DICOM. Each DICOM was recorded numerically and several associated characteristics were manually transcribed including: medical record number (MRN), account number, accession number, date of the study, effusion status (present or absent), equality status (presence or absence of strain), exit status (dilated or normal), ejection fraction status (depressed, <50%, normal, 50 - 65%, hyperdynamic >65%), and number of studies associated with each encounter. This resulted in a dataset consisting of 1545 videos from 1515 patients. For this study, only those videos that specifically commented on the presence or absence of pericardial effusion were included. This resulted in 272 videos.

These videos were then fed through an image preprocessing Docker package created by collaborator Adrian Haimovich. Preprocessing included anonymization by stripping of all identifying metadata and splitting into still frames. Our utlimate dataset consisted of 12,942 still frames. Our training dataset consisted of 80% of these frames (10,299) and our test dataset consisted of the remaining 20% (2643).

1.2.2 ResNet 20

Work for building and tuning the ResNet architecture was done primarily by Nicholas Chedid

A neural network was developed using Python scripts and programs using Keras packages running on a Theano backend. Specifically the subtype of convolutional neural network created was a 20-layer residual network, a gold-standard neural network for image classification and computer vision tasks, which won the Imagenet challenge in 2015 (ResNet-20) [14]. Plain deep networks can be difficult to train because of vanishing and exploding gradients. By using stacked Residual Blocks, which

use skip connections that take activations of one layer and feed them to much deeper layers, ResNets are able to be built much deeper.

Our network has 20 weighted layers with shortcut connections inserted. Our convolutional layers mostly have 3x3 filters and are designed so that: (1) for the same output feature map size, the layers have the same number of filters; and (2) if the feature map size is halved, the number of filters is doubled so as to preserve the time complexity per layer. Downsampling was performed directly by convolutional layers with a stride of 2. The network ends with a global average pooling layer and a 1000-way fully-connected layer with softmax. L2 regularization was utilized to reduce overfitting. Batch normalization was used to speed up training and increase accuracy. Activation functions primarily consisted of recitified linear units (ReLU) except for the softmax classifier layer.

Many different training iterations were run for hyperparameter tuning in order to optimize the neural network's accuracy on the test set. Tunable variables included: number of epochs, ResNet Model (i.e. number of layers), learning rate, L2 Regularization coefficient, batch size, and data augmentation features including: featurewise center, samplewise center, featurewise standard normalization, samplewise standard normalization, zca whitening, rotation range, width shift range, height shift range, horizontal flip, and vertical flip.

The optimal neural net was one in which epochs were set to 50, ResNet Model (i.e. number of layers) was set to 20, learning rate was set to 0.001, L2 Regularization coefficient was set to 1.00E-03, batch size was 16, and the following data augmentation features were set to: featurewise center = off, samplewise center = off, featurewise standard normalization = off, samplewise standard normalization = off, zca whitening = off, rotation range = 180, width shift range = 0.15, height shift range = 0.15, horizontal flip = on, and vertical flip = on.

Training was performed over nearly 19 hours on a desktop computer with 3 Titan X NVIDIA graphics cards with 8 GB RAM each.

The number of layers and the batch size were not able to be further increased due to system constraints . Fortunately deepening of the ResNet past 20 layers did not seem to significantly improve tests accuracy across 200 epochs (accuracy remained at 92%) from ResNet20 to ResNet110 as seen in He *et al*. [14]. Code availability: ResNet is publicly available on Github.

1.3 Results

After running the aforementioned ResNet for 200 epochs, we were able to achieve a final test accuracy of 71%. Our results can be seen in Table 1.1.

The table is organized into three columns with percentage of our total dataset used for training in the left column, our final test accuracy in the middle column, and our final train accuracy in the right hand column. The final train accuracy remained fairly consistent ranging from 74 - 81%. More importantly, it can be seen that the test accuracy improved from 49 to 71% as the amount of data used to train our ResNet increased from 20 to 80% illustrating the continual improvement associated with increasing available training data.

TABLE 1.1: Neural Network Performance in Identifying Presence or Absence of Pericardial Effusion

% of dataset used	Final Test Accuracy	Final Train Accuracy
25% of 80 = 20%	49%	75%
50% of 80 = 40%	42%	81%
75% of 80 = 60%	57%	78%
Full(80%)	71%	74%

1.4 Discussion

We have demonstrated the creation of a proof-of-concept neural network for a clinical decision support tool for pericardial effusion in the emergent setting with an accuracy

of 71% compared to a sensitivity of 73% and specificity of 44% for the detection of pericardial effusions by academic emergency medicine physicians [5]. We are currently in the process of writing code that would allow us to assess the sensitivity and specificity of our program as well.

The accuracy of our neural network showed step-wise improvement as we used increasing percentages of our available data. Given the fact that our training data came from one of the highest volume EDs in the United States (Yale New Haven Hospital has a very high volume ED with the 3rd most ER visits in 2016 [15]) and our results suggest likely continual improvement with even more data, this highlights the need for multicenter collaboration to aggregate sufficient training data to train very high performance algorithms that can aid in clinical decision making.

Future steps include writing code that would allow us to assess the sensitivity and specificity of our program as well as several steps that may help improve our accuracy further such as incorporating transfer learning from a ConvNet pre-trained on ImagNet, reformatting input data from still frames to short video clips as this may improve performance, using a Generative Adversarial Network (GAN) instead of a ResNet, and using segmentations to improve performance.

Chapter 2

Fracture X-Ray Synthesis with Generative Adversarial Networks

2.1 Introduction

2.1.1 Fractures in the Emergency Department

Fractures are among the most common reasons for emergency department visits. While some fractures are easily discernible on x-ray, many others are subtle enough to require a radiologist's inspection for a definitive diagnosis. In the fast-paced environment of the emergency department, the subtleties in fracture diagnosis can sometimes be overlooked or misinterpreted, leading to medical error. This phenomenon has been quantified before: A four-year study in a busy district general emergency department found 953 diagnostic errors, of which 760 (79.7%) were missed fractures [16]. The primary reason for diagnostic error in 624 of 760 (82.1%) of these patients with fractures was a failure to interpret radiographs correctly [16].

The annual incidence of fractures has been estimated to be as high as 100.2 per 10,000 in males and 81.0 per 10,000 in females [17].

Additionally delay in appropriate diagnosis may lead to worsened clinical outcomes and increased healthcare costs. Medical errors cost the United States $17 billion in 2008 [18].

A technology that can automatically detect fractures has the potential to reduce emergency department medical errors, costs, and waiting times. However, training image analysis algorithms often requires hundreds or thousands of manually annotated examples. The process of annotating these examples can be labor and time intensive.

The process of developing automatic fracture detectors is even more burdensome given that there are many different types of fracture, which would require training many different types of detectors. Hundreds to thousands of images would have to be manually annotated to train each of these detectors. Fortunately, here we describe a method to greatly simply the training of a multitude of automatic fracture detectors. This method entails the creation of synthetic x-rays from procedurally generated segmentations, thereby creating annotated datasets with minimal human time expenditure.

Data augmentation is the process of increasing the total information provided by a training dataset by generating many variants of datapoints within the dataset. In the context of images, this often involves simple transformations such as rotation, scaling, and translation. Training an algorithm on many examples of the same images that are rotated by different amounts can teach that algorithm rotational invariance; training it on many resized examples of an image can teach invariance to scale and so on.

However simple image transformations are unable to teach invariance to more subtle features. Generating synthetic images to augment training data sets may improve invariance to these more subtle features. In this work, we demonstrate that it is possible to generate synthetic x-ray images using image-to-image synthesis for the purpose of data augmentation.

2.1.2 Image-to-Image Synthesis

Image-to-Image synthesis is the process of converting an image from an element of one domain to an equivalent image from an element of another domain.

Training image-to-image synthesis algorithms is notoriously difficult because image-to-image synthesis is an underconstrained problem. There are many correct solutions to any image-to-image synthesis problem.

A generative adversarial network (GAN) is a generative model that is trained in an adversarial process between two sub-networks: a generative model G and a discriminative model D. G learns to generate synthetic simulations of images from a particular domain while D learns to discriminate between true images from that domain and synthetic imitations generated G. This is an adversarial process in the sense that these two networks are trained in opposition. Generally optimization of one's network's performance will lead to deterioration of the other's. Thus an ideal, unique solution exists where G recovers the training data distribution and D is equal to $\frac{1}{2}$ everywhere.

2.1.3 Prior Work

Chuquicusma *et al*. [19] used generative adversarial networks (GANs) to create synthetic lung cancer nodules and place them in computed tomography (CT) images. The overall quality of these synthesized nodules was then evaluated using a "Visual Turing test," which consisted of having two radiologists evaluate images with either real or synthetic nodules and try to distinguish between the two. The creation of synthetic lung nodules via GANs was a novel concept. Possible next steps might include: using quantitative measures of image synthesis such as a structural similarity index in addition to the qualitative Visual Turing test, using the pix2pixHD method which may be an interesting way of generating higher resolution images, and generating entirely synthetic images as opposed to a component within the image (*e.g.* lung nodules).

Korkinof *et al*. [20] used GANs to generate synthetic mammograms. The overall quality of these synthetic images was evaluated qualitatively by comparing them visually to real mammograms. The creation of synthetic high-resolution mammograms via GANs was a novel concept. Possible next steps might include: using more rigorous qualitative assessments of image synthesis such as the Visual Turing Test by experts used by Chuquicusma *et al*., using quantitative measures of image synthesis such as a structural similarity index, and by using the pix2pixHD method which may be an interesting way of generating even higher resolution images.

2.2 Methods

2.2.1 Network Architecture

This work uses the pix2pixHD network architecture described by Wang *et al*. [21]. The pix2pixHD method improves upon GANs by introducing a coarse-to-fine generator and multi-scale discriminator architecture which allows for image generation at a much higher resolution with an order of magnitude less memory.

The coarse-to-fine generator consists of a global generator network G_1 and a local enhancer network G_2. The architecture of the global generator G_1 is that proposed by Johnson *et al*. [22]: a convolutional front-end, a set of residual blocks, and a transposed convolutional back-end. The architecture of the local enhancer network G_2 is the same except that the input to the residual blocks consists of the element-wise sum of not only the feature maps from the convolutional front-end of G_2 but also the last feature map of the transposed convolutional back-end of G_1, which helps integrate information from the global network.

The coarse-to-fine moniker describes the training method of the generator. First, G_1 is trained on lower resolution versions of the original training images, then G_2 is appended to G_1, and finally the two networks are trained together on the full resolution,

original images.

Utilizing the coarse-to-fine generator to produce higher resolution synthetic images poses a novel challenge however. Traditional GAN discriminator design does not perform as well on these higher resolution images because to distinguish between higher resolution real and synthetic images it would be necessary to use a discriminator with a large receptive field. This could be accomplished by either using a deeper network or larger convolutional kernels both of which could potentially cause overfitting and would require significantly more memory for training. This was addressed by Wang *et al*. in the design of their multi-scale discriminator consisting of three discriminators—D_1, D_2, and D_3—with identical network structures but organized in a pyramid structure in which each discriminator operates at different image scales funneling from lower to higher image resolutions as seen in Figure 2.1.

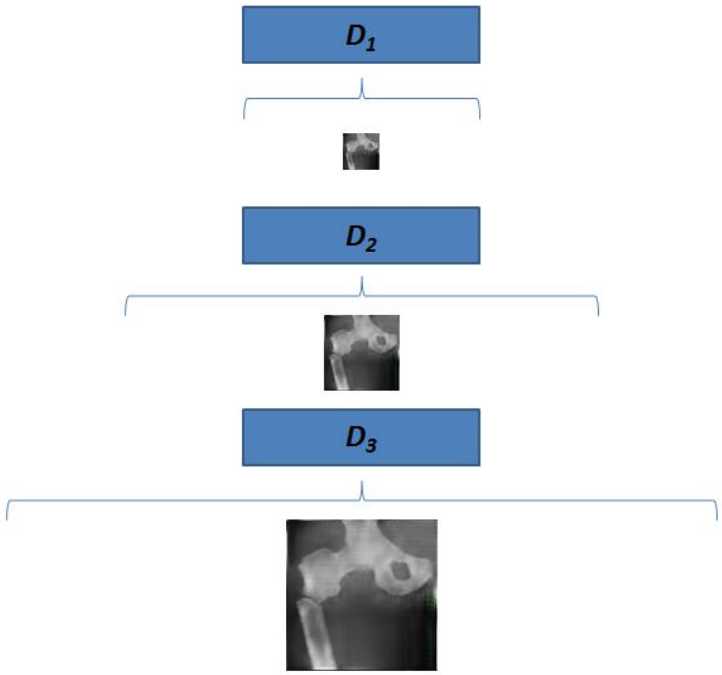

FIGURE 2.1: Multi-scale Discriminator: D_1, D_2, and D_3 are the three discriminators that make up the multi-scale discriminator. Each has the same architecture. They are multi-scale in that they form a pyramid structure with each operating at a smaller scale with correspondingly smaller receptive fields from D_3 to D_1.

2.2.2 Image Acquisition and Preprocessing

Image Acquisition and preprocessing work was done primarily by Nicholas Chedid.

50 x-rays of femoral fractures were downloaded from an internet search. Using a small initial dataset of 50 images aligns with the goals of this work to show how a pix2pixHD pipeline could allow for a rapidly scalable tool to aid in data augmentation of many fracture types while reducing manual work and the need for very large databases. The difficulty in acquiring and the manual work necessary to use a dataset of this size is much less than what would be needed to acquire and label a traditional dataset of several hundred to thousands of images for training a single fracture detection algorithm. Not only might time and manual labor be significantly reduced via the creation of a pix2pixHD pipeline, but the training of accurate neural networks that may have previously been hampered by a lack of original data may be possible.

The 22 highest quality images were then chosen for training and testing purposes. Afterwards, artifacts and labels were removed from these 22 x-rays using the GNU Image Manipulation Program (GIMP). Segmentations of these images were created using the GIMP software package by drawing arcs and lines to represent bones and soft tissue. Both the x-ray images as well as the segmentations were then converted to squares and resized to 1024 x 1024 pixels in order to be input into the pix2pixHD model. The segmentations were further processed by having their RGB pixels programmatically converted to all 0s and 1s as the final step in order to utilize them as input to the pix2pixHD model. This work can be seen in Figures 2.2 and 2.3.

2.2.3 Training

Coding, debugging, and training of the pix2pixHD neural network was done by both Nicholas Chedid and collaborator Praneeth Sadda.

Given our limited dataset of 22 x-rays and in order to improve the accuracy of our pix2pixHD model, we utilized the leave-one-out cross-validation method, which is

FIGURE 2.2: X-ray Preprocessing: The first row contains the original x-ray images. The second row contains x-rays that were cleaned of artifacts and labels by using the GIMP software package. The third row contains the final version of the x-rays that have been programmatically resized into 1024 x 1024 pixel squares in order to be input into the pix2pixHD model.

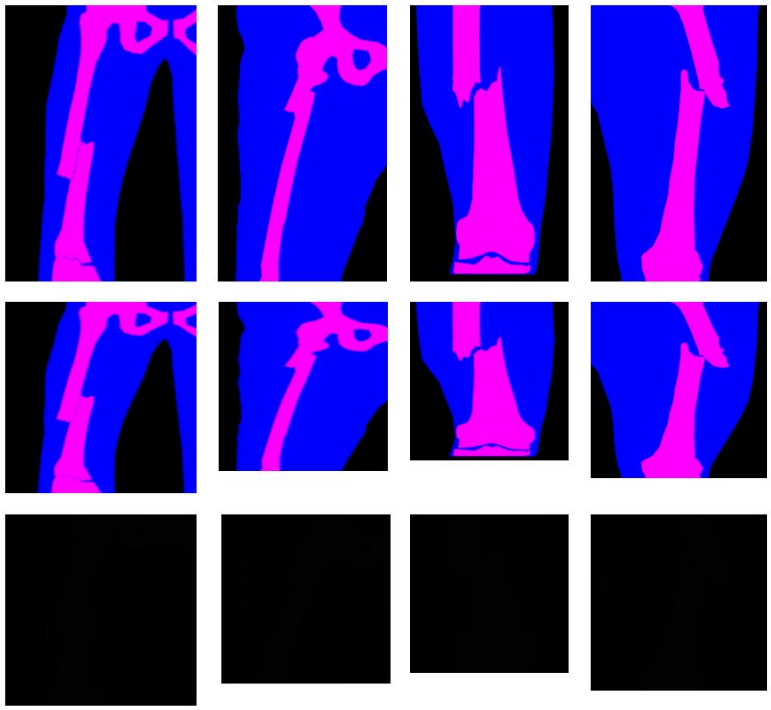

FIGURE 2.3: Segmentation Preprocessing: The first row contains the bone and soft tissue segmentations of the x-ray images created using the GIMP software package. The second row contains segmentations that have been programmatically resized into 1024 x 1024 pixel squares. The final row contains the resized segmentations, which have had their RGB pixels programmatically converted to all 0s and 1s in order to be input into the pix2pixHD model.

commonly used in machine learning research to improve accuracy for models trained on smaller databases [23, 24].

The leave-one-out cross-validation method works in the following manner. The machine learning model, pix2pixHD in our case, is trained on all data points (images in our case) except one which is reserved for testing. This process is then repeated until every data point has been used for testing. So in our case, we utilized 21 images for training while reserving 1 for testing. This process was repeated for a total of 22 variations in order to utilize every image as a testing image. This method helps improve performance when one's dataset is smaller by increasing the computational burden. Namely, the parameters of the model are re-calculated repeatedly according to the number of data points in the dataset. This means that if a machine learning model such as pix2pixHD were to need a certain number of calculations (n) proportional to the dataset size, then utilizing the leave-one-out cross-validation method would instead require n^2 calculations.

Fortunately this is the optimal method for this project since we are aiming to show the scalability of our method for many fracture types and are therefore intentionally more data limited than computationally limited.

Another advantage of the leave-one-out cross-validation method is that, by using nearly the complete dataset for training for each iteration, it is believed to give the most accurate estimate of the parameters, and, accordingly, the best estimation for how the model would perform on new data (generalizability) [24].

The training data was assembled by pairing the segmentations with their associated x-ray images while leaving one out for testing in the method described above.

Our networks were trained over 200 epochs. A learning rate of 0.0002 was used for the first 100 epochs. The learning rate was then decayed in a linear manner to zero over the next 100 epochs. Weights were initialized randomly following a Gaussian distribution with a mean of 0.

2.2.4 Postprocessing: Denoising

In order to further improve image quality and reduce noise artifacts, the images produced by the pix2pix model will then be input into a convolutional denoising autoencoder before being assessed for quality via the Visual Turing Test and the Structural Similarity Index Measurement Algorithm. Convolutional denoising autoencoders have already shown great utility for the denoising of medical images [25].

2.2.5 Visual Turing Test

Recruitment of radiologists into this study was done primarily by Nicholas Chedid. Code for displaying real vs synthetic x-rays to radiologists for assessment was written by my collaborator Praneeth Sadda.

A Visual Turing Test for assessment of synthetic image quality produced by GANs was proposed by Chuquicusma *et al*. [19]. We follow a similar methodology here to evaluate our synthetic images.

We designed 10 Visual Turing Test experiments. Our experiments will be conducted with three radiologists (one resident and two attendings). A radiology resident and two attending MSK radiologists including the division chief have been recruited The code for displaying the x-rays in these experiments has already been written.

Our experiments consist of 5 experiments of all generated x-rays and 5 of mixed generated and real x-rays. Each experiment contains 9 images in a 3 by 3 grid. Radiologists will be allowed to zoom in or change the view of the image. For each experiment the radiologists will be informed that the presented grid of images could consist of all generated, all real, or a mixture of images. Radiologists will then be asked to identify which images are real and which are generated. It is estimated that the total time for each radiologist to complete these experiments will be less than 30 minutes.

We will quantitatively measure the results from our Visual Turing Test and therefore the quality of our synthetic x-rays by measuring inter-observer variations, False Recognition Rate (FRR), and True Recognition Rate (TRR).

2.2.6 Structural Similarity Index Measurement (SSIM)

Assessment of pix2pix accuracy using the structural similarity assay will be done primarily by Nicholas Chedid.

A more quantitative assessment of image synthesis quality can be performed using a structural similarity index measurement (SSIM) as described by Wang *et al*. [26]. The SSIM is an objective method for assessing perceptual image quality. Previous methods for assessing image quality such as mean squared error (MSE) and peak signal-to-noise ratio estimate absolute errors, while SSIM is a quantitative model that predicts perceived image quality, which is of more value given our work.

Once post-processing using a convolutional denoising autoencoder is completed, I will run the SSIM.

2.3 Results

Our work has progressed through several stages. In my initial work I used plain GANs to synthesize x-ray images from segmentations. In order to further improve this work, we moved on to using the pix2pixHD method. This preliminary work utilized the pix2pixHD method without the leave-one-out method and was also prior to removal of artifacts and labels via the GIMP software package.

I presented these preliminary qualitative results (*i.e.* our synthetic x-ray images without the leave-one-out method, with minimal preprocessing, without postprocessing, and without the Visual Turing tests or SSIM data) as a poster titled, *Deep-Learned Generation of Synthetic X-Rays from Segmentations*, at the International Conference on

Medical Imaging and Case Reports in Baltimore, Maryland. These results can be seen in Figure 2.4. It can be seen that synthetic x-rays closely resembling their associated segmentations were able to be generated. However, ideally both improved resolution and reduction in artifacts could be achieved. To this end, I have increased our dataset from 13 to 22 x-rays and their segmentations and have removed artifacts and labels from the original x-rays. Additionally as mentioned in Section 2.2.3, I am now implementing the leave-one-out method and postprocessing using a denosing concolutional autoencoder.

Preliminary results incorporating these changes can be seen in Figure 2.5. As can be seen in said figure, artifacts have been decreased and resolution increased. Ideally current work on postproccessing should further increase image quality.

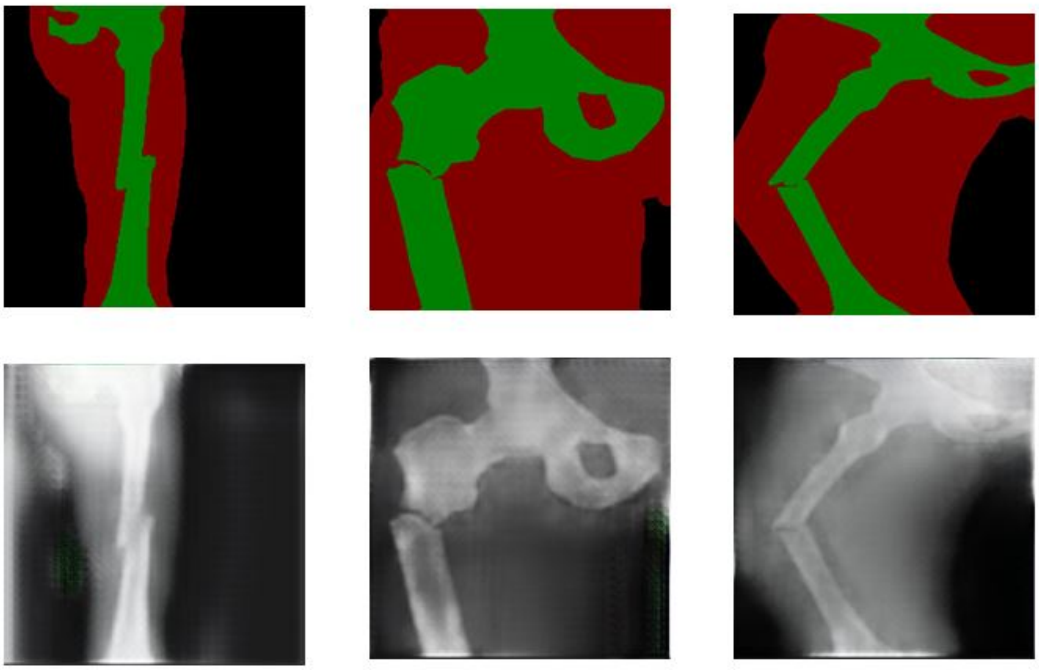

FIGURE 2.4: The top row displays our previous programmatically generated segmentations from x-ray tracings and the bottom row displays the corresponding synthetic x-rays generated from these segmentations using the pix2pix method prior to our implementation of the leave-one-out method and prior to the clean up of artifacts and labels from our x-ray images.

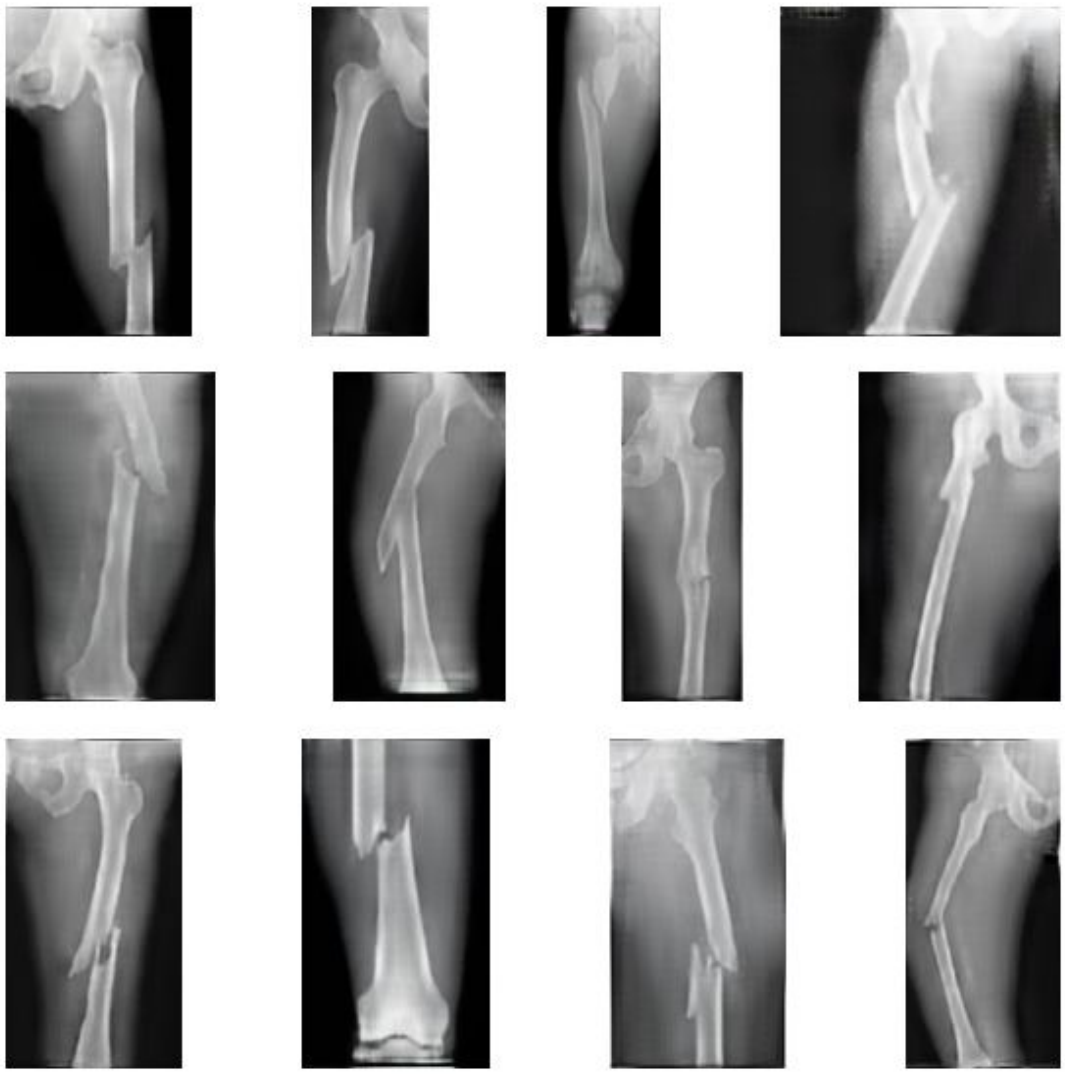

FIGURE 2.5: Examples of Generated X-rays: Here is a random selection of the synthetic x-rays generated using the pix2pix method with implementation of the leave-one-out method. Following the completion of postprocessing, quality should improve even further.

Given valuable feedback from that conference, I also decided to incorporate both the Visual Turing Test and the SSIM for evaluation of results.

2.3.1 Visual Turing Test

Once postprocessing using a convolutional denoising autoencoder is completed, these updated synthetic x-rays will be used to conduct our Visual Turing Tests as outlined in section 2.2.5. Code to conduct these tests has already been written. One of the 3 by 3 grids of generated vs real x-ray images to be used in the Visual Turing Test can be can be seen in Figure 2.6.

We envision displaying our results from the Visual Turing Test experiments (including FRR) in a manner similar to Chuquicusma *et al*.

FIGURE 2.6: Generated vs Real X-rays Visual Turing Test Grid: This is one of the 3 x 3 grids that will be utilized in the Visual Turing Tests. It consists of both generated and real x-ray images. For illustrative purposes here (i.e. in order to compare real and generated x-rays), the generated x-ray images have been given a pink highlighted outline.

2.3.2 Structural Similarity Index Measurement (SSIM)

Once post-processing using a convolutional denoising autoencoder is completed, I will implement a SSIM to quantitatively evaluate the perceptual similarity of these updated synthetic x-rays to the original x-rays as described in section 2.2.6.

SSIM results will be reported in the format seen in Table 2.1.

Algorithm	Avg. MSE	Avg. SSIM
pix2pixHD	97.1 ± 34.6	97.1 ± 34.6

TABLE 2.1: Example table for SSIM results.

2.4 Discussion

It is possible to synthesize realistic x-rays from procedurally generated segmentations with the pix2pix method. This can be seen qualitatively when comparing the synthesized x-rays to the original x-rays. The quality of these synthesized x-rays will be further quantified by our Visual Turing Test experiments and SSIM. Our study will be the first to quantify generated medical images to such a rigorous extent.

This is the first work that we know of to quantify synthetically generated medical images with a SSIM as well as the first to generate entire, synthetic x-ray images de novo using the pix2pixHD method (a higher resolution modality) as well as the first to measure the quality of entire, synthetic x-ray images using the Visual Turing Test.

We envision several ways in which the image synthesis method we have demonstrated may be useful in improving automated fracture detectors. Neural networks depend on supervised learning and are therefore limited by the availability of labeled data. As mentioned above, image synthesis can be useful for data augmentation. For example, when a fracture detector fails at classifying an image (e.g. as fracture present or absent), one would wish to ideally retrain that detector on multiple closely related

examples to improve its accuracy. Unfortunately one is limited by available data. Fortunately using image synthesis would allow to generate closely related examples to retrain the detector. Additionally our method could be a valuable way of generating training data for x-ray segmentation algorithms.

Additionally the generation of synthetic examples using GANs has been used for improving out of domain or novelty detection, meaning the ability of a classifier to recognize unknown inputs, which could be another way to utilize our method for improving automated fracture detectors particularly their generalizability.

Our trained synthesizer can also be used to better describe images (*e.g.* by learning features from the trained synthesizer).

Finally, these results were achieved by using only a small dataset compiled from readily accessible data from an internet search. This was done to address our goal of demonstrating how a pix2pixHD pipeline could be utilized as a rapidly scalable tool to aid in data augmentation of automated fracture detectors for many fracture types while reducing manual work and the need for very large databases.

The difficulty in acquiring and the manual work necessary to use a dataset of this size is far less than what would be needed to acquire and label a traditional dataset of several hundred to thousands of images for training a single fracture detection algorithm. Additionally, this tool can easily be used for different fracture types with much less work needed to switch between fracture types compared to the aforementioned traditional methods.

An interesting next step would be to train several neural networks at different tasks first by using small to moderately sized original databases and then by using those same databases further augmented with synthetic images generated via rapidly iterable customized versions of this pix2pixHD pipeline and then observe whether neural network performance was improved via this time and manual labor saving method.

Chapter 3

Neural Networks for Depression Screening & Treatment Monitoring

3.1 Introduction

3.1.1 Depression and it's Diagnosis

Depression is a disease with tremendous impact upon the human race. Globally, over 350 million individuals suffer from depression per year [27], and The Substance Abuse and Mental Health Services Administration estimates that approximately 16.2 million adults in the United States had at least one major depressive episode in 2016 [28]. One in five US adults are estimated to have experienced depression in their lifetime [29]. This problem is compounded in certain populations such as high functioning adults with demanding careers (such as medical trainees or professionals), adolescents, and the chronically ill [30].

In addition to being widespread, depression is associated with significant suffering, disability, and mortality. It is estimated that depression accounts for more "years lost" from disability than any other chronic disease by a wide margin [27]. This includes traditionally disabling conditions such as back pain, lung disease, and alcohol abuse. Recent studies have also demonstrated those with depression have higher rates

of obesity, heart disease, and diabetes [31, 32]. Finally, untreated major depression is well known to be the highest risk factor for suicide [33, 34].

While many psychopharmacologic and psychotherapeutic treatments exist to treat depression, self-recognition and diagnosis remain a formidable challenge. It is estimated that around two-thirds of all cases of depression in the United States are undiagnosed [35]. First, the diagnoses of depression and most psychiatric conditions are based on clinical assessment by a physician, leaving potential for bias and inter-physician variability. For example, according to the field trials of the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) there was high variability between physicians even when assessing the same patient for depression with an intraclass Kappa of 0.28 (95% CI: 0.20–0.35) [36, 37, 38]. Secondly, a formal diagnosis of depression requires a visit with a clinician, and individuals with limited access to healthcare have higher levels of undiagnosed depression [39]. Finally, the subjective nature of the assessment of depression and infrequent or difficult to obtain mental health appointments may lead to difficulties in longitudinal tracking of incremental changes or declines in psychiatric state.

After diagnosis, one complexity in the treatment of major depressive disorder is tracking response to therapy. After initiating a treatment, clinicians have to wait weeks in order to assess the effectiveness of that specific therapy. This often leads to lapses in care, and unfortunately even those with significant depression symptoms can be unintentionally neglected. It is estimated that over a quarter of those who complete suicide are undergoing treatment for their disease [33]. Clearly there remains an unfulfilled need to carefully monitor patients who are not so ill as to qualify for mandatory inpatient admission.

Given the prevalence and severity of undiagnosed and undertreated depression, readily available and widely-implemented screening tools to recognize depression symptoms are vital to enhance overall public health. Acknowledging this issue,

the U.S. Preventive Services Task Force (USPSTF) has recommended routine depression screening in primary care clinical practices [40]. Survey-based methods such as the Patient Health Questionnaire (PHQ)-2 and PHQ-9 are the most commonly used screening tools in the primary care setting [41]. However, these surveys take time, rely upon patient interaction with a primary care doctor, and do not adequately address the risk of developing depression symptoms in between often infrequent clinical visits. They also do not allow for monitoring of treatment in between difficult-to-obtain clinical visits. What is needed, and does not currently exist, is a solution to depression screening that is scalable, easy to administer, timely, and allows continual assessment.

Here, we propose a digital tool that uses an AI-powered facial and language recognition algorithm to screen for depression. The technology will use a 30-second video of a face, which will be recorded by the user, and notably can be integrated into any smartphone with a front-facing camera. The AI algorithm will then analyze the video and audio samples to provide a real-time measurement of that user's depression risk. Those who are found to be at high risk for a major depressive episode will be provided appropriate resources and/or a possible referral to a clinician or telepsyhiatry service for follow-up care. The technology will be user-friendly, accessible, and accurate.

3.1.2 Prior Work

Even though major depression is a tremendous public health concern, it has lagged far behind other diseases in recognition and strategy for research and cure. The study of artificial intelligence is a rapidly growing field and, while its use has been explored broadly in medicine, only recently has its use been directly studied to improve mental health. Experienced psychiatrists rely on microexpressions and auditory cues when evaluating patients for mental health disease; our hypothesis is that machine learning can detect similar patterns to add to the growing armamentarium of screening tools

in mental health. Others have suggested AI analysis can predict and diagnose depression [42], but most applications focus on using only a single modality, such as audio or text and do not track changes in mood longitudinally.

Previous work on automatic depression detection has been conducted on the Audio-Visual Emotion recognition Challenge (AVEC) datasets [42, 43, 44, 45, 46]. AVEC is currently the only available source of audiovisual data with associated depression ground truth labels. AVEC 2013 and 2014 provide video samples with corresponding BDI-II scores, while more recent AVEC challenges provide interactive video samples and transcriptions associated with PHQ-8 scores. We are focused on the BDI-II survey (thus the AVEC 2014 dataset), as it is the depression survey used most commonly for research purposes and because it is more granular (0-63 range of BDI-II scores versus 0-23 range on PHQ-8; these additional questions provide a possible source of metadata).

The AVEC 2014 dataset consists of 150 interviews designed to evaluate patients for depression. The dataset is divided into a 50 video training set, a 50 video development set, and a 50 video testing set. Each video is associated with a BDI-II score that can range from 0 to 63. The actual scores in the dataset range from 0 to 45, highlighting the skew towards non-depressed patients. A higher BDI-II score is correlated with a greater risk of depression. A depression cutoff of 14 has been used in our preliminary studies, per NIH guidelines (NINDS).

However, the accuracy of previous approaches using AVEC to study depression were limited by the use of outdated techniques. These methods require inefficient feature engineering, such as combining speech style, eye activity and head pose modalities in a Support Vector Machine [47], inputting hand-engineered features to Random Forests [48], using facial movement, head movement dynamics and vocal prosody with logistic regression [49], and utilizing topic modeling-based learning [50].

Neural networks have shown far greater accuracy than the aforementioned methods at analyzing complex behaviors, and our hypothesis is that the same will hold

true in depression. Thus, we believe that they may outperform these previous techniques in screening for depression. While, the advent of neural networks has offered a novel opportunity to improve accuracy, current approaches using neural nets have methodological flaws of their own. Two recent papers have used neural networks to predict PHQ-8 scores over audio, text and visual data with a best reported Root Mean Square Error (RMSE) of 5.4 on a 27 point scale [51, 52]. RMSE can be simply understood as the average distance of a predicted value from the true value. For example, an RMSE of 6 on a 27 point scale would indicate that the values predicted by the neural network are on average 6 points away from the true values. Using text and audio analysis, Hanai and colleagues [53] achieved an RMSE of 6.7 on the same scale. By leaving out video analysis, their peak RMSE and ability to expand to other areas of application is limited. Additionally the decision to use PHQ-8 as the ground truth misses out on the the granularity inherent to the more comprehensive BDI-II. The best predictive results were achieved by Jan *et al*. [54]. They used a multi-modal approach (video and audio), used BDI-II as ground truth, and have achieved the best RMSE in the literature (7.4 for the 63 point BDI-2 scale). However, further improvement for clinical utility will require an algorithm to reach an RMSE of 5 or less since a BDI-II change of five points is considered "minimally clinically significant" [55]. We believe that further improvement is hampered by the small size of the AVEC training set and that the way forward is not only refinement of the algorithm but also the creation of larger and more robust datasets.

3.1.3 Proposed Solution

We are developing multi-modal deep neural networks to predict BDI-II scores. We incorporate visual and audio neural networks into a single master neural network which combines the features and/or predicted scores of each individual model. We will also be incorporating text-based NLP analysis in the future as well. Crucially, our models

will make predictions using longitudinal data from multiple sessions, taking into account previous history and predictions, a novelty in this area of work. Additionally, our data collection will alleviate data sparsity problems associated with current deep learning models and enable more expressive models to be developed.

Our proposed technology seeks to integrate multiple inputs for predictive capability and will track individual user mood changes over time. Our approach will improve predictive power by accounting for intra-user variation by allowing each participant to serve as his or her own control and will help reduce the subjectivity and inter-user variability inherent to psychiatric diagnosis.

Our work is particularly significant because, while not everyone has access to psychiatric care or even knows if they have a developing mental health disorder, 75% of Americans use a smartphone [56]. With minimal intrusion into individuals' daily lives, we believe our solution can assess mood changes longitudinally, predict signs of an upcoming major depressive episode, monitor efficacy of treatment, and appropriately offer connections to care when needed. The technology will allow for more frequent assessment of disease status than possible with often infrequent clinic visits, and, by providing a readily available system, we can democratize mental healthcare to the most in-need populations. This technology has the potential to significantly impact the epidemic of undiagnosed depression.

Our technology may reveal novel insights into treatment efficacy, disease segmentation, and alleviating and exacerbating environmental factors. This may be particularly useful during evaluation of patients undergoing clinical trials of novel agents. Additionally, after developing an algorithm that can use a multimodal inputs to detect signs of depression, we hope to expand towards screening and treatment monitoring for other mental health disorders including burnout, bipolar disorder, schizophrenia, Alzheimer's, and Parkinson's Disease. Additionally, our active methodology innovation will allow us to expand our technology to the clinic and acute care settings. In

the clinic, we imagine our active video analysis functionality to provide valuable, non-invasive immediate diagnostic and clinical support information useful during telepsych, telemedicine, and in-person clinical encounters. Passive tracking, such as monitoring weeks of phone activity, will not be able to do this.

In addition to our goal of developing a "state-of-the-art" neural network to predict depression, we are also excited to develop a longitudinal audio-visual database correlated with depression scores. This sort of data would be the first in existence, since it has historically been very difficult to obtain. Obtaining videos that are tied to ground truth will generate a promising source of information that can be used in later studies outside of our own, and we believe that it could be the source of significant academic and translational productivity both for our project and others to come.

3.2 Methods

3.2.1 Overview

As described in the introductory sections, our overall neural network analysis can currently be split primarily into video and audio analysis.

Current input data consists of videos (including audio in German) from the academic AVEC 2014 database. It is a dataset of 150 videos (50 for training, 50 for testing, and 50 for development) of individual people speaking to a webcam. Each video is labelled with an individual's BDI-II score (0-63). As mentioned before, the actual scores range from 0 to 45, highlighting the skew towards non-depressed patients.

Currently we have two separate networks for audio and for visual data. We are planning to output a final score by learning a weight between the scores of the two networks and the desired BDI-II score. Another option we are considering exploring is to see how we can combine the audio features and the video features into a single neural network which combines features to learn BDI-II score.

The primary future aim of this work is to obtain orders of magnitude more data to significantly improve accuracy beyond that seen in the literature. Not only will the quantity of data be significantly increased (one of the best ways to improve neural net performance), but the quality will also be higher than anything previously gathered as will be further described in section 3.2.4.

3.2.2 Video Analysis

Work for building and tuning the video neural net architecture was done primarily by my colleague Michael Day.

A neural network was developed using Python scripts and programs using Keras packages running on a TensorFlow backend. Specifically the neural net created was a 19-layer convolutional neural network, a gold-standard neural network for image classification and computer vision tasks.

Our network has 19 weighted layers. We apply three sets of three convolution layers with each followed by 2x2 max-pooling layer. Our 2D convolution layers apply a 3x3 convolution with 32, 64, and 128 output filters. Then we flatten our 3D feature maps to a 1D feature vector, and feed those vectors through two 64-node fully-connected dense layers with 20% dropout each. The network ends with a single-node dense layer that will output a single predicted BDI-II score per set of inputs. Dense layers perform classification on the features extracted by the convolutional layers and downsampled by the pooling layers. Faces are extracted from input images using Haar cascade classifiers, and the faces are reduced to 48x48 grayscale images to improve runtimes and normalize inputs across different resolutions (per state-of-the-art practices). The model is compiled using an Adam optimizer and measures mean squared error as its loss metric. Activation functions primarily consisted of recitified linear units (ReLU).

Many different training iterations were run for hyperparameter tuning in order to optimize the neural network's accuracy on the test set. Tunable variables included: number of epochs, batch size, number of samples taken per video, and data augmentation features including: minimum face size, Haar cascade classifier scale factor, Haar cascade classifier minimum number of neighbors, horizontal flip, face image rescale size.

The optimal neural net was one in which the epochs were infinite and the model ceased training after no improvements were made in many (unsure of number) epochs. Only the best models having the lowest mean squared error were saved. The following parameters were ultimately used: batch size = 128, 500 samples taken per video, minimum face size = 30x30, Haar cascade classifier scale factor = 1.1, Haar cascade classifier minimum number of neighbors =5, horizontal flip = off, face image rescale size = 48x48.

Training was performed over 781 epochs for more than 72 hours on a desktop computer with one NVIDIA 980ti graphics card with 6 GB RAM.

3.2.3 Audio Analysis

Work for building and tuning the audio neural net architecture was done primarily by my colleague Alexander Fabbri.

Currently the audio architectures extracts MFCC features based on the Fourier transform of a speech signal (13 features per timestep) from the entire audio stream; additional features including Mel-frequency cepstral coefficients (MFCCs) will be extracted moving forward. Feature selection is then completed to determine the most relevant features. The features are then fed into a Gated Recurrence Unit (GRU) of dimension 100 as part of a Gated Recurrence Neural Network. By training on audio data labeled with BDI-II, our networks learn to output BDI-II scores for our test set. Regarding the datasets, we have data split into train, development and test sets.

Alex is currently implementing the following techniques to further improve accuracy. We plan to use cross validation to examine the results of our algorithms. In the cases where our data is skewed, we plan to explore undersampling and oversampling methods. We will also explore binning the samples to alleviate data sparsity. However, our end goal is to be able to perform fine-grained analysis of the spectrum of depression. Additionally, we plan to use L1 regularization, L2 regularization, their combination (also known as Elastic Nets) as well as dropout to allow better generalization of our neural networks. We will use standard optimization techniques such as gradient descent and its variants (Adam, AdaGrad, Nesterov gradient descent, RMSprop) as well as cutting-edge optimization techniques such as super-convergence and 1-cycle learning rate scheduling. Related to combining various modes of input, we would like to analyze textual input related to BDI-II scores, either by converting speech to text and then using Natural Language Processing techniques or through additional meta-data. More generally, we want to explore the effect of "out of domain" data on our neural network; how training with additional data taken from a different setting or pilot study affect our results. We would like to see how we can unbiasedly identify subgroups of patients. Also, we want to be able for our algorithms to continually learn and not have to re-train our algorithms every single time a new data point is added.

3.2.4 Pilot Studies for Gathering of First-in-Class Data

Work for designing and implementing these pilot studies was done primarily by Nicholas Chedid.

3.2.5 Need for Additional Data

The only currently openly available audiovisual datasets correlated to depression come from AVEC and this data has several significant weaknesses, which highlight many of the needs for additional data in this space.

1) The AVEC datasets correlated to BDI-2 scores consist of only 150 videos; it is a generally accepted maxim in machine learning that increasing training data is one of the most effective ways to improve algorithm performance.

2) The audio in these videos is only in German; having audio in several languages could enhance the generalizability of our algorithms.

3) This dataset, similar to many in medical research and the facial recognition space, consists of a relatively racially and ethnically monolithic participant population; facial recognition algorithms and medical research in general are hampered by non-diverse data which limits generalizability and applicability of such research. For example, prior studies show that the accuracy of facial recognition algorithms is sensitive to the demographic composition of both training and test data [57, 58]. Numerous papers describe the importance of diverse patient populations in medical studies in general as well [59, 60, 61]. Inclusion of minority participants in NIH funded research continues to be an ongoing issue; for example, since the NIH passed the Revitalization Act in 1993 to address this, less than 2% of the greater than 10,000 cancer clinical trials funded by the National Cancer Institute included sufficient minority participants to meet the NIH's own criteria [60].

4) The AVEC database does not contain longitudinal data i.e. multiple videos and BDI-2 scores from participants over time. We hypothesize that audiovisual data collected longitudinally allows for more accurate prediction of BDI-II compared to data from a single encounter. One possible reason for improved performance would be the ability to measure a patient's delta or relative change from assessment to assessment as opposed to relying on just an absolute BDI-2 score.

5) The AVEC dataset lacks extreme BDI-2 scores particularly at the higher end of scoring with the majority of scores clustered in the low to intermediate range. This lack of data significantly hampers the ability of any algorithms trained on this data to detect more significant depression.

We are in the final stages of IRB approval to begin our first pilot study at Ponce Health Sciences University in Puerto Rico to address many of these shortcomings as seen in section 3.2.7. We are also in the process of beginning several other pilot studies to further address these issues with emergency room patients (section 3.2.8) and medical residents (section 3.2.6).

3.2.6 Pilot Study with Medical Residents

I will begin with the development of our pilot studies with medical residents because this was the original motivation to develop an artificial intelligence mental health screening technology. As a medical student, it was easy to see the epidemic of depression and burnout among medical trainees. The evidence is sobering. In a meta-analysis of 54 studies by Mata *et al*. published in JAMA, the estimated prevalence of depression among resident physicians was 28.8% [62]. In another multicenter-study focusing on surgical residents by Williford *et al*. and published in JAMA surgery, the estimated prevalences of burnout and depression among surgical residents were 75% and 39% respectively [63]. For context, the estimated prevalence of depression among the general population is approximately 9% and, more specifically, between 12-13% in adults of residency age (25-34) [64]. Additionally, the Accreditation Council for Graduate Medical Education (ACGME) has recently rolled out new Common Program Requirements that require every residency program to address physician well-being, burnout, self-care, and mental health issues including requirements to improve mental health screening of residents.

In addition to the aforementioned significant need for work improving mental health in medical trainees, a pilot in this population would also address many of the shortcomings of the AVEC dataset as described in section 3.2.5. Specifically, it would allow us to greatly increase the quantity of our data, have data in English in addition to German, increase ethnic and racial diversity of our data, obtain longitudinal data, and likely gather more more varied data with possibly higher BDI-2 scores given that the

prevalence of depression in medical trainees is more than double that of their same-age peers (29% vs 12-13%) as described above within this same subsection.

Realizing this, I met with Dr. Rosemary Fischer, Director of Resident and Fellow Well-Being at Yale New Haven Hospital, to discuss implementing a pilot study to gather additional training data to improve the accuracy of our neural networks. Following these discussions, I went on to give an oral presentation of our proposed pilot study and technology at the Yale Innovation Summit where I also presented a poster titled, *Artificial Intelligence for the Detection of Psychiatric Disease*, which won Best Tech Poster.

The feedback from these presentations was invaluable. Incorporating this feedback, I then gave an oral presentation titled, *An AI-enabled mobile gaming platform for the early detection of psychiatric disease*, at the Stanford Medicine X ED conference. There I was fortunate to meet with the President of Ponce Health Sciences University (PHSU) in Puerto Rico to discuss our work. Given our mutual interest in conducting a pilot study at PHSU, he invited me to present a pilot proposal to the Deans of PHSU. This is further discussed in section 3.2.7.

Our specific aims are:

- Specific Aim 1: Confirm that our algorithm is capable of accurately predicting whether an individual has mild depression or greater, as defined by the BDI-II instrument. Criteria for Acceptance: Our algorithm will achieve a sensitivity of 75% and specificity of 85% in predicting a BDI-II score greater or equal to 14.
  - Rationale:
    * A BDI-II score of 14 or greater corresponds to depression ranging from mild to severe
    * Primary care physicians have a sensitivity of 51% and specificity of 87% at detecting depression without an instrument [65]

    - The most common screening instrument (PHQ-9) has a sensitivity of 74% and specificity of 91% at detecting depression [66]
    - Given the above three points, we are aiming to maintain comparable specificity while exceeding PHQ-9 and primary care sensitivity, which is our primary focus given that we are initially developing a screening technology

- Specific Aim 2: Demonstrate that longitudinal analysis of a user's audiovisual data can detect clinically important BDI-II changes. Criteria for Acceptance: Our algorithm will predict BDI-II scores with a root mean square error (RMSE) of less than 7.
  - Rationale:
    - The ability to measure changes over time is essential to identify at-risk subjects who transition between depressive and non-depressive states and monitor improvement in depressed patients undergoing treatment
    - An approximate 5-point change in the BDI-II score corresponded to a minimal clinically meaningful change in severity according to DSM-IV [55]
    - Given that a corresponding RMSE of 5 in real-world data is so far above current standards, we are aiming for an RMSE < 7 at the conclusion of Phase 1 of our STTR grant with the goal of reaching RMSE5 twelve months after concluding Phase 1

Eligible participants for this study include all residents at Yale New Haven Hospital. Recruitment will follow a consecutive sampling strategy with a recruitment goal of 150 residents. We aim to have 100 medical residents complete this study, with completion defined as completing one survey per month. Participants will be reimbursed upon completion of the study.

Study participants will be directed to download the Sol application via an email link. For participants who do not use a smartphone, a link to a weekly Qualtrics survey will be provided. Via the Sol app or Qualtrics, participants will be asked to record their answer to a simple question like, "How was your day yesterday?" There will be both a Spanish and English version of the application and survey. Users can choose which language they prefer. Care will be made to not ask questions that could be potentially triggering to participants. Following completion of the video response and successful upload (either automatically via the Sol app or manually uploaded through Qualtrics), each participant will be presented with a BDI-2 survey. Each response will be tagged to the associated video and delivered to secure, HIPAA compliant servers for subsequent analysis by the predictive AI algorithms.

We are currently applying for an NIMH STTR grant to fund this pilot and the Emergency Department pilot. Our aim is to begin this pilot in October as funds from the grant disburse. Pilot duration will be 12 months: 3 months enrollment, 6 months data collection, and 3 months data analysis.

3.2.7 Pilot Study at Ponce Health Sciences University

As mentioned in section 3.2.6 , I was invited to give an oral presentation to the President and Deans of PHSU in Puerto Rico to discuss a pilot study proposal. I presented this as: *An AI technology for the screening of depression in healthcare students*.

Recognizing the strong need for improved mental health among healthcare trainees and excited by our proposal, PHSU was excited to collaborate. Two psychology PhD students were recruited to administer and run the pilot study locally with Dr. Nydia Ortiz, Dean of the School of Behavioral and Brain Sciences and the former Director of the Puerto Rico Mental Health and Substance Abuse Administration, serving as the site PI.

The data gathered from the PHSU pilot will serve to further address the weaknesses of the AVEC dataset and also strengthen areas of our resident pilot. Specifically, it will allow us to increase our data size beyond what would be possible with the resident pilot, to add Spanish data to our English and German data, to even more significantly increase the racial and ethnic diversity of our data particularly among Hispanic participants, to obtain longitudinal data, and to likely gather more varied data with higher BDI-2 scores than found in AVEC. The resident pilot would still likely allow for gathering of higher BDI-2 scores given the high prevalence of depression among resident trainees.

Our pilot is titled: *An AI-enabled mobile application for the rapid assessment and risk stratification of depression in medical professionals*.

Our objectives are:

- Objective 1: Collect audiovisual data which can be used to identify patterns of facial and linguistic expression, as well as other relevant predictors, useful in the identification of depression in the study population
- Objective 2: Compare the effectiveness of an AI-powered facial and linguistic analysis algorithm to detect signs of depression as compared to a BDI-2 questionnaire.
- Objective 3: Validate the feasibility and utility of rapid, automated psychiatric risk stratification via a mobile interface

Eligible participants for this study include any healthcare students aged 21 or older, the legal age of medical consent in Puerto Rico, enrolled at Ponce Health Sciences University (PHSU). Recruitment will follow a consecutive sampling strategy with an estimated sample size of 300 - 400 students. We aim to have 150 students complete the study with completion defined as completing one survey per month. There will be two primary arms to this study that will have equal numbers of participants: one conducted in English and the other conducted in Spanish.

Study participants will be directed to download the Sol application via an email link. The application is a simple touch-based interface that will allow for the video recording of a user. For this study this app is meant to be a data gathering tool and not a diagnostic tool. For participants who do not use a smartphone, a link to a weekly Qualtrics survey will be provided. There will be both a Spanish and English version of the application and survey. During study registration, users will answer a 5-point Likert language proficiency question for both English and Spanish, with scores ranging from basic to native. Users who score 3 and above in only one language will complete the study in that language. Those who score 3 and above in both English and Spanish will be randomized and complete the study in either language.

Via the Sol app or Qualtrics, participants will be asked to record their answer every other week to a simple question such as, "How was your day yesterday?" Care will be made to not ask questions that could be potentially triggering to participants. Participants will also be asked each week if they are clinically diagnosed with depression or are in treatment for depression. Following completion of the video response and successful upload (either automatically via the Sol app or manually uploaded through Qualtrics), each participant will be asked to complete a BDI-2 survey We anticipate the entire interaction with the application will take approximately 5 minutes.

Each response will be tagged to the associated video and delivered to secure, HIPAA compliant servers for subsequent analysis by the predictive AI algorithm. Servers are specifically run through Amazon Web Services on their HIPAA secure platform. Only the study programmers will have access to the information on these servers, as they will use the data to improve the AI algorithm.

The IRB is in the final stages of approval. We aim to begin recruitment this March. Pilot duration will be 6 months.

3.2.8 Pilot Study with Yale Emergency Department Patients

As we prepared our NIMH STTR grant, we realized that one limitation of the AVEC data was still not being sufficiently addressed by our other two pilot studies: the lack of more extreme BDI-2 scores particularly at the higher end of scoring. This lack of data could significantly hamper the ability of our algorithms to detect more significant depression.A pilot study in the emergency department would allow us to selectively recruit depressed patients to address this.

One drawback of a pilot in the emergency department would be the lack of longitudinal data. Fortunately, this ability to provide longitudinal data is a strength of the two previously described pilot studies. This drawback can also be seen to provide some benefit. Namely, given the non-longitudinal nature of participation in this pilot, which would allow for lower reimbursement per participant, it will be much easier to have a significantly higher number of participants. So while the data may not be longitudinal, there is benefit to be gained from having a much greater variety of faces and voices for analysis.

Our specific aim is the same as specific aim 1 in section 3.2.6, since both pilots are part of the same NIMH STTR grant application:

- Specific Aim 1: Confirm that our algorithm is capable of accurately predicting whether an individual has mild depression or greater, as defined by the BDI-II instrument. Criteria for Acceptance: Our algorithm will achieve a sensitivity of 75% and specificity of 85% in predicting a BDI-II score greater or equal to 14.
  - Rationale:
    * A BDI-II score of 14 or greater corresponds to depression ranging from mild to severe
    * Primary care physicians have a sensitivity of 51% and specificity of 87% at detecting depression without an instrument [65]

* The most common screening instrument (PHQ-9) has a sensitivity of 74% and specificity of 91% at detecting depression [66]

* Given the above three points, we are aiming to maintain comparable specificity while exceeding PHQ-9 and primary care sensitivity, which is our primary focus given that we are initially developing a screening technology

Eligible participants for this study include all patients in the Yale New Haven Hospital Emergency Department and Crisis Intervention Unit (CIU) over the age of 18 with a clinic suspicion of depression. Exclusion criteria: excessive agitation or a history of schizophrenia or schizoaffective disorder. Enrollment and data collection periods will occur simultaneously as each participant will immediately complete the study after being enrolled (i.e. recording a video response to a question and completing the BDI-II survey). Completing those steps will take less than 5 minutes. Participants will be reimbursed upon completion of the study. The enrollment goal is 400 participants. The simultaneous enrollment and data collection periods will last for 7 months.

Study participants will be directed to complete a survey on either the Sol app or Qualtrics on one of the Emergency Department iPads designated for research. Participants will be asked to record their answer to a simple question like, "How was your day yesterday?" There will be both a Spanish and English version of the application and survey. Users can choose which language they prefer. Care will be made to not ask questions that could be potentially triggering to participants. Following completion of the video response and successful upload, each participant will be presented with a BDI-2 survey. Each response will be tagged to the associated video and delivered to secure, HIPAA compliant servers for subsequent analysis by the predictive AI algorithms.

We will be submitting an NIMH STTR translational grant on April 1st to fund this pilot and the medical resident pilot. Our aim is to begin this pilot in October as

funds from the grant disburse. Pilot duration will be 12 months: 9 months simultaneous enrollment and data collection and 3 months data analysis.

3.3 Results

Currently, several pilot studies have been designed. The first, our pilot study at Ponce Health Sciences University in Puerto Rico, is aimed at acquiring more diverse data. We are in the final stages of IRB approval and are aiming to begin recruiting in March. Using this new data, we hope to update our neural network results prior to the start of our other pilot studies.

Additionally, after several presentations and the associated feedback and the formation of the several collaborations over time, we have designed two other pilot studies incorporating medical residents and ED patients, which we are applying for an STTR grant for in April.

The measure most commonly used to test the accuracy of a neural network is root-mean-square error (RMSE), which is a measure of the average difference between a predicted and actual value (BDI-II score in this case). Our previous best results are displayed in 3.1. Our video neural network had an RMSE of 10.1 and our audio neural network had an RMSE of 11.6 with accuracies of 74% and 70% respectively. When considering these RMSE values, it is important to remember that the range of BDI-II scores is from 0 to 63. In addition, what is more important than getting the BDI-II score exactly correct is knowing clinically which individuals need help. Using a BDI-II score of 20–which indicates moderate depression–as a cutoff, our video analysis correctly binned users 74% of the time, and the audio analysis correctly binned users 70% of the time.

Since then we have improved our video neural network to an RMSE of 9.53 giving us the second best values in the literature. Currently we are implementing new architectures for our audio and video algorithms in the next few weeks and are aiming

to begin enrollment of participants in our Puerto Rico pilot this March; these updates will allow us to further improve our accuracy.

Another next step will be the submission of our NIMH STTR grant in April, which has already undergone many drafts.

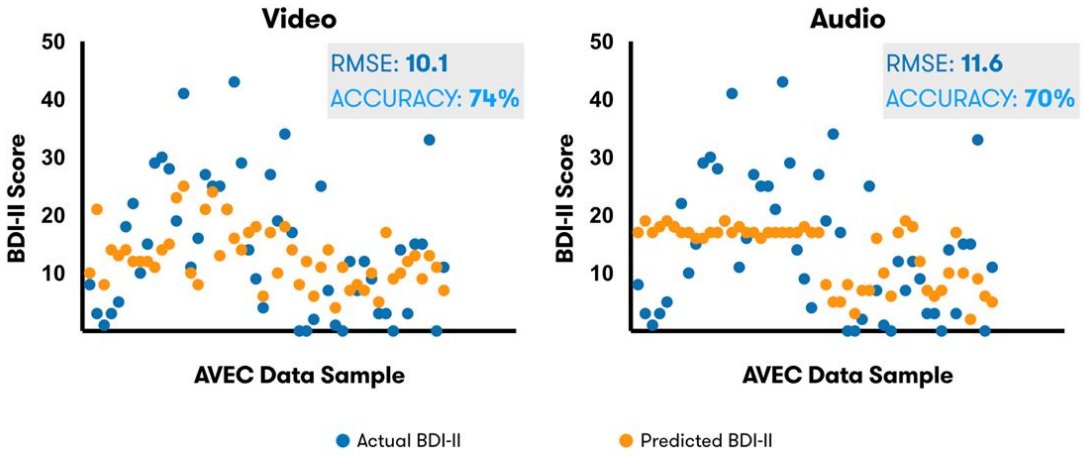

FIGURE 3.1: Video and Audio Neural Networks Accuracy

3.4 Discussion

Regarding, improving the performance of our neural nets: increasing the amount of input data as our pilot studies progress would likely result in an improved model from an increased quantity of data, quality of data (longitudinal, more diverse participants, and more diverse BDI-2 scores including more extreme values). Future enhancements may include a similar approach for change in pupil size over time, change in emotional sentiment over time, minimum and maximum emotional sentiment of an entire video, and other techniques including incorporating other meta-data such as time of day or location or lighting when video is taken. We also plan to incorporate natural language processing analysis of text from our audio recordings to ideally further improve accuracy.

As previously described, we will be creating the first longitudinal audio-visual database correlated with depression scores. Given the critical importance of data in machine learning and the fact that just our Puerto Rico pilot study will provide an order of magnitude more data than the non-longitudinal AVEC database, we are confident that we will be able to significantly outperform current best prediction tools. We also currently plan to incorporate the longitudinal data from our pilots studies in two ways. First, we plan to assess not just absolute BDI-2 scores but relative changes to the delta of their scores as another possible way for predicting depressive episodes. Additionally, a user's scores to any and all of these neural nets may be considered as a time-series. For example, considering the same user's video score over a period of weeks as they take the test multiple times.

In summary, we aim to develop a digital biomarker for depression. Developing such a digital biomarker for depression can serve as proof of concept for AI-based diagnosis, disease segmentation, and monitoring of other mental health disorders and of non-psychiatric diseases. Our platform will allow us to identify objective nuances in subjectively established psychiatric disease categories and facilitate personalized treatment regimens. Currently, the evaluation of chronic diseases such as depression relies on longitudinal evaluation. The active, video nature of our technology offers the potential to rapidly assess depression and other diseases instantaneously unlike current passive techniques. Furthermore, audiovisual samples may yield valuable insights into complex disorders such as burnout, bipolar disorder, schizophrenia, Alzheimer's Disease, and potentially non-psychiatric conditions including Parkinson's Disease, cerebrovascular accidents, and myocardial infarctions. Finally, our platform could be useful for screening, diagnosis, treatment monitoring, and patient selection and monitoring in clinical trials of novel agents.

Bibliography

[1] John L Kendall, Stephen R Hoffenberg, and R Stephen Smith. History of emergency and critical care ultrasound: the evolution of a new imaging paradigm. *Critical care medicine*, 35(5):S126–S130, 2007.

[2] R Andrew Taylor, Isabel Oliva, Reinier Van Tonder, John Elefteriades, James Dziura, and Christopher L Moore. Point-of-care focused cardiac ultrasound for the assessment of thoracic aortic dimensions, dilation, and aneurysmal disease. *Academic Emergency Medicine*, 19(2):244–247, 2012.

[3] M Kennedy Hall, EC Coffey, Meghan Herbst, Rachel Liu, Joseph R Pare, R Andrew Taylor, Sheeja Thomas, and Chris L Moore. The "5es" of emergency physician–performed focused cardiac ultrasound: a protocol for rapid identification of effusion, ejection, equality, exit, and entrance. *Academic Emergency Medicine*, 22(5):583–593, 2015.

[4] Elisa Ceriani and Chiara Cogliati. Update on bedside ultrasound diagnosis of pericardial effusion. *Internal and emergency medicine*, 11(3):477–480, 2016.

[5] Michael Blaivas, Daniel DeBehnke, and Mary Beth Phelan. Potential errors in the diagnosis of pericardial effusion on trauma ultrasound for penetrating injuries. *Academic Emergency Medicine*, 7(11):1261–1266, 2000.

[6] Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216, 2016.

[7] Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.

[8] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[9] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[10] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Tomonari Cho, Shinichi Kataoka, Akihiro Yamauchi, Yushi Ogawa, Yasuharu Maeda, Kenichi Takeda, Katsuro Ichimasa, et al. Artificial intelligence-assisted polyp detection for colonoscopy: Initial experience. *Gastroenterology*, 154(8):2027–2029, 2018.

[11] Jeffrey Zhang, Sravani Gajjala, Pulkit Agrawal, Geoffrey H Tison, Laura A Hallock, Lauren Beussink-Nelson, Eugene Fan, Mandar A Aras, ChaRandle Jordan, Kirsten E Fleischmann, et al. A computer vision pipeline for automated determination of cardiac structure and function and detection of disease by two-dimensional echocardiography. *arXiv preprint arXiv:1706.07342*, 2017.

[12] Johan PA van Soest, Andre LAJ Dekker, Erik Roelofs, and Georgi Nalbantov. Application of machine learning for multicenter learning. In *Machine Learning in Radiation Oncology*, pages 71–97. Springer, 2015.

[13] Rachel L Richesson, Jimeng Sun, Jyotishman Pathak, Abel N Kho, and Joshua C Denny. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artificial intelligence in medicine*, 71:57–61, 2016.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] *52 hospitals with the most ER visits*, 2016 (accessed January 16, 2019). https://www.beckershospitalreview.com/lists/50-hospitals-with-the-most-er-visits-2016.html.

[16] H. R. Guly. Diagnostic errors in an accident and emergency department. *Emergency Medicine Journal*, 18(4):263–269, 2001.

[17] Liam J. Donaldson, Amanda Cook, and Richard G. Thomson. Incidence of fractures in a geographically defined population. *Journal of Epidemiology & Community Health*, 44(3):241–245, 1990.

[18] Charles Andel, Stephen L Davidow, Mark Hollander, and David A Moreno. The economics of health care quality and medical errors. *Journal of health care finance*, 39(1):39, 2012.

[19] Maria JM Chuquicusma, Sarfaraz Hussein, Jeremy Burt, and Ulas Bagci. How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 240–244. IEEE, 2018.

[20] Dimitrios Korkinof, Tobias Rijken, Michael O'Neill, Joseph Yearsley, Hugh Harvey, and Ben Glocker. High-resolution mammogram synthesis using progressive generative adversarial networks. *arXiv preprint arXiv:1807.03401*, 2018.

[21] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *arXiv preprint arXiv:1711.11585*, 2017.

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*,

pages 694–711. Springer, 2016.

[23] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[24] Alan Moses. *Statistical Modeling and Machine Learning for Molecular Biology*. Chapman and Hall/CRC, 2017.

[25] Lovedeep Gondara. Medical image denoising using convolutional denoising autoencoders. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 241–246. IEEE, 2016.

[26] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[27] Kerri Smith. Mental health: a world of depression. *Nature News*, 515(7526):180, 2014.

[28] Substance Abuse, Mental Health Services Administration, et al. 2015 national survey on drug use and health. 2016.

[29] Deborah S Hasin, Renee D Goodwin, Frederick S Stinson, and Bridget F Grant. Epidemiology of major depressive disorder: results from the national epidemiologic survey on alcoholism and related conditions. *Archives of general psychiatry*, 62(10):1097–1106, 2005.

[30] AH Weinberger, M Gbedemah, AM Martinez, D Nash, S Galea, and RD Goodwin. Trends in depression prevalence in the usa from 2005 to 2015: widening disparities in vulnerable groups. *Psychological medicine*, 48(8):1308–1315, 2018.

[31] Ryan J Anderson, Kenneth E Freedland, Ray E Clouse, and Patrick J Lustman. The prevalence of comorbid depression in adults with diabetes: a meta-analysis. *Diabetes care*, 24(6):1069–1078, 2001.

[32] Floriana S Luppino, Leonore M de Wit, Paul F Bouvy, Theo Stijnen, Pim Cuijpers, Brenda WJH Penninx, and Frans G Zitman. Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies. *Archives of general psychiatry*, 67(3):220–229, 2010.

[33] Keith Hawton, Carolina Casañas i Comabella, Camilla Haw, and Kate Saunders. Risk factors for suicide in individuals with depression: a systematic review. *Journal of affective disorders*, 147(1-3):17–28, 2013.

[34] Takeaki Takeuchi and Mutsuhiro Nakao. The relationship between suicidal ideation and symptoms of depression in japanese workers: a cross-sectional study. *BMJ open*, 3(11):e003643, 2013.

[35] Sarah P Wamala, John Lynch, Myriam Horsten, Murray A Mittleman, Karin Schenck-Gustafsson, and Kristina Orth-Gomer. Education and the metabolic syndrome in women. *Diabetes care*, 22(12), 1999.

[36] John A Bilello. Seeking an objective diagnosis of depression. *Biomarkers in medicine*, 10(8):861–875, 2016.

[37] Darrel A Regier, William E Narrow, Diana E Clarke, Helena C Kraemer, S Janet Kuramoto, Emily A Kuhl, and David J Kupfer. Dsm-5 field trials in the united states and canada, part ii: test-retest reliability of selected categorical diagnoses. *American journal of psychiatry*, 170(1):59–70, 2013.

[38] Robert Freedman, David A Lewis, Robert Michels, Daniel S Pine, Susan K Schultz, Carol A Tamminga, Glen O Gabbard, Susan Shur-Fen Gau, Daniel C Javitt, Maria A Oquendo, et al. The initial field trials of dsm-5: New blooms and old thorns, 2013.

[39] Sharifa Z Williams, Grace S Chung, and Peter A Muennig. Undiagnosed depression: A community diagnosis. *SSM-population health*, 3:633–638, 2017.

[40] Albert L Siu, Kirsten Bibbins-Domingo, David C Grossman, Linda Ciofu Baumann, Karina W Davidson, Mark Ebell, Francisco AR García, Matthew Gillman, Jessica Herzstein, Alex R Kemper, et al. Screening for depression in adults: Us preventive services task force recommendation statement. *Jama*, 315(4):380–387, 2016.

[41] Douglas M Maurer. Screening for depression. *Depression*, 100:23, 2012.

[42] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM, 2013.

[43] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2014.

[44] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. Avec 2015: The 5th international audio/visual emotion challenge and workshop. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1335–1336. ACM, 2015.

[45] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2016.

[46] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja

Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9. ACM, 2017.

[47] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Matthew Hyett, Gordon Parker, and Michael Breakspear. Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing*, 2016.

[48] Aven Samareh, Yan Jin, Zhangyang Wang, Xiangyu Chang, and Shuai Huang. Predicting depression severity by multi-modal feature engineering and fusion. *arXiv preprint arXiv:1711.11155*, 2017.

[49] Hamdi Dibeklioğlu, Zakia Hammal, Ying Yang, and Jeffrey F Cohn. Multimodal detection of depression in clinical interviews. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 307–310. ACM, 2015.

[50] Yuan Gong and Christian Poellabauer. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 69–76. ACM, 2017.

[51] Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 53–59. ACM, 2017.

[52] Le Yang, Hichem Sahli, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Dongmei Jiang. Hybrid depression classification and estimation from audio video and text information. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 45–51. ACM, 2017.

[53] Tuka Al Hanai, Mohammad Ghassemi, and James Glass. Detecting depression with audio/text sequence modeling of interviews. In *Proc. Interspeech*, pages 1716–1720, 2018.

[54] Asim Jan, Hongying Meng, Yona Falinie Binti A Gaus, and Fan Zhang. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Transactions on Cognitive and Developmental Systems*, 10(3):668–680, 2018.

[55] Takahiro Hiroe, Masayo Kojima, Ikuyo Yamamoto, Suguru Nojima, Yoshihiro Kinoshita, Nobuhiko Hashimoto, Norio Watanabe, Takao Maeda, and Toshi A Furukawa. Gradations of clinical severity and sensitivity to change assessed with the beck depression inventory-ii in japanese patients with depression. *Psychiatry research*, 135(3):229–235, 2005.

[56] Jelle Kooistra. Global Mobile Market Report by Newzoo. 2018.

[57] P Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J O'Toole. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):14, 2011.

[58] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

[59] Sam S Oh, Joshua Galanter, Neeta Thakur, Maria Pino-Yanes, Nicolas E Barcelo, Marquitta J White, Danielle M de Bruin, Ruth M Greenblatt, Kirsten Bibbins-Domingo, Alan HB Wu, et al. Diversity in clinical and biomedical research: a promise yet to be fulfilled. *PLoS medicine*, 12(12):e1001918, 2015.

[60] Moon S Chen Jr, Primo N Lara, Julie HT Dang, Debora A Paterniti, and Karen Kelly. Twenty years post-nih revitalization act: Enhancing minority participation in clinical trials (empact): Laying the groundwork for improving minority clinical trial accrual: Renewing the case for enhancing minority participation in cancer clinical trials. *Cancer*, 120:1091–1096, 2014.

[61] Esteban G Burchard, Sam S Oh, Marilyn G Foreman, and Juan C Celedón. Moving toward true inclusion of racial/ethnic minorities in federally funded studies. a key

step for achieving respiratory health equality in the united states. *American journal of respiratory and critical care medicine*, 191(5):514–521, 2015.

[62] Douglas A Mata, Marco A Ramos, Narinder Bansal, Rida Khan, Constance Guille, Emanuele Di Angelantonio, and Srijan Sen. Prevalence of depression and depressive symptoms among resident physicians: a systematic review and meta-analysis. *Jama*, 314(22):2373–2383, 2015.

[63] Michael L Williford, Sara Scarlet, Michael O Meyers, Daniel J Luckett, Jason P Fine, Claudia E Goettler, John M Green, Thomas V Clancy, Amy N Hildreth, Samantha E Meltzer-Brody, et al. Multiple-institution comparison of resident and faculty perceptions of burnout and depression during surgical training. *JAMA surgery*, 2018.

[64] Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173, 2009.

[65] Mariko Carey, Kim Jones, Graham Meadows, Rob Sanson-Fisher, Catherine D'Este, Kerry Inder, Sze Lin Yoong, and Grant Russell. Accuracy of general practitioner unassisted detection of depression. *Australian & New Zealand Journal of Psychiatry*, 48(6):571–578, 2014.

[66] Bruce Arroll, Felicity Goodyear-Smith, Susan Crengle, Jane Gunn, Ngaire Kerse, Tana Fishman, Karen Falloon, and Simon Hatcher. Validation of phq-2 and phq-9 to screen for major depression in the primary care population. *The Annals of Family Medicine*, 8(4):348–353, 2010.