Master's thesis

Master's thesis

2021

Kristoffer Vanebo

NTNU
Norwegian University of
Science and Technology
Faculty of Information Technology and Electrical
Engineering
Department of Computer Science

Kristoffer Vanebo

Multi Image Bilirubin Estimation

May 2021

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

Multi Image Bilirubin Estimation

Kristoffer Vanebo

Computer Engineering
Submission date: May 2021
Supervisor: Donn Morrison

Norwegian University of Science and Technology
Department of Computer Science

Multi Image Bilirubin Estimation

Kristoffer Vanebo

May 2021

Foreword

This bachelor thesis was written during the spring semester 2021 at the Department of Computer Science (IDI) of the Norwegian University of Science and Technology (NTNU), with close cooperation with Picterus AS.

The reason I chose this thesis, was that it presented an unique challenge with regards to the exploring of the hot technology of multi input machine learning methods. It was also a thesis structured around what I consider a very worthy cause, in helping the field of neonatal jaundice diagnostics. It has been a memorable and educational experience, with lots of challenges with regards to studying the field of machine learning and ai.

First, I would like to express sincere thanks to my guidance councilor, Donn Morrison, for all help, and especially his patience with giving me the extended time to complete the writing of this thesis.

I would also give special thanks to Tárik S. Salem and Vegard Blindheim from Picterus, who provided lots of great feedback and help with accomplishing my goals. Writing this thesis on my own has been a daunting task, but their help severely lessened the burden. The access to the GPU server that Tárik set up for me, has also been priceless.

Lastly, I am also most grateful to Picterus AS as a whole for allowing me to work on this thesis, and that they granted me access to the dataset needed for this thesis. Working on this thesis has been a most enjoyable experience.

Trondheim, May 27. 2021

Kristoffer Vanebo

Problem

Picterus AS, a Norwegian company developing an application for diagnosing jaundice in newborns by smartphone images, are looking into different models to accomplish this task. As neonatal jaundice in the worst cases might be lethal, accurate and reliable diagnoses are required. The task of this thesis, is to look into machine learning models, which takes in multiple images of newborns at once, doing bilirubin prediction on these sets of images. The goal is to explore whether such models might achieve better accuracy in these estimations than existing models.

Abstract

Jaundice in newborn babies is a common occurrence. It is usually harmless, and caused by elevated levels of the pigment bilirubin. In severe cases, however, the condition may become lethal, or lead to permanent brain damage of the infant. Diagnosing this condition is therefore an important task. This is complicated in some regions of the world, as access to equipment for diagnosing is lacking. Therefore, much effort has been made into creating systems using digital images, which can be taken by smartphones, to estimate bilirubin levels. This thesis looks into the use of the hot field of multi input machine learning models, for the use in bilirubin estimation. Two such models are proposed, and compared both to a popular existing model, and a similar single image machine learning model. Though conclusive results were not obtained, these proposed models showed great promise.

Sammendrag

Gulsott hos nyfødte babyer er vanligvis et normalfenomen, som er forårsaket av økte nivåer i blodet av fargestoffet bilirubin. Unntaksvis kan disse nivåene stige til såpass høyde at tilstanden kan føre til permanente hjerneskader, eller i værstefall død. Diagnostisering av tilstanden er derfor svært viktig. Dessverre er dette ofte vanskelig å gjennomføre i enkelte verdensdeler, der tilgangen til nødvendig utstyr er liten. Derfor er det de siste årene vært mye forskning på å skape systemer som kan bruke digitale bilder fra smarttelefoner til å estimere bilirubin nivåer. I denne oppgaven undersøkes det om teknologi fra det nye og populære feltet om multi input maskinlære modeller kan brukes til bilirubin estimering. To slike systemer blir foreslått, og deretter sammenliknet med en populær eksisterende modell, og en liknende singel input maskinlæremodell. Selv om resultatene ikke kunne gi en difinitiv konklusjon, viser de foreslåtte modellene lovende egenskaper.

Contents

Foreword **i**

Problem **ii**

Abstract **iii**

Sammendrag **iv**

1 Introduction **1**

1.1 Motivation 1

1.2 Goals of this research 2

1.3 Structure of the thesis 3

Acronyms **4**

Glossary **5**

2 Background **6**

2.1 Neonatal Jaundice 6

2.2 Jaundice Diagnostics 8

2.2.1 Transcutaneous Bilirubin (TcB) 9

2.3 Related work 10

2.3.1 Smartphone Solutions for Jaundice Detection 10

2.3.2 Existing Machine Learning Approaches 12

2.3.3 Support Vector Machine (SVM) 12

2.3.4 Convolutional Neural Network (CNN) 13

2.3.5 Multi Image Machine Learning Models 14

3 Method **16**

3.1 Dataset 17
3.2 Baseline Models 19
3.2.1 Support Vector Regression (SVR) Model 19
3.2.2 Single Image CNN (SICNN) 19
3.3 Multi Image Machine Learning Models 20
3.3.1 Simple Multi Image CNN (SMICNN) 20
3.3.2 Complex Multi Image CNN (CMICNN) 21

4 Results **22**
4.1 Performance Comparison 22
4.2 Hyperparameter Optimization 28

5 Discussion **29**
5.1 Performance Comparison 29
5.2 Hyperparameter Optimization 31
5.3 Project Reflection 32

6 Conclusion **33**
6.1 Further work 34

References **I**

Appendix A - Full Network Graphs of CNN Models **VI**

List of Tables

3.1 Distribution of captured images over trials. 16
3.2 Hyperparameters used for grid search on SVR model 19

4.1 Performance comparison between the machine learning models. 22
4.2 Caption 23
4.3 Optimal hyperparameters for the SVR model, found with a grid search. 28
4.4 Best found hyperparameters for the CNN models using Optuna [38] random hyperparameter search. 28

List of Figures

2.1 Norwegian jaundice treatment determination chart. It shows the age and weight bounds for the different treatments for EHB. 8

3.1 Distribution of bilirubin values. 17
3.2 Skin color distribution, color coded by trial and plotted in RGB space. 18
3.3 Diagram showing the SMICNN. Multiple image inputs is fed through separate convolution layers, before merging and completing the Wide ResNet50-2. 20
3.4 Diagram showing the CMICNN. Multiple image inputs is fed through Wide ResNet50-2 models, before merging and completing a couple fully connecting layers. . 21

4.1 Linear regression of the correlation between predicted values by the SVR model and actual values. 23
4.2 Loss charts for the SICNN, where loss is evaluated in MSE. 25
4.3 Loss charts for the SMICNN, where loss is evaluated in MSE. 26
4.4 Loss charts for the CMICNN, where loss is evaluated in MSE. 27

6.1 Tensorboard graph of the full SICNN network, which is just an implementation of Wide Resnet-50-2. VII
6.2 Opened up Tensorboard graph of the SICNN, showcasing the block based structure, with layers within Bottlenecks, and Bottlenecks within Layer blocks. VIII
6.3 Tensorboard graph of the full SMICNN network, which is a modified Wide Resnet-50. It takes 6 images as a set, does a single convolution on each, before merging the outputs and completing the network. IX
6.4 Tensorboard graph of the full CMICNN network, which is a modified Wide Resnet-50. It takes 6 images as a set, sends each through a full Wide Resnet-50 network, before merging and doing a couple fully connected layers. X

Chapter 1

Introduction

1.1 Motivation

Neonatal jaundice is the name of a common condition in newborn babies, where their skin and whites of the eyes is yellowed. The cause of this condition, is an elevated level of a pigment called bilirubin. This condition is usually completely harmless, and a reflection of a normal transitional phenomenon that usually clears out after 10-14 days [1]. Sometimes, however, the accumulation of unconjugated bilirubin might rise excessively, leading to the condition known as extreme hyperbilirubinemia (EHB), which is a cause for concern. This condition may lead to permanent brain damage, or even death [2]. Bhutani et al. [3] estimated that in 2010, around 114 000 infants had died from lack of treatment for EHB, and many others suffered from disability caused by permanent brain damage. More than three quarters of these deaths occurred in sub-Saharan Africa and South Asia; some of the poorest regions in the world.

Current standards for diagnosing jaundice, are the measuring of either total serum bilirubin (TSB) or through transcutaneous bilirubin (TcB). Since measuring TSB requires a blood sample to be taken, this is a quite invasive method for diagnostics. TcB provides a less invasive approach, by using bilirubinometers which estimates the bilirubin level by scanning the skin of the infant. There are major shortcomings to these methods, however, as they both require skilled personnel and expensive equipment: either laboratory equipment or the aforementioned bilirubinometers. Because of this, these methods are not widely available in poorer regions of the world, where expensive equipment is hard to come by and there is a lack of trained personnel. This leads to visual assessment usually being the only method available for recognizing jaundice, but as studies have found, this is an unreliable form of diagnostics [4, 5]. Therefore, a need for affordable, non-invasive methods for estimating bilirubin has been recognized [6], as these regions is also some of

the most vulnerable to jaundice and EHB.

As mobile phones has become a necessity in modern society, the availability and access to them has steadily increased, helping the field of mobile phone health (mHealth) emerging. This field, which uses mobile phones in the service of medical and public health, has grown to become a field of interest with great potential [7]. Drawing upon the emergence of this field, several publications have proposed novel ways of diagnosing neonatal jaundice through digital images, often captured by smartphones [8–22]. One of the more recent efforts, Aune et al. [20], showed that their physics based system achieved bilirubin estimates correlating closely to the measured TSB. Several of these studies utilized machine learning to estimate bilirubin values, though most only applied different forms of regression models. As machine learning is an expansive field with many possibilities, this is an area that still could be explored even further. This is especially true, as studies have shown the aforementioned methods hold great promise [12].

1.2 Goals of this research

As previously mentioned, the use of machine learning for bilirubin estimation, is a field open for exploration. Though many of these proposed systems uses machine learning approaches for bilirubin estimation [8–16, 18, 19, 21, 22], most rely on different forms of regression models, meaning that this is still a field with many unexplored possibilities. Machine learning has grown expansive over the last few years, with many different methods and technologies. One of the areas currently unexplored for use in bilirubin estimation, is multi-input methods. These have been shown to effective in, amongst others, flower grading [23] and image super-resolution reconstruction [24]. The goal of this thesis is to explore the possibilities of multi image processing machine learning methods for bilirubin estimation. To this end, the research question asked is:

> *R.Q: Can multi image machine learning methods compete with current single image machine learning methods for bilirubin estimation in accuracy?*

1.3 Structure of the thesis

The **Introduction** provides the motivation for the thesis, as well as presenting the research questions to be answered. In the **Background** chapter relevant material appertaining to neonatal jaundice, basic theory behind the methodology used in the thesis and related work is presented. Starting with some background on neonatal jaundice, its treatment and diagnostics, before discussing related work within the field of mobile phone health (mHealth). Lastly theory behind the machine learning methods used in this thesis is also presented. These methods is then detailed in the **Method** chapter. This chapter breaks down the implemented machine learning methods and how they were constructed. The findings from the experimentation with the machine learning models is then presented in the **Results** chapter. Here the results from the different models is presented and compared. The **Discussion** chapter then analyses the results, and reflects upon the viability of the methods. How these results relate to the research question is also discussed. Lastly, the **Conclusion** answers the research question.

Acronyms

CMICNN Complex Multi Image CNN. viii, X, 21, 22, 24, 27, 28, 30

CNN Convolutional Neural Network. 4, 12–16, 19, 22, 24, 28–34

EHB extreme hyperbilirubinemia. viii, 1, 2, 6–9, 17

kNN k-Nearest Neighbor. 11

mHealth mobile phone health. 2, 3, 10

MSE Mean Squared Error. viii, 22–27, 30

RBF Radial Basis Function. 13, 19

ResNet Residual Neural Network. viii, 14, 19–21, 30, 34

SICNN Single Image CNN. viii, VII, VIII, 20, 22, 24, 25, 28–30, 33

SMICNN Simple Multi Image CNN. viii, IX, 20, 22, 24, 26, 28, 30, 31

SVM Support Vector Machine. 12, 13

SVR Support Vector Regression. vii, 11, 13, 16, 19, 22–24, 28–31, 33

TcB transcutaneous bilirubin. 1, 5, 8–11

TSB total serum bilirubin. 1, 2, 8–10, 12

Glossary

artificial intelligence Computer systems able to perform tasks which usually requires human intelligence. Commonly has such abilities as: visual perception, speech recognition, decision making, and translation. 5

artificial neural network A form of computing system which is meant to simulate the way the human brain functions, by having connected nodes which mimics neurons in the brain. It is a founding structure upon which artificial intelligence is built. 13

bilirubinometer Device using multiple wavelengths of light to measure transcutaneous bilirubin (TcB) through light absorption properties of bilirubin. 1, 9

KFold cross-validation A form of cross-validation where the dataset is randomly shuffled and subdivided into k groups, where for each group the machine learning model is trained on the remaining groups, and tested on the hold out. 19–22, 29

Chapter 2

Background

2.1 Neonatal Jaundice

The condition of jaundice is a common occurrence in newborn babies, and usually not a cause for concern. It affects around 60%-80% of infants [1], and is often recognizable by yellowing of the skin and whites of the eyes, which is caused by a yellow compound called bilirubin. Bilirubin is a product of the breakdown of red blood cells, which is usually conjugated by enzymes in the liver and passed out of the body through faeces. In newborn babies, there is often the case that the liver has not attained full efficiency at conjugating all the bilirubin, leading to a build up of unconjugated bilirubin in the infants blood. This is because the placenta has the job of removing the unconjugated bilirubin from the fetus while it is growing in the mothers womb. Therefore, it takes some time for the enzymes to properly activate in the liver, which only starts after the umbilical cord is cut. This is rarely an issue, as the condition usually clears out within 10-14 days, when the enzymes in the liver gains full functionality to conjugate the bilirubin [1].

The cases where jaundice is a cause for concern, is when the levels of bilirubin in the blood rise excessively, in which case it leads the condition of extreme hyperbilirubinemia (EHB). In this condition, unconjugated bilirubin may accumulate in the basal ganglia of the brain, causing severe damage because of its neurotoxicity [2]. The damage caused by excessive amounts of bilirubin in the brain may lead to permanent brain damage, and in the worst case scenarios, even death. The form of brain damage suffered by those who survive is known as kernicterus, and may manifest itself as cerebral palsy, hearing loss or learning disabilities. Therefore, these extreme cases require treatment, where the goal is to reduce the level of bilirubin in the infant's blood. The most common method for treatment is called phototherapy, where the infant is irradiated with blue light. The blue light is absorbed by bilirubin, which in turn breaks down into substances that can be processed

and excreted. These processes are called photooxidation and photoisomerization, where the former completely breaks apart the bilirubin molecule, and the latter deforms it into a less toxic compound [25]. Phototherapy is accomplished through two main methods: either the infant's eyes is covered before it is placed under halogen spotlights or fluorescent lamps, or it is wrapped in a so called "biliblanket", which is a blanket laid with fiber-optic cables which shines blue light onto the infant's back and body.

As discussed, the potential harm caused by neonatal jaundice is severe, yet with proper intervention it can relatively easily be treated. However, neonatal jaundice and EHB is still a major problem worldwide. Researcher found that in 2010, more than a 100 000 deaths were caused by EHB, whilst around 83% of survivors of kernicterus suffered from at least one impairment [3]. Of these deaths, 6% belonged to the Eastern Europe/Central Asia region, 7% to Latin America/Caribbean, 35% to sub-Saharan Africa, and 39% to South Asia. On a worldwide basis, with around 134 million live births, 85/100 000 deaths were associated with jaundice. In these regions, the combined prevalence were 119/100 000, compared to the 1/100 000 in high-income countries [3]. This disparity seems to stem from the lack of timely access to resources for both identifying the severe cases, and treating them [26]. The previously discussed treatments for jaundice requires both trained personnel, as well as expensive equipment, both of which is hard to come by in the poorer regions of the world. Methods for jaundice diagnostics, which will be discussed further, also requires skilled, trained personnel and expensive equipment; explaining why EHB cases are much more adverse in these regions.

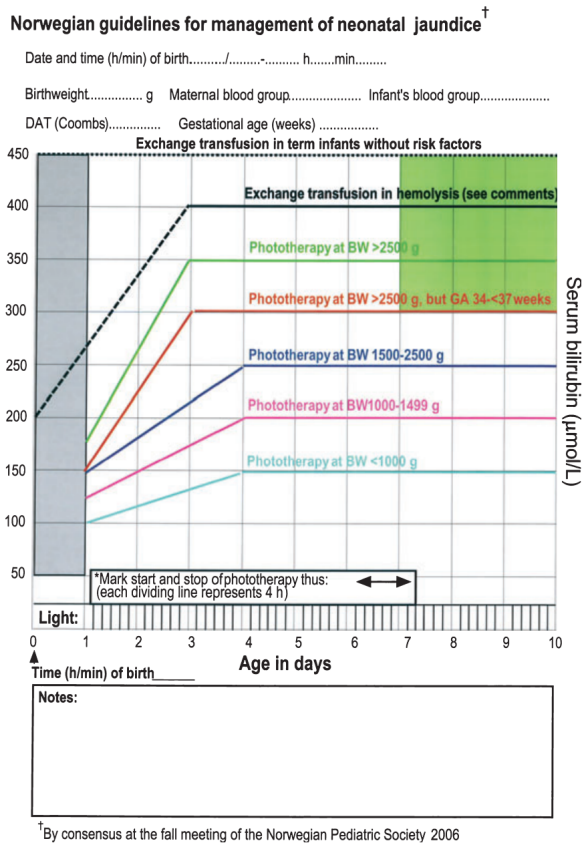


Figure 2.1: Norwegian jaundice treatment determination chart. It shows the age and weight bounds for the different treatments for EHB.

2.2 Jaundice Diagnostics

The current standards for measuring the level of bilirubin in the infant, and thereby diagnosing jaundice, is either measuring total serum bilirubin (TSB) or transcutaneous bilirubin (TcB). The most accurate way to ascertain the level of bilirubin and diagnosing the severity of the condition, is to measure TSB. To do this, a blood sample is taken; usually by pricking the infants heel. Then the amount of bilirubin is measured in the part of the blood called serum, which is the liquid part not containing any clotting factors or blood cells. The concentration of bilirubin in $\mu mol/l$ determines whether the infant need treatment, and what sort of treatment. This assessment is dependant upon the age and body weight of the baby, with certain thresholds as shown in figure 2.1.

2.2.1 Transcutaneous Bilirubin (TcB)

As taking a blood sample to measure TSB is quite intrusive, TcB has become the more common method for diagnosing jaundice. This method, however, is not a precise measurement for the levels of jaundice. Using light absorption properties of bilirubin, TcB measuring devices, called bilirubinometers, emit light at different wavelenghts, and record the reflection properties as a function of wavelengths. The measurements can then be used to calculate a TcB value, based upon differences in optical densities for different wavelengths of light. The main drawback, is that this method measures the bilirubin in the extravascular tissue, not the blood. There is agreement upon a linear correlation between TcB and TSB, but there is uncertainty of the reliability of the estimations. It is at least certain that TcB is a good way of screening for jaundice, and thereby identify which cases needs to undertake a TSB test [27].

Several bilirubinometers are in production and use in hospitals around the world. Two of the most state-of-the-art bilirubinometers are the BiliChek device and the Dräger JM-series, where the newest is the JM-105 model. The main difference between the two bilirubinometers is that the Dräger uses a dual wavelength of 460 nm and 540 nm, wheras BiliChek uses data from multiple wavelength readings between 400 and 760 nm [28]. The multiple wavelength readings of the BiliChek allows for corrections for differences in skin pigmentation and hemoglobin, it does have a drawback however, as it needs a disposable tip for each measurement, increasing cost. BiliChek and the predecessor of the Dräger JM-105, the Dräger JM-103, was found to perform equivalently, with comparable performance [29], and is considered reliable screening tools for EHB.

2.3 Related work

Mobile phone health (mHealth) is an emerging field of study, which utilizes mobile phones and tablets to provide support to both medical and public health practices [30]. The recent years have had rapid advances in mobile phone technology, as well as increasingly more widespread availability of mobile phones, which has helped garnering interest into this field. This is a field with great potential, especially as these technologies have been shown to be accessible even for low- and middle-income countries [7]. There has been great interest in utilizing this field in the effort to diagnose jaundice, as this may offer affordable means to this end. Therefore, several researchers and publications have proposed novel systems using digital images, often capture by smartphones, to try to estimate bilirubin and assessing jaundice [8–22].

2.3.1 Smartphone Solutions for Jaundice Detection

Several smartphone systems for assessing and detecting jaundice have been proposed, which have focused on using digital images captured by smartphones, to either predict bilirubin levels, or assess jaundice in some other way. One of the first of these, were the BiliCam system presented in 2014 by De Greef et al. [8], which were shown to have the potential of becoming a great screening tool comparable to TcB measuring devices. This system used an advanced algorithm for predicting bilirubin levels, which color corrected the digital images before doing a feature extraction and regression on the color corrected RGB values. To color correct the taken images, the BiliCam system both normalized the RGB values, and used a color calibration card placed on the infant before capturing the image. Using this card, they were able to correct for hue and saturation differing because of different illumination, as well as white balance the taken images. Using both an image with flash, and one without, they converted these color corrected images into both the YCbCr and Lab color spaces, calculating a mean value for each color space, giving each image a total of 9 features. Pooling these together, and adding a color gradient in the RGB space for the "with flash" image, they were able to extract 21 features i total. Using these features, they trained a custom regression algorithm, which consisted of eight regressions, using five different regression algorithms, where the output were determined by an agreement in the ensemble. Using the maximum and minimum values, the output gave the mean value where the difference were less than the empirically derived threshold of $2\,\mathrm{mg/dl}$, and the second highest value (90th percentile) when above. This helping to bias the model towards rather giving false positives than false negatives. A later study on a larger study sample was performed by Taylor et al. [12], found the BiliCam to produce quite accurate TSB estimates, that could be effectively used for screening newborns for jaundice.

Another similar system was presented in 2016 by Aydin et al. [11]. Like the BiliCam sys-

tem, this also used a color calibration card, which was printed with an advanced laser printer, on uncoated paper. First the calibration card was used in a segmentation process, where it was used to compare colors with each of the segments of the captured images. Undesired segments were discarded, and the remaining sent through a Gauss filter. After this, noise reduction and thresholding were applied to the images, before using pixel similarity to restore some lost colors to the images, and white balancing to counter environmental and lightning differences. This was followed by feature extraction, where they color mapped the images to YCbCr and Lab color spaces, and calculated average values for each color channel, giving 9 features. Gradient calculation was also applied for each of the RGB channels, leaving 3 more features, totalling 12. After this, both k-Nearest Neighbor (kNN) and Support Vector Regression (SVR) algorithms were applied. As with BiliCam, a threshold value were applied, in this case 1.91, where the mean value was selected when the difference between maximum and minimum were below the threshold, otherwise the 90th percentile of the difference were used. Aydin et al. [11] only had a small dataset of 40 jaundiced infants, but achieved a linear correlation of 0.81, and a success rate of 85%.

In 2020 Aune et al. [20] presented a somewhat different system. Based upon previous research by Randberg et al. [31], they were able to create a library of simulated reflectance spectra of newborn skin using diffusion theory. The model also factored in influence from such factors as: skin reflectance, skin thickness, haemoglobin and melanin levels. They were also able to adapt the model to work with images taken by smartphone cameras, which only feature RGB. As with previously discussed systems, Aune et al. [20] used a color calibration card. Selecting colors which reflects light similarly to newborn skin, they printed these cards with spectral printing on a seven ink printer. After printing, each card were measured with a color spectrometer, storing the readings for use in later calibrations. The system then worked by using the color calibration card to calibrate smartphone images, which were then compared to a large database of color and bilirubin pairs. This database could either run through cloud based storage, or be downloaded directly to the device for offline use. Further Aune et al. [20], found in a test with 302 newborns, where 76 had severe jaundice, that their system had a sensitivity of a 100% and a spcifity of 69% for severe cases over 250 $\mu mol/l$. The correlation between image estimates and TcB were .81, and they were able to conclude with that because of its specificity and accuracy, it could provide a screening tool for neonatal jaundice.

2.3.2 Existing Machine Learning Approaches

Most of the currently proposed smartphone systems for jaundice detection and diagnosis, features machine learning methods to estimate bilirubin values. The use of regression models were the most common approach, as these systems traced correlation between observed features in the digital images and the measured TSB values. Most of these systems follow along the lines of: some sort of color correction, into some feature extraction before applying a form of regression model. There have been some work, however, to try and use the ever more popular machine learning method of Convolutional Neural Network (CNN) to detect jaundice. Falk and Jensen [16] tried this approach, but were hindered by having problems with both configuring their model, as well as having a training set which were not large enough to properly train all the layers of the network. Chakraborty et al. [22] proposed a CNN which could detect images of newborns with jaundice. This proposed system were quite limited, however, as it only produced a binary model which could only classify images as either jaundice or not jaundice. This meant that the model could not be used in any way to either estimate the TSB value, or assess the severity of the jaundice in any other way. As such, this study should be treated only as a proof of concept to usability of CNN to detect differences in images of jaundice.

2.3.3 Support Vector Machine (SVM)

First proposed and formally described by Cortes and Vapnik [32] in 1995, Support Vector Machine (SVM)s are one of the most robust prediction methods within the field of supervised learning. Based upon statistical learning theory, SVM is a linear classifier, which tries to divide a n-dimensional feature space into two parts, belonging to two separate classes. To accomplish this, the SVM accepts n-dimensional data belonging to these two classes, and attempts to find a (n-1)-dimensional hyperplane which separates the data. As several hyperplanes might accomplish this devision, the SVM tries to find the one with the largest possible possible margin; that is to say, the largest distance to the nearest datapoint within each class. This is accomplished through whats called "training the machine", where an optimizer function adjusts the hyperplane to maximise its distance to the nearest datapoints. These datapoints are called **support vectors**, and are the only data stored after training, both giving SVMs their name, as well as reducing storage requirement and computational time for testing. This hyperplane then defines the classification model, and can gauge the position of new unlabeled data within the feature space, and thereby classify these new datapoints.

There are three main advantages to SVMs: the first, as mentioned, is the reduced storage and computational cost stemming from the support vectors. The second is that SVMs are able to use something known as "the kernel trick". This trick lets SVMs use a non-linear kernel to map the

datapoints into a higher dimensional feature space. This allows for data which previously not were linearly separable to maybe become linearly separable, thus reducing complex problems into linear classification. Some of the most popular kernels are as follows: RBF, Sigmoid and Polynomial. There is no strictly "best" kernel, and each kernel might be better for different applications. There is therefore a bit of trial and error finding the one most suitable for a certain case. The third advantage to SVMs are a result of the hyperplane optimization. As SVMs try to maximise the distance between the hyperplane and both classes, it is therefor able to better generalize as it is forced to sit at an equal and maximized distance from both classes.

Support Vector Regression (SVR)

Regression is a generalization of the classification problem, where a continuous-valued multivariate function is estimated, instead of an output from a finite set. Drucker et al. [33], proposed in 1996 such a generalization of SVMs, called Support Vector Regression (SVR). To accomplish this generalization, an ε-insensitive region is applied around the function, defining an ε-tube. This then redefines the problem into one where the goal is to find the tube which best approximates the continuous-valued function. The loss function therefore ignores values within the margins of error defined by ε, whilst penalizing those falling outside. As with the hyperplane for SVMs, the ε-tube is defined by support vectors, which are the training datapoints falling just outside the tube.

2.3.4 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) are a popular form of artificial neural network, which is commonly used for analysing visual imagery. These networks were made popular by Krizhevsky et al. [34] after their success in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), achieving much greater accuracy than previous state-of-the-art solutions. As with all artificial neural networks, CNNs are structured with several layers which uses edge weights, biases and activation functions to alter the input data. An input layer is where the input data is sent in, and an output layer receives the final evaluated output, which has gone through the network of layers. The layers between the input layer and the output layer are called hidden layers. These networks can then be trained by altering the weights used in the layers, in response to the inputs and corresponding ground truth values. This is done by using gradients calculated from the loss of the models, and backpropagate these gradients through the network.

The most distinguishing feature, and what names the CNN, is the use of so called "Convolutional layers". These layers are based upon the mathematical operation convolution. Convolution works like this: two functions are taken in, whereupon a third function is produced, which describes how the shape of one of the input functions is modified by the other. Translating this into

the convolutional layers used by CNNs, these layers produce a feature map constructed from the input tensor. Alongside these convolution layers, CNNs usually rely upon pooling layers, and fully connected layers; the former being layers which reduce dimensionality, and the latter being layers where all neurons in one layer is connected to another.

Residual Neural Network (ResNet)

A big problem within the field of CNNs, are with the construction of ever deeper neural networks. As these networks get even deeper to solve complex tasks and improve accuracy, the training becomes more difficult, and accuracy starts to saturate, and eventually degrade as well. To solve these issues, Residual Neural Network (ResNet) were introduced by He et al. [35] in 2015. ResNet introduced residual blocks, which are blocks of layers, which includes direct connections bypassing some of the layers within the block. These connections are called "skip connections". This solves one of the main causes of the aforementioned problems, which are that as the network gets deeper, the gradients used to update the model layers get vanishingly small as it backpropogates through the model. The skip connections allow the gradient to flow through and not vanish. These connections also ensures that the higher layers perform at least as good as the lower ones, as they help the model learn the identity function.

2.3.5 Multi Image Machine Learning Models

In real life, humans make distinctions and recognition based upon visual information received over time. Small changes in features can to a large degree determine how things appear. As a logical consequence, this applies to computer vision as well. Video cameras has become more and more widespread in its use, which grants us data in the form of sets of images taken over time. The use of these image sets in machine learning tasks, has become a significant area of research in the last few years.

A major review of the field was undertaken by Zhao et al. [36] in 2019. They found that image set classification methods were suitable for the use of analysing image sequences from video surveillance, or multi-angle cameras. These images could then vary drastically, with variations to illumination, posture, viewing angle etc.; providing more detail rich features for more complex classification tasks. Further Zhao et al. [36] found that, though challenging, the use of more recent and advanced forms of machine learning, such as CNNs, could be used for image set classification tasks.

Multi-Input Convolutional Neural Network (CNN)

Though both image set classification, and CNNs are some of the hottest fields of research, there has been few efforts to combine the two. The complexity of both fields, combined with the need for computational power, might be the main causes for this. However, a few efforts has been done to create multi input CNNs. One of the first of these, were described by Sun et al. [23], where a CNN handling three images as an input were created. Used to grade flowers for handling and marketing, it was able to classify sets of three images of a plant, taken from three different angles. As grading of flowers is a difficult task single image CNNs proved unsuitable, as single images could not cover the whole plant. Therefore, they created a CNN which took in three distinct images, ran these through separate convolutional layers, with each followed by a pooling layer, before merging the extracted features. The merged features where then sent through a more conventional CNN. They found this multi-input CNN to outperform single image CNN by 5% on average, achieving up to a 93.9% accuracy in grading the flowers.

Sun et al. [23] showed that multi-input CNNs could have great advantages over conventional single input CNNs. The added distinguishing features extracted from the use of multiple images, could in their case at least, help with achieving better accuracy in classification. As other problems may also benefit from the added distinguishing features a multi-input CNN may provide, this seems to be a great proof of concept, with great promise for other similar problems.

Chapter 3

Method

The main goal of this thesis is to study the viability of multi-input machine learning methods for bilirubin estimation. To achieve this goal, the effectivity of multi-input machine learning models need to be compared to that of existing solutions. To this end, both a regression model and a single input CNN is used as baselines. Both are trained and tested with the same dataset as the multi-input machine learning models. Two approches to multi-input CNN models is proposed: one with less feature extration per image in the imageset taken as an input, and one with heavy single image feature extraction before merging features of the imageset. **Appendix A** provides more detailed graphs of the CNN architectures. Also, to be able to train the computationally heavy CNN models, a GPU server running two NVIDIA Tesla V100 with 32 GB VRAM each are used. As the SVR are implemented using scikit-learn, which has no GPU support, the SVR are being run on an Intel i7 6700k with 3.4Ghz quad core, and 16 GB of 2133 Mhz DDR4 memory.

Trial	1	2	3	4	5	6	7
Subjects	32	24	33	41	40	46	36
Image Sets	32	24	33	41	43	57	36
Images	192	144	198	246	258	342	216

Table 3.1: Distribution of captured images over trials.

3.1 Dataset

The dataset used in this thesis is provided by Picterus As. It consists of 1596 images of newborn babies, structured into 266 image sets belonging to 252 subjects. 14 subjects had 2 sets of images taken. Each set contains 6 images, where half were taken with flash, and half without. The images were gather at 7 distinct trials, with the distribution as given in table 3.1. As table 3.1 shows, the extra image sets stems from trial 5 and 6. Every image is already color calibrated, represented by normalized, linear RGB values.

Target Bilirubin Distribution

For each subject in the dataset, a standarized ground truth bilirubin measurement is recorded. Figure 3.1 shows the distribution of bilirubin values, which quite closely follows a bell curve, which is as expected for standarized values. It also shows that the most severe outliers lies on the positive side of 0, which is also expected, as these are the cases with EHB.

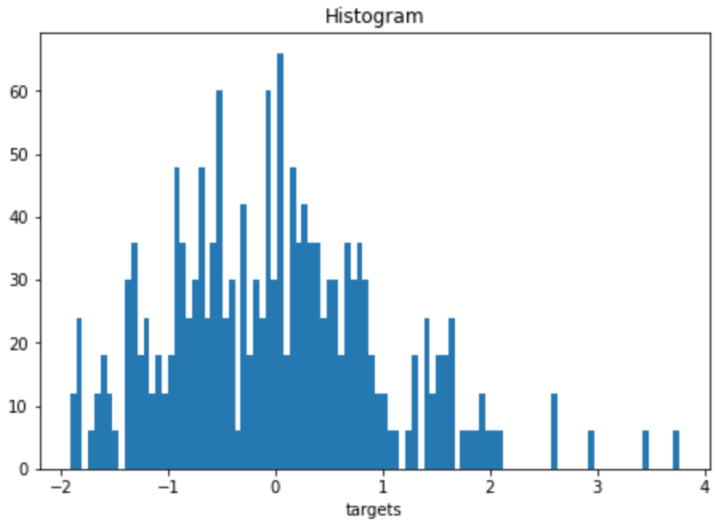
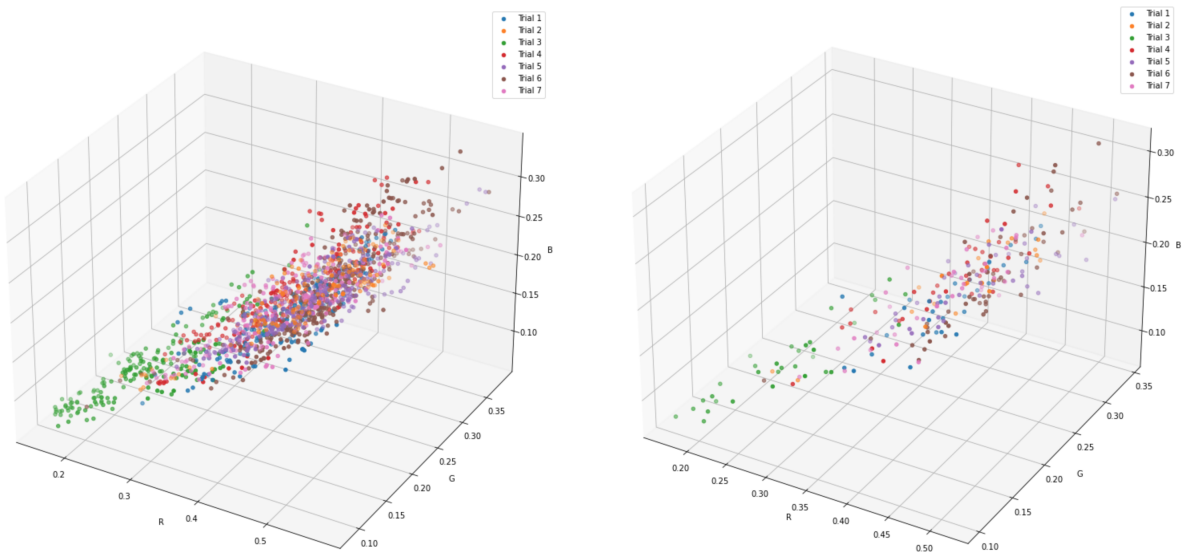


Figure 3.1: Distribution of bilirubin values.

Skin Color Distribution

There is no data recorded as to the distribution of ethnicities: whether the dataset mainly consists of light toned Caucasian subjects, or if it includes darker and more deeply toned ones as well. Using the mean RGB value of each image, we can get an idea of at least how skin tones are differently distributed across the trials, these are shown in figure 3.2. No definitive conclusions may be drawn up from these graphs, as the mean RGB value is not an accurate measurement of skin tone. However, these graphs suggest that certain trials skew either more towards the darker shades as trial 3 clearly does, and some towards the lighter shades as it seems trial 6 does. These skews might also correlate to the different bilirubin distributions across the trials as well.



(a) Single image mean RGB value

(b) Image set mean RGB value

Figure 3.2: Skin color distribution, color coded by trial and plotted in RGB space.

3.2 Baseline Models

C	1	10	100	1000	2000	5000	-
degree	2	3	4	-	-	-	-
ε	0.005	0.01	0.05	0.1	0.2	0.3	-
γ	0.05	0.1	0.2	0.5	0.7	0.8	0.9
Kernel	Linear	Polynomial	RBF	Sigmoid	-		-

Table 3.2: Hyperparameters used for grid search on SVR model

3.2.1 Support Vector Regression (SVR) Model

As previously discussed, most already existing methods for estimating bilirubin use a form of regression to predict bilirubin values based upon features extracted from digital images. One form of regression that was used by several of these papers [9, 11, 16, 19], is Support Vector Regression (SVR). Being a robust regression model, with fast runtimes, and providing generally solid results, there is no wonder why it has been a popular option for this task on previous occasions. The implementation used in this thesis is the scikit-learn implementation based on the work of Chang and Lin [37]. Up to date with the most stat-of-the-art kernels, and with a simple framwork written in Python, this a solid and flexible implementation. The setup follows the same pattern as Falk and Jensen [16], with grid search, KFold cross-validation and polynomial RGB expansion of mean RGB values. 10 times repeating KFold cross-validation with 5 folds is applied, alongside a grid search using the hyperparameters from table 3.2. KFold cross-validation means that the dataset is divided into k randomly shuffled groups, in this case 5 groups, where for each group, the others are used for training, and the hold out for testing. This means that in this case 50 repeats of training and testing is being used to validate each combination of hyperparameters. The main difference made from that of Falk and Jensen [16], is that the degree of the polynomial expansion is not subject to the hyperparameter search, and instead the degree of $\theta_{1,3}$ is chosen.

3.2.2 Single Image CNN (SICNN)

As the multi image machine learning models used in this thesis is multi-input CNNs, a second baseline in the form of a single image CNN is used. Using PyTorch, the Wide ResNet50-2 provided by Torchvision is implemented, and trained with the previously discussed dataset, using KFold cross-validation with 5 folds. Optuna [38] hyperparameter search, a random, automated hyperparameter search framework, is used to run several trials to optimize batch size and learning rate. Early stopping is also implemented to avoid overfitting, with manual tuning of patience. The fully connected layer is modified to be optimized for the regression task.

The Torchvision implementation of ResNet, is based upon the work of He et al. [35], which won the ImageNet Large Scale Visual Recognition Challenge 2015 (ILSVRC2015). Torchvision provides several ResNet models with the amount of layers the names suggest: ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152. These layers are sorted into Bottlenecks, which in turn are sorted into Layer blocks. A graph illustrating this is provided in **Appendix A** as figure 6.2. Wide ResNet expands upon the normal ResNet, by providing double the Bottleneck channels within each layer block. E.g. where the last convolution in ResNet-50 has 2048-512-2048 channels, the Wide ResNet-50-2 has 2048-1024-2048 channels.

3.3 Multi Image Machine Learning Models

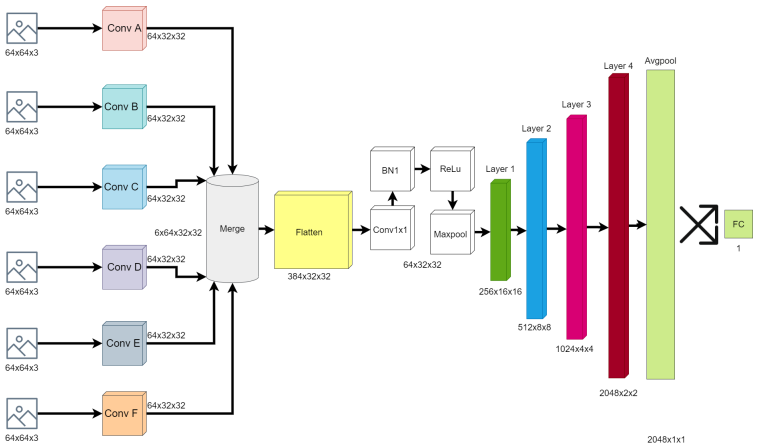


Figure 3.3: Diagram showing the SMICNN. Multiple image inputs is fed through separate convolution layers, before merging and completing the Wide ResNet50-2.

3.3.1 Simple Multi Image CNN (SMICNN)

The first effort to achieve a multi image machine learning model, was by taking inspiration from Sun et al. [23]. Using the same methodology, we modify the Wide ResNet-50 model by duplicating the first convolution layer, and feeding each of the six images in the image set through separate convolution layers. After this single convolution per image, the output is merged, and fed through a 1x1 convolution layer, to reduce the channels. The resulting output is then fed through the rest of the network, using a modified fully connected layer as with the SICNN. As with the SICNN, we use KFold cross-validation with 5 folds, Optuna [38] hyperparameter search and early stopping with manual tuning of patience. A simple representation is seen in figure 3.3. In the **Appendix A** a more detailed representation of the SMICNN is given as figure 6.3.

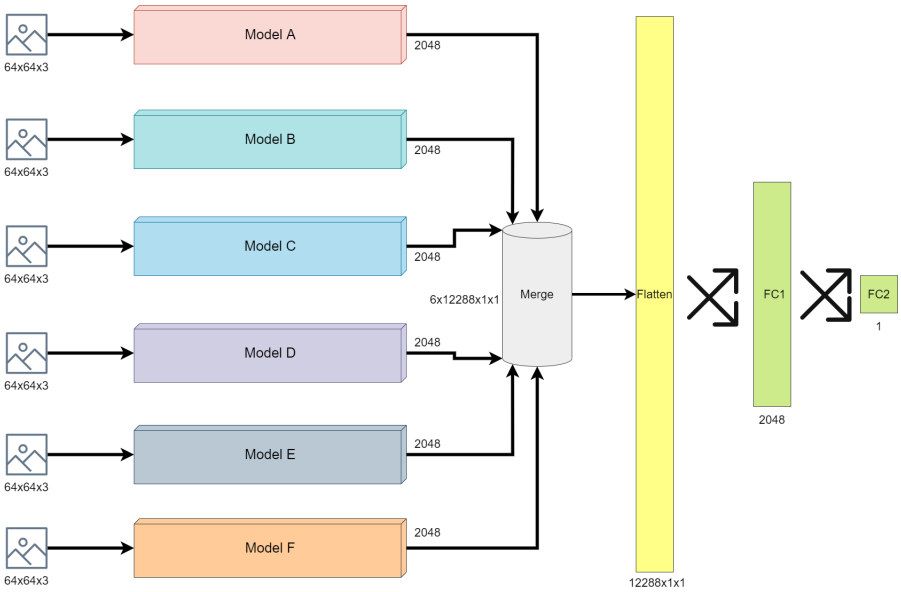


Figure 3.4: Diagram showing the CMICNN. Multiple image inputs is fed through Wide ResNet50-2 models, before merging and completing a couple fully connecting layers.

3.3.2 Complex Multi Image CNN (CMICNN)

As a second multi image machine learning model, CMICNN is presented. To try to get even more discriminate features from the single images in the image sets, the CMICNN model runs each image through a full Wide ResNet50-2. After this, the output features are merged, and run through two fully connected layers. Two layers are needed because the output of the merging results in $6 * 2048 = 12288$ channels, which is far too much for a single fully connected layer to handle for regression tasks. Again we train and evaluate with KFold cross-validation, use Optuna [38] for hyperparameter search, and early stopping with manually tuned patience. Figure 3.4 shows the CMICNN network. An even more detailed representation is listed in **Appendix A** as figure 6.4.

Chapter 4

Results

The results from the experiments detailed in the Method chapter are presented here. First the performance of the models are detailed, and secondly the steps to try to optimize the models are detailed.

4.1 Performance Comparison

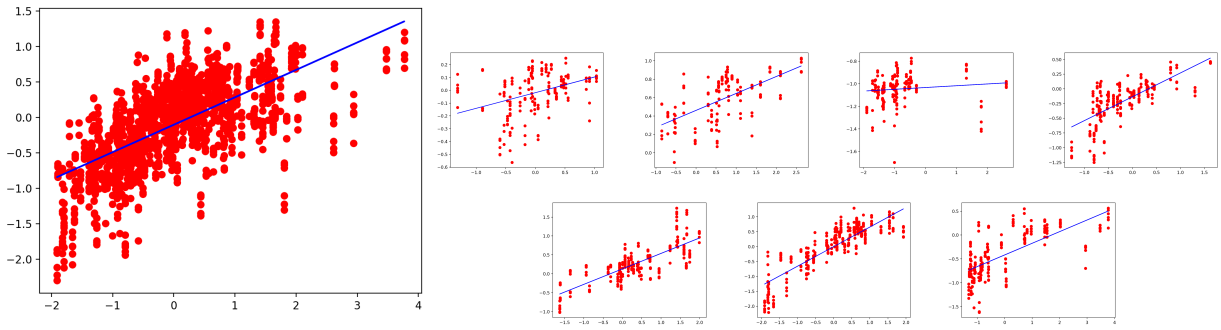
Four models are laid forward to compare their ability to predict bilirubin values: a Support Vector Regression (SVR) model, a single image CNN (SICNN), and two multi image CNNs (SMICNN and CMICNN). All models use KFold cross-validation with 5 folds, where each fold trains the models on 80% of the data, retaining the hold out fold of 20% of the data for testing. Each test generates predicted values using the training set, and compares these against the ground truth values recorded. The final scores are calculated as a mean of the evaluation scores from each fold. Several measurements are taken, the most prevalent of which are recorded in table 4.1.

Model	Max err.	MSE	Corr.
SVR	3.308	0.5898	0.6513
SICNN	3.064	0.4943	0.7139
SMICNN	2.9062	0.6877	0.5746
CMICNN	2.486	0.6253	0.6236

Table 4.1: Performance comparison between the machine learning models.

Trial	Max err.	MSE	Corr.
1	1.441	0.2274	0.3659
2	1.746	0.5387	0.6277
3	3.643	0.9718	0.1198
4	1.425	0.1934	0.7046
5	1.502	0.3608	0.7608
6	1.643	0.2594	0.8600
7	3.6393	1.3985	0.6905

Table 4.2: Caption



(a) Full dataset

(b) Per trial

Figure 4.1: Linear regression of the correlation between predicted values by the SVR model and actual values.

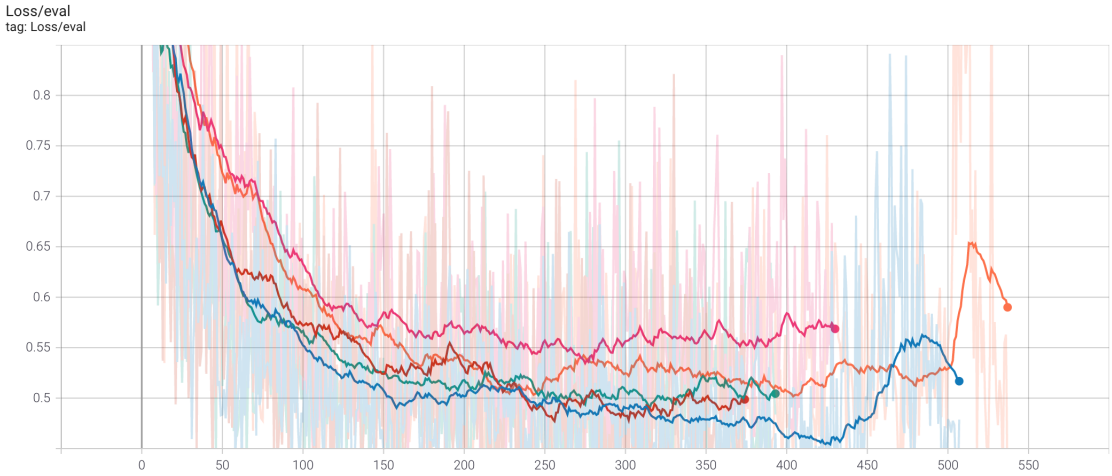
Support Vector Regression (SVR) Model

After doing the grid search detailed in section 3.2.1, the optimal hyperparameters were found for the SVR model. Using these hyperparameters, the values in table 4.1 were recorded. Doing several runs, resulted in a bit varying values, as the different splits of the dataset affected the performance to some degree. Still, the max error stayed around the same value of around $3.1 - 3.5$, the MSE around $0.55 - 0.59$, and the correlation around $0.65 - 0.71$. The graphs shown in figure 4.1 shows a plotted linear regression between the estimated values produced by the SVR model, and the ground truth values. The leftmost graph shows the SVR trained and tested on the full dataset, whereas the seven graphs on the right were produced by training and testing the SVR with only data from each trial. Table 4.2 records the measurements taken for each trial.

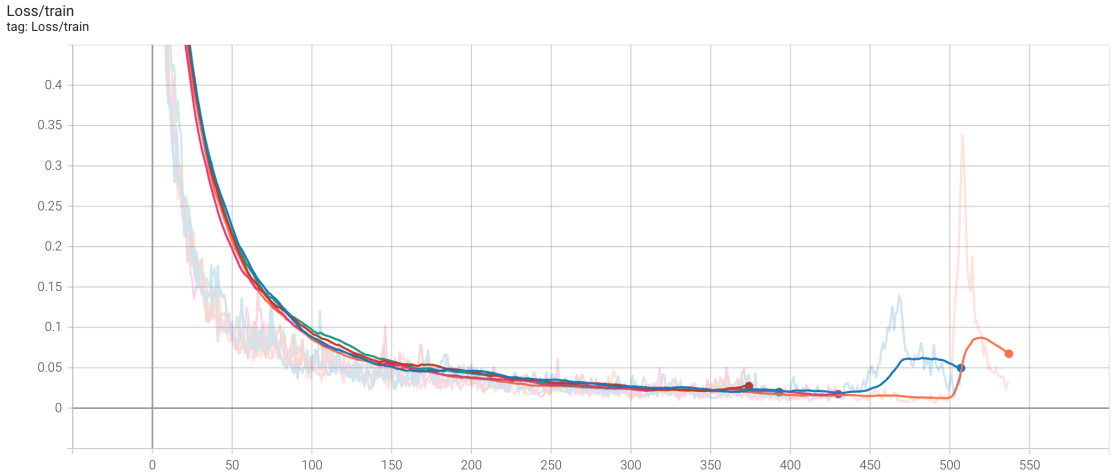
Convolutional Neural Network (CNN) Models

The best values achieved by the CNN models during testing, are also recorded in table 4.1. The SICNN managed to get pretty good both MSE and correlation values, outperforming the SVR. But neither of the multi image CNNs were able to do this, scoring worse in both MSE and correlation. Notably however, both models scored lower max errors than both the SVR and the single image CNN.

Figures 4.2, 4.3 and 4.4 shows the loss measured in MSE for both training and evaluation over time for each of the models. The SICNN from figure 4.2 performs quite well, with the exception of the two folds which spike somewhat towards the end. This, however, may only indicate a saddlepoint for the model at around this number of epochs, as the folds without spikes stops short of this number of epochs. The SMICNN evaluation loss is quite interesting, as there is a clearly two folds performing much better than the rest. This is somewhat puzzling, as this was unique to the SMICNN when running all three CNN models with the same randomstate for fold division. Other than this, figure 4.3 seems to show that the performance of the best folds shows great promise if the model could generalize and achieve these scores with the other folds as well. The CMICNN is quite volatile, and as figure 4.4 shows, the loss is fluctuating heavily.

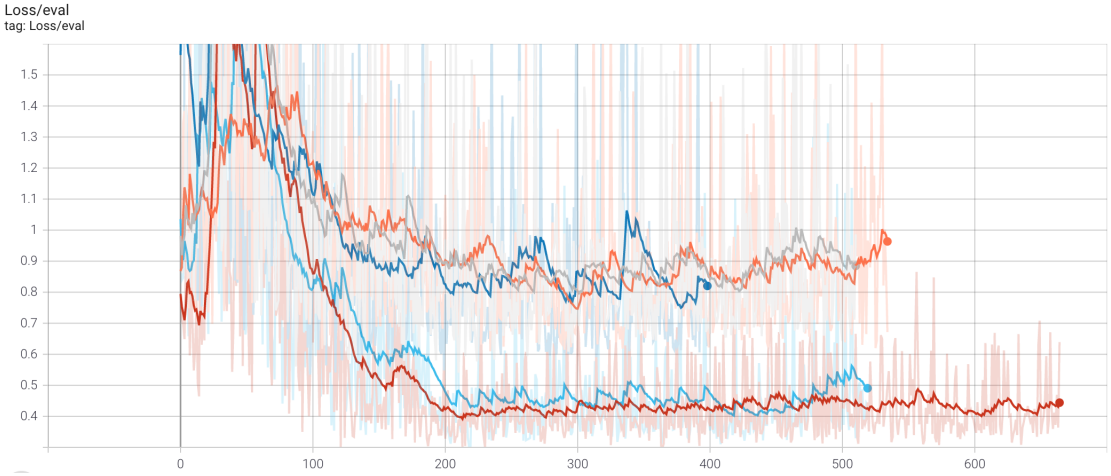


(a) Evaluation loss

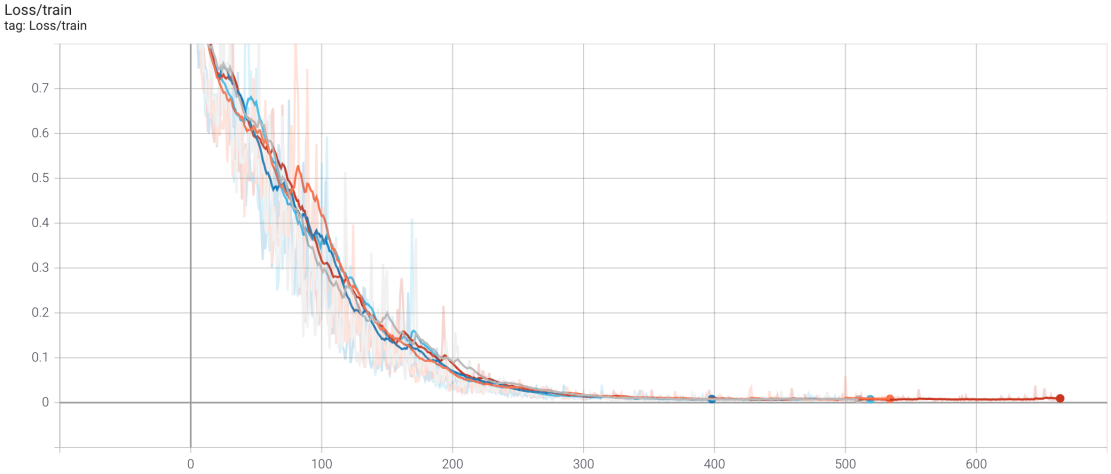


(b) Training loss

Figure 4.2: Loss charts for the SICNN, where loss is evaluated in MSE.

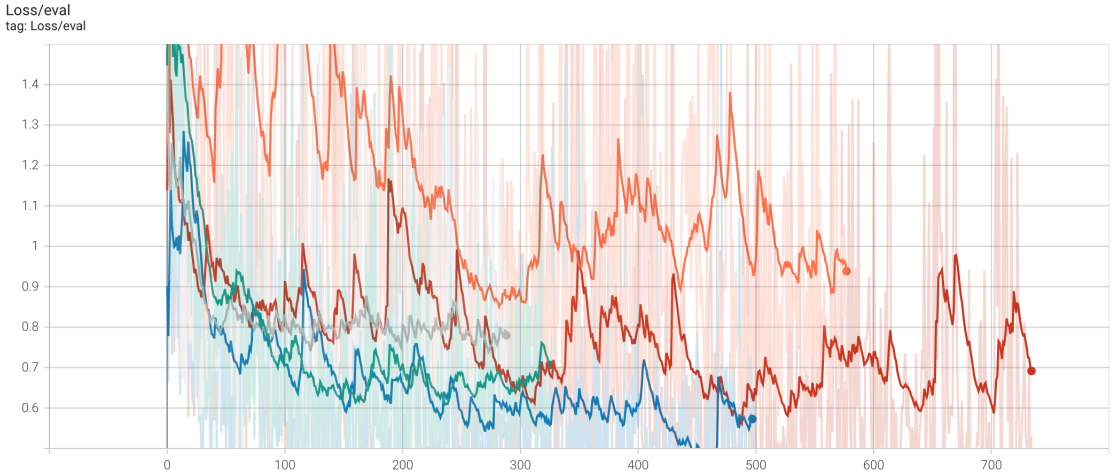


(a) Evaluation loss

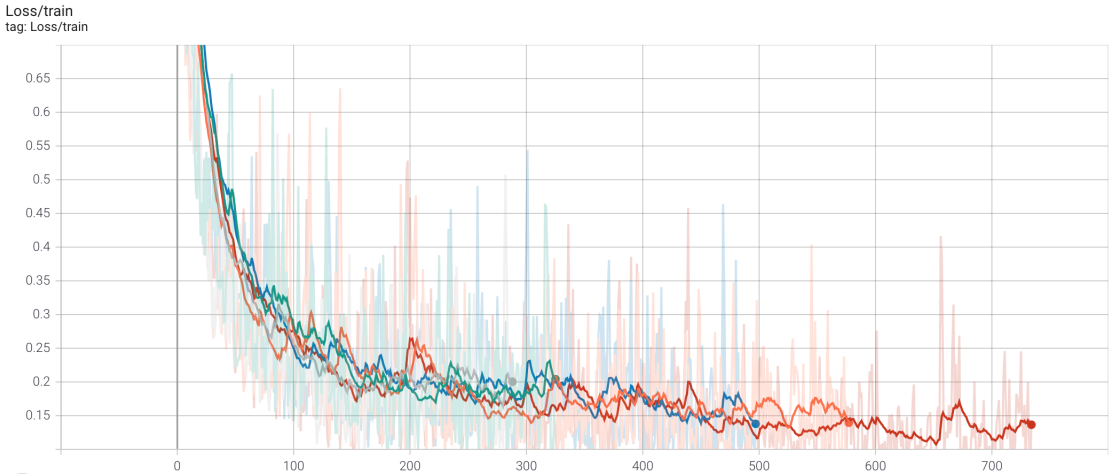


(b) Training loss

Figure 4.3: Loss charts for the SMICNN, where loss is evaluated in MSE.



(a) Evaluation loss



(b) Training loss

Figure 4.4: Loss charts for the CMICNN, where loss is evaluated in MSE.

4.2 Hyperparameter Optimization

An important step in getting results from the models, were optimization of the hyperparameters. A full extensive grid search were conducted for the SVR model, whilst several optimization attempts were conducted on the Convolutional Neural Network (CNN) models.

C	degree	ε	γ	Kernel
5000	4	0.3	0.9	Polynomial

Table 4.3: Optimal hyperparameters for the SVR model, found with a grid search.

Support Vector Regression (SVR) Model

A grid search was used to find the best hyperparameters for the SVR model. The hyperparameter values included in this grid search, is detailed in table 3.2. The best found hyperparameters are detailed in table 4.3. As the kernel found to be best suited to this task is the Polynomial kernel, the ε and γ values are irrelevant, as the Polynomial kernel only uses the degree and C value.

Model	Learning rate	Batch Size
SICNN	1.330e-4	150
SMICNN	5.894e-4	25
CMICNN	5.293e-6	25

Table 4.4: Best found hyperparameters for the CNN models using Optuna [38] random hyperparameter search.

Convolutional Neural Network (CNN) Models

Several hyperparameter search trials were conducted to try to optimize the CNNs. As detailed in the Method chapter, these were conducted using Optuna [38] hyperparametersearch, which is a random, automated hyperparameter search framework. Both the SICNN and the SMICNN were also manually tuned to some degree, as to try to narrow down the best hyperparameters a bit further. As the CNN models required much more computing time than the SVR, a complete optimization of the models were not completed. This was especially the case for the CMICNN, which as the most complex model, required the longest running times of about 2-3 hours per trial. In contrast the SMICNN ran in a much more timely manner of around 20 mins per trial. Optimization trials were run until the end of the project, and the recorded values in tables 4.1 and 4.4 are the best values found during this search. As the SMICNN were the fastest running model, some attempts were made to use a learning rate scheduler, but time made this difficult, and the use of such schedulers were not explored to its fullest.

Chapter 5

Discussion

In this chapter, the results presented in the previous chapter, Results, are discussed. The performance of the models, how they compare, and the hyperparameter optimization will be discussed. Also, this chapter will reflect upon the presented solutions.

5.1 Performance Comparison

To compare the models, we use with 5 folds for all the models. We also set the randomstate for the shuffeling of the folds, so that they should be shuffled the same for all models. There is one problem to this, however, and that is that the multi image models separate the dataset into imagesets before shuffeling, while the SVR and SICNN shuffles using all the images. Therefore, the folds are not equal between the models, and may be a source for discrepancy between the models. The set randomstate at least ensured comparability between the different runs of the CNNs, which helped towards optimizing the models.

Support Vector Regression (SVR) Model

The SVR model was run using both the full dataset, and using only the data for each trial which the data were collected from. Figure 4.1 b) shows the plotted linear regression between the estimated values and the ground truth values for each trial. Few conclusions may be drawn, however, as each of these trials had much fewer images to be trained from, and to predict upon. This only serves as an indication that the data from the different trials skews the final result in either a positive or negative direction. I.e the data from trial three seems to have much less correlation than the other trials, which will will impact the performance of the SVR on the full dataset as well. Of the data in table 4.2, only the correlation is of real importance, as the range of bilirubin values are different

between the different trials. This affects both the max error and the MSE, which is clearly seen as the correlation of trial seven is quite good, however the MSE is much worse than the other trials, as the range of bilirubin values in this trial is much greater than for the others.

Convolutional Neural Network (CNN) Models

The performance of the SICNN shows that CNNs are able to distinguish more distinct features than those extracted for the SVR, and thereby outperform such a more conventional model. But as figure 4.2 shows, there is still room for improvement with this model, as it is overfitting a bit, as the difference between the training and evaluation loss shows.

Neither of the multi image CNNs could achieve as good performance as either the SVR or the SMICNN. This is however, not necessary an indication that these models are worse, as before achieving the results listed in table 4.1, the SICNN were performing quite similarily as the SMICNN. A lucky trial hitting much better hyperparameters than those found for the SMICNN might be the cause of this. An other factor may be the difference in which images falls within the different folds, as figure 4.3 shows, a couple of the folds of the SMICNN outperformed the SICNN. A different shuffeling of the imagesets may have lead to there being more consitency between the folds, and the SMICNN might have performed much better.

The CMICNN on the other hand, is difficult to evaluate. As figure 4.4 shows, the model is very irratic, and much more volatile than the other two models. This is an indication that the model complexity does not allow it to generalize well, at least with the used datset. It might be that the dataset used was much too small to be able to train this quite complex model. Also, because of its complexity, it required the longest run times, which also made it much more difficult to optimze, as far fewer trials were conducted with it.

The choice of creating the models as ResNets might also not have been the best choice. Though optimized for very deep networks which can distinguish a large number of distinct features, this task may not have required such deep networks and so much feature extraction. The skip connections may also have been problematic, as all models show indication of very quickly learning the training data. The quick training provided my these skip connections may have been detrimental to allowing the models to generalize well.

5.2 Hyperparameter Optimization

As the SVR does not require too much computational time, we chose to do an exhaustive grid search to find optimal hyperparameters. Doing a grid search for the CNNs would not be feasible, as the number of runs necessary would be too much to do in a timely manner. Therefore, using the automated, random search framework Optuna [38], we were able to do trials using random hyperparameter values, within set ranges, to try to find the most optimal parameters.

Support Vector Regression (SVR) Model

A full hyperparameter optimization was conducted on the SVR model, using grid search. This optimization, however, was only conducted using the full dataset, and was not conducted for each trial. Therefore the results from the trial by trial run uses the same hyperparameters as each other, and as the run with the full dataset.

Convolutional Neural Network (CNN) Models

The optimization of the CNN models were time consuming, as each trial of hyperparameters took a long time. The hyperparameters to be optimized for these models were the learning rate, and the batch size. The patience for the early stopping is also a parameter that could be tuned, however this was only tuned manually with a bit of trial and error. All models seem to stop at reasonable times, at points where the loss isn't really improving much more, and before it starts to degrade. None of the CNNs were optimized fully, and there is room for improving the hyperparameter optimization for all. Also adding learning rate schedulers to the models might help performance as well. The SMICNN was experimented with using a couple learning rate schedulers, both reduce on plateau scheduler and a one cycle policy scheduler, which ran over 400 epochs. They were experimented with at different stages of the optimization, and no clear experiments were done to compare learning rate schedulers. However, both did marginally improve performance somewhat, indicating that adding such schedulers might help the models performances.

5.3 Project Reflection

Going into the project, I had only the cursory knowledge of machine learning attained through the previous semesters course TDAT3025 Applied Machine Learning. I quickly realized that exploring multi-input machine learning methods were a vast field, which would require much study. Starting to imagine solutions were difficult, but after coming across the papers by Sun et al. [23] Kawulok et al. [24] and Falk and Jensen [16], I started gravitating towards solutions involving CNNs. As I had previous knowledge and experience with CNNs, solutions with this method quickly became the main focus of the report. The advantage of this, is that CNNs are powerful tools for image processing, which had not really been taken advantage of in previous research into bilirubin estimation, other than by Falk and Jensen [16] who struggled with them. Their struggle served as a motivator to accomplish what had not been accomplished previously, though it also foreshadowed my own problems with these solutions. Locking myself into focusing on CNNs, meant having to spend the time on these computationally heavy models, which required much time to try to optimize and get results from. It also meant that I only really ended up exploring a single small area of a much vaster field of multi-input machine learning models.

The project has been a valuable learning experience which have provided me with much knowledge regarding machine learning. It has been a somewhat daunting challenge to work on this project by myself, especially as I have a hard time forcing myself to work. However, this project has been much fun, and an overall great experience, as this is technology which really interests me, and work done may contribute to a great cause.

Chapter 6

Conclusion

As neither of the proposed multi image CNNs were completely optimized, no definitive conclusion with regards to the research questing given in the Goals of this research section may be drawn. If using only the best found values from table 4.1, there would seem like the answer would be "no". However, as both models have room for improvement, they might be able to outperform both the SVR and SICNN. The conclusion will be as follows: multi image machine learning models, and more specifically, the models proposed by this theses, have great potential to become competitive models for bilirubin prediction.

6.1 Further work

As discussed, there was not enough time to fully optimize and improve the performance of the multi image CNNs. Therefore, a natural course for further work, is investing time and effort into optimizing these models. As also discussed, the dataset might have been too small to fully train the multi image CNNs. By using a larger dataset, the multi image CNNs might be able to perform better, as they would have more data to learn from.

The multi image CNNs proposed in this thesis are only two different experiments using multi-input CNN. Other architectures might produce better results than the ones explored in this thesis. The structuring of when to merge, how much processing should be applied before and after the merge, are areas which might be explored. Also, all the CNNs in this thesis is based off of ResNets. Many other CNN architectures are detailed and implemented into Torchvision, which might perform better or worse than ResNet.

This thesis ended up really only focusing on multi-input CNNs, and did not experiment with any other forms of multi-input/image processing machine learning methods. Zhao et al. [36] details a lot more approaches to machine learning models using image sets, which also could be explored. This thesis is only scratching at the surface of the field of multi image machine learning methods, showing that there might be some benefits to these kinds of models for bilirubin estimation. Further work should therefore look into other completely different approaches than those detailed in this thesis.

References

[1] R. C. Amos, H. Jacob and W. Leith, 'Jaundice in newborn babies under 28 days: Nice guideline 2016 (cg98),' *Archives of Disease in Childhood - Education and Practice*, vol. 102, no. 4, pp. 207–209, 2017, ISSN: 1743-0585. DOI: 10.1136/archdischild-2016-311556. eprint: https://ep.bmj.com/content/102/4/207.full.pdf. [Online]. Available: https://ep.bmj.com/content/102/4/207.

[2] T. W. R. Hansen and D. Bratild, 'Bilirubin and brain toxicity,' *Acta Paediatrica*, vol. 75, no. 4, pp. 513–522, 1986. DOI: https://doi.org/10.1111/j.1651-2227.1986.tb10242.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1651-2227.1986.tb10242.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1651-2227.1986.tb10242.x.

[3] V. Bhutani, A. Zipursky, H. Blencowe, R. Khanna, M. Sgro, F. Ebbesen, J. Bell, R. Mori, T. Slusher, N. Fahmy, V. Paul, L. Du, A. Okolo, M. F. de Almeida, B. Olusanya, P. Kumar, S. Cousens and J. Lawn, 'Neonatal hyperbilirubinemia and rhesus disease of the newborn: Incidence and impairment estimates for 2010 at regional and global levels,' *Pediatric research*, vol. 74 Suppl 1, pp. 86–100, Dec. 2013. DOI: 10.1038/pr.2013.208.

[4] R. Keren, K. Tremont, X. Luan and A. Cnaan, 'Visual assessment of jaundice in term and late preterm infants,' *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 94, no. 5, F317–F322, 2009, ISSN: 1359-2998. DOI: 10.1136/adc.2008.150714. eprint: https://fn.bmj.com/content/94/5/F317.full.pdf. [Online]. Available: https://fn.bmj.com/content/94/5/F317.

[5] A. Riskin, A. Tamir, A. Kugelman, M. Hemo and D. Bader, 'Is visual assessment of jaundice reliable as a screening tool to detect significant neonatal hyperbilirubinemia?' *The Journal of pediatrics*, vol. 152, 782–7, 787.e1, Jun. 2008. DOI: 10.1016/j.jpeds.2007.11.003.

[6] T. M. Slusher, A. Zipursky and V. K. Bhutani, 'A global need for affordable neonatal jaundice technologies,' *Seminars in Perinatology*, vol. 35, no. 3, pp. 185–191, 2011, Newborn Jaundice Technologies, ISSN: 0146-0005. DOI: https://doi.org/10.1053/j.semperi.

2011.02.014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0146000511000437.

[7] S. F. V. Sondaal, J. L. Browne, M. Amoakoh-Coleman, A. Borgstein, A. S. Miltenburg, M. Verwijs and K. Klipstein-Grobusch, 'Assessing the effect of mhealth interventions in improving maternal and neonatal care in low- and middle-income countries: A systematic review,' *PLOS ONE*, vol. 11, no. 5, pp. 1–26, May 2016. DOI: 10.1371/journal.pone.0154664. [Online]. Available: https://doi.org/10.1371/journal.pone.0154664.

[8] L. de Greef, M. Goel, M. J. Seo, E. C. Larson, J. W. Stout, J. A. Taylor and S. N. Patel, 'Bilicam: Using mobile phones to monitor newborn jaundice,' in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '14, Seattle, Washington: ACM, 2014, pp. 331–342, ISBN: 978-1-4503-2968-2. DOI: 10.1145/2632048.2632076. [Online]. Available: http://doi.acm.org/10.1145/2632048.2632076.

[9] J. Castro-Ramos, C. Toxqui-Quitl, F. V. Manriquez, E. Orozco-Guillen, A. Padilla-Vivanco and J. Sánchez-Escobar, 'Detecting jaundice by using digital image processing,' in *Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing XXI*, T. G. Brown, C. J. Cogswell and T. Wilson, Eds., International Society for Optics and Photonics, vol. 8949, SPIE, 2014, pp. 290–296. DOI: 10.1117/12.2041354. [Online]. Available: https://doi.org/10.1117/12.2041354.

[10] Z. Rong, F. Luo, L. Ma, L. Chen, L. Wu, W. Liu, L. Du and X. Luo, 'Evaluation of an automatic image-based screening technique for neonatal hyperbilirubinemia,' *Zhonghua Er Ke Za Zhi (Chinese)*, vol. 54, no. 8, pp. 597–600, Aug. 2016. DOI: doi:10.3760/cma.j.issn.0578-1310.2016.08.008..

[11] M. Aydin, F. Hardalaç, B. Ural and S. Karap, 'Neonatal jaundice detection system,' *Journal of Medical Systems*, vol. 40, May 2016. DOI: 10.1007/s10916-016-0523-4.

[12] J. A. Taylor, J. W. Stout, L. de Greef, M. Goel, S. Patel, E. K. Chung, A. Koduri, S. McMahon, J. Dickerson, E. A. Simpson and E. C. Larson, 'Use of a smartphone app to assess neonatal jaundice,' *Pediatrics*, vol. 140, no. 3, 2017, ISSN: 0031-4005. DOI: 10.1542/peds.2017-0312. eprint: https://pediatrics.aappublications.org/content/140/3/e20170312.full.pdf. [Online]. Available: https://pediatrics.aappublications.org/content/140/3/e20170312.

[13] B. Yang, D. Huang, X.-y. Gao, M. Su, M. Li, H.-l. Lei, Y. Ren and C. Shi, 'Neonatal and early infantile jaundice: Assessment by the use of the smartphone,' 2018.

[14] S. Swarna, S. Pasupathy, B. Chinnasami, N. D. and B. Ramraj, 'The smart phone study: Assessing the reliability and accuracy of neonatal jaundice measurement using smart phone application,' *International Journal of Contemporary Pediatrics*, vol. 5, no. 2, pp. 285–289,

2018, ISSN: 2349-3291. DOI: 10.18203/2349-3291.ijcp20175928. [Online]. Available: https://www.ijpediatrics.com/index.php/ijcp/article/view/1290.

[15] S. B. Munkholm, T. Krøgholt, F. Ebbesen, P. B. Szecsi and S. R. Kristensen, 'The smartphone camera as a potential method for transcutaneous bilirubin measurement,' *PLOS ONE*, vol. 13, no. 6, pp. 1–11, Jun. 2018. DOI: 10.1371/journal.pone.0197938. [Online]. Available: https://doi.org/10.1371/journal.pone.0197938.

[16] H. H. Falk and O. D. Jensen, *A machine learning approach for jaundice detection using color corrected smartphone images*, eng, 2018. [Online]. Available: http://hdl.handle.net/11250/2566507.

[17] T. S. Leung, F. Outlaw, L. W. MacDonald and J. Meek, 'Jaundice eye color index (jeci): Quantifying the yellowness of the sclera in jaundiced neonates with digital photography,' *Biomed. Opt. Express*, vol. 10, no. 3, pp. 1250–1256, Mar. 2019. DOI: 10.1364/BOE.10.001250. [Online]. Available: http://www.osapublishing.org/boe/abstract.cfm?URI=boe-10-3-1250.

[18] P. Padidar, M. Shaker, H. Amoozgar, M. Khorraminejad, F. Hemmati, K. Najib and S. Poorarian, 'Detection of neonatal jaundice by using an android os-based smartphone application,' *Iranian Journal of Pediatrics*, vol. In Press, Feb. 2019. DOI: 10.5812/ijp.84397.

[19] Y. Karamavuş and M. Özkan, 'Newborn jaundice determination by reflectance spectroscopy using multiple polynomial regression, neural network, and support vector regression,' *Biomedical Signal Processing and Control*, vol. 51, pp. 253–263, 2019, ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2019.01.019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809419300199.

[20] A. Aune, G. Vartdal, H. Bergseng, L. L. Randeberg and E. Darj, 'Bilirubin estimates from smartphone images of newborn infants' skin correlated highly to serum bilirubin levels,' *Acta Paediatrica*, vol. 109, no. 12, pp. 2532–2538, 2020. DOI: https://doi.org/10.1111/apa.15287. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/apa.15287. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/apa.15287.

[21] F. Outlaw, M. Nixon, O. Odeyemi, L. W. MacDonald, J. Meek and T. S. Leung, 'Smartphone screening for neonatal jaundice via ambient-subtracted sclera chromaticity,' *PLOS ONE*, vol. 15, no. 3, pp. 1–17, Mar. 2020. DOI: 10.1371/journal.pone.0216970. [Online]. Available: https://doi.org/10.1371/journal.pone.0216970.

[22] A. Chakraborty, S. Goud, V. Shetty and B. Bhattacharyya, 'Neonatal jaundice detection system using cnn algorithm and image processing,' *International Journal of Electrical Engineering and Technology*, vol. 11, no. 3, pp. 248–264, Jun. 2020. [Online]. Available: https://ssrn.com/abstract=3636169.

[23] Y. Sun, L. Zhu, G. Wang and F. Zhao, 'Multi-input convolutional neural network for flower grading,' *Journal of Electrical and Computer Engineering*, vol. 2017, pp. 1–8, Aug. 2017. DOI: 10.1155/2017/9240407.

[24] M. Kawulok, P. Benecki, S. Piechaczek, K. Hrynczenko, D. Kostrzewa and J. Nalepa, 'Deep learning for multiple-image super-resolution,' *CoRR*, vol. abs/1903.00440, 2019. arXiv: 1903.00440. [Online]. Available: http://arxiv.org/abs/1903.00440.

[25] A. F. McDonagh and D. A. Lightner, 'Like a shrivelled blood orange: Bilirubin, jaundice, and phototherapy,' *Pediatrics*, vol. 75, no. 3, pp. 443–455, 1985, ISSN: 0031-4005. eprint: https://pediatrics.aappublications.org/content/75/3/443.full.pdf. [Online]. Available: https://pediatrics.aappublications.org/content/75/3/443.

[26] B. O. Olusanya, T. A. Ogunlesi and T. M. Slusher, 'Why is kernicterus still a major cause of death and disability in low-income and middle-income countries?' *Archives of Disease in Childhood*, vol. 99, no. 12, pp. 1117–1121, 2014, ISSN: 0003-9888. DOI: 10.1136/archdischild-2013-305506. eprint: https://adc.bmj.com/content/99/12/1117.full.pdf. [Online]. Available: https://adc.bmj.com/content/99/12/1117.

[27] A. Carceller-Blanchard, J. Cousineau and E. Delvin, 'Point of care testing: Transcutaneous bilirubinometry in neonates,' *Clinical Biochemistry*, vol. 42, no. 3, pp. 143–149, 2009, ISSN: 0009-9120. DOI: https://doi.org/10.1016/j.clinbiochem.2008.09.106. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0009912008005146.

[28] S. N. El-Beshbishi, K. E. Shattuck, A. A. Mohammad and J. R. Petersen, 'Hyperbilirubinemia and Transcutaneous Bilirubinometry,' *Clinical Chemistry*, vol. 55, no. 7, pp. 1280–1287, Jul. 2009, ISSN: 0009-9147. DOI: 10.1373/clinchem.2008.121889. eprint: https://academic.oup.com/clinchem/article-pdf/55/7/1280/32675577/jlm.2010-2.pdf. [Online]. Available: https://doi.org/10.1373/clinchem.2008.121889.

[29] F. Raimondi, S. Lama, F. Landolfo, M. Sellitto, A. Borrelli, R. Maffucci, P. Milite and L. Capasso, 'Measuring transcutaneous bilirubin: A comparative analysis of three devices on multiracial population,' *BMC pediatrics*, vol. 12, p. 70, Jun. 2012. DOI: 10.1186/1471-2431-12-70.

[30] W. G. O. for eHealth, *Mhealth: New horizons for health through mobile technologies: Second global survey on ehealth*, 2011.

[31] L. Lyngsnes Randeberg, E. Bruzell Roll, L. T. Norvang Nilsen, T. Christensen and L. O. Svaasand, 'In vivo spectroscopy of jaundiced newborn skin reveals more than a bilirubin index,' *Acta Paediatrica*, vol. 94, no. 1, pp. 65–71, 2005. DOI: https://doi.org/10.1111/j.1651-2227.2005.tb01790.x. eprint: https://onlinelibrary.wiley.

com/doi/pdf/10.1111/j.1651-2227.2005.tb01790.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1651-2227.2005.tb01790.x.

[32] C. Cortes and V. Vapnik, 'Support-vector networks,' *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, ISSN: 0885-6125. DOI: 10.1023/A:1022627411411. [Online]. Available: https://doi.org/10.1023/A:1022627411411.

[33] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola and V. Vapnik, 'Support vector regression machines,' in *Proceedings of the 9th International Conference on Neural Information Processing Systems*, ser. NIPS'96, Denver, Colorado: MIT Press, 1996, pp. 155–161.

[34] A. Krizhevsky, I. Sutskever and G. E. Hinton, 'Imagenet classification with deep convolutional neural networks,' in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12, Lake Tahoe, Nevada: Curran Associates Inc., 2012, pp. 1097–1105.

[35] K. He, X. Zhang, S. Ren and J. Sun, 'Deep residual learning for image recognition,' Jun. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[36] Z.-Q. Zhao, S.-T. Xu, D. Liu, W.-D. Tian and Z.-D. Jiang, 'A review of image set classification,' *Neurocomputing*, vol. 335, pp. 251–260, 2019, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2018.09.090. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523121831258X.

[37] C.-C. Chang and C.-J. Lin, 'Libsvm: A library for support vector machines,' *ACM Transactions on Intelligent Systems and Technology*, vol. 2, Jul. 2007.

[38] T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, *Optuna: A next-generation hyperparameter optimization framework*, 2019. arXiv: 1907.10902 [cs.LG].

Appendix A

Full Network Graphs of CNN Models

Figure 6.1: Tensorboard graph of the full SICNN network, which is just an implementation of Wide Resnet-50-2.

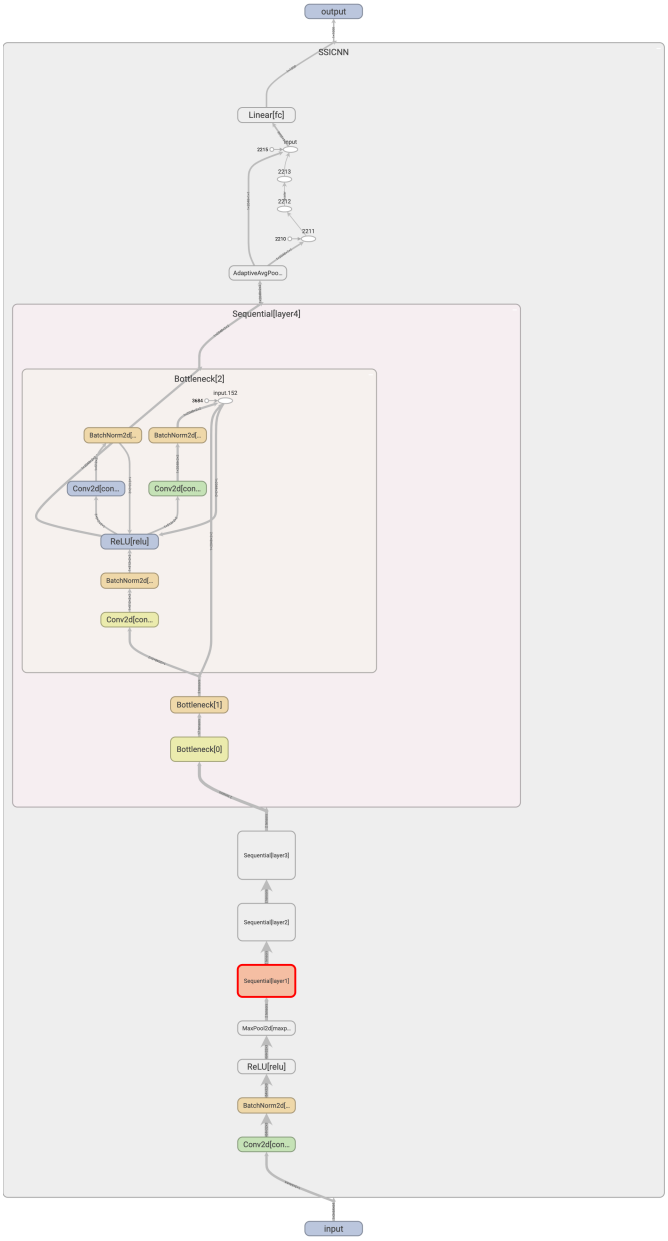


Figure 6.2: Opened up Tensorboard graph of the SICNN, showcasing the block based structure, with layers within Bottlenecks, and Bottlenecks within Layer blocks.

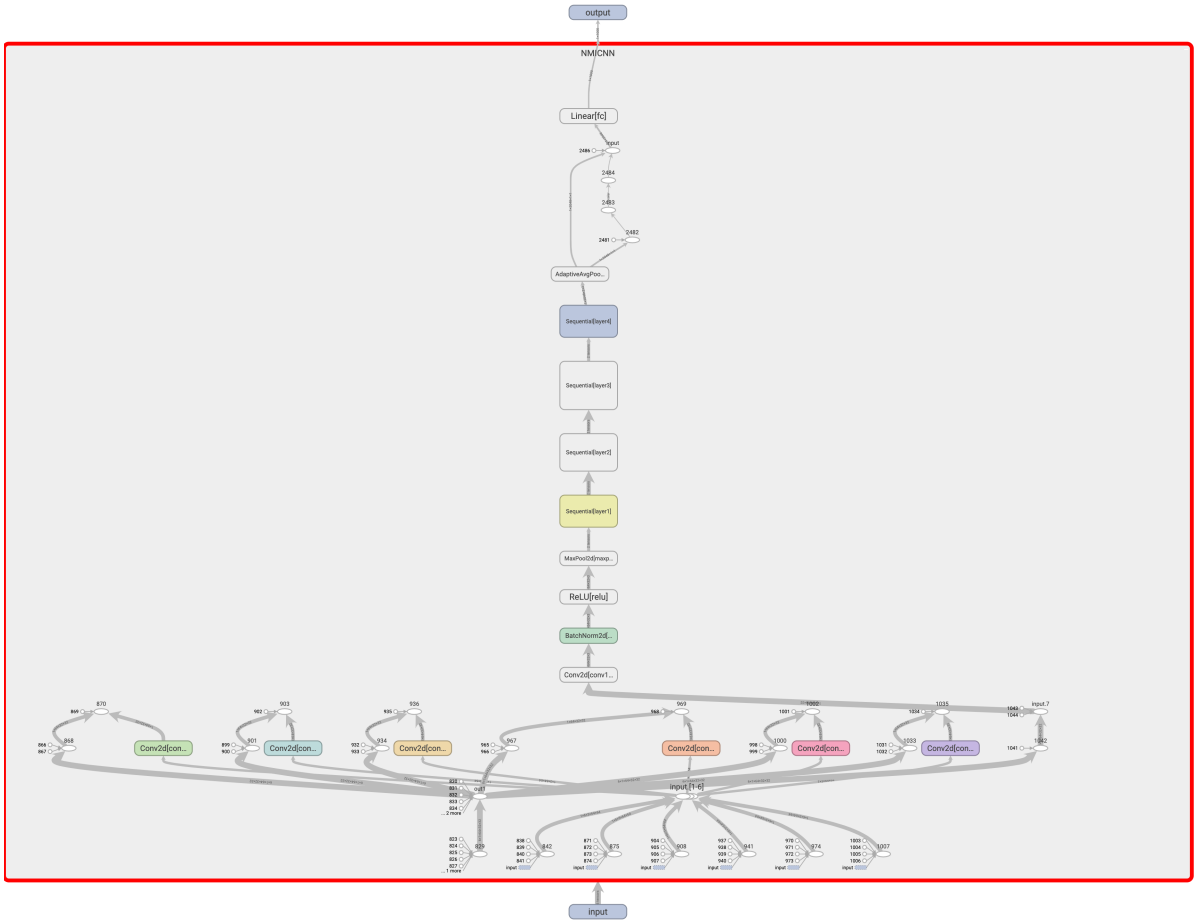


Figure 6.3: Tensorboard graph of the full SMICNN network, which is a modified Wide Resnet-50. It takes 6 images as a set, does a single convolution on each, before merging the outputs and completing the network.

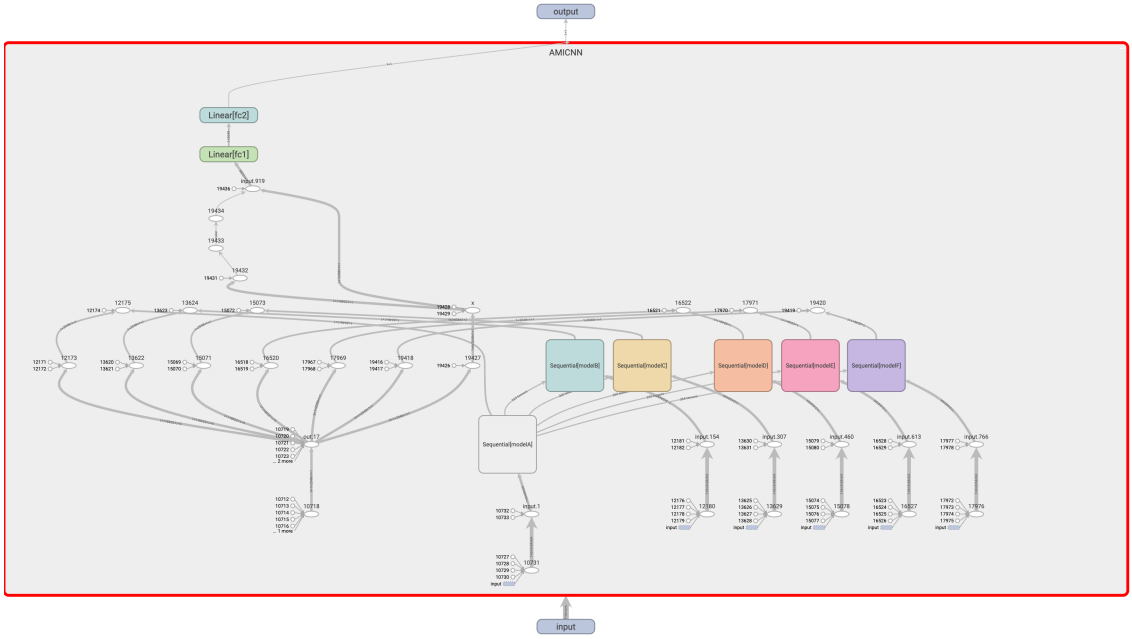


Figure 6.4: Tensorboard graph of the full CMICNN network, which is a modified Wide Resnet-50. It takes 6 images as a set, sends each through a full Wide Resnet-50 network, before merging and doing a couple fully connected layers.