2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Use of Convolutional Neural Nets and Transfer Learning for Prediction of Surgical Site Infection from Color Images

Richard Ribón Fletcher, *Member IEEE*, Gabriel Schneider, Bethany Hedt-Gauthier, Theoneste Nkurunziza, Barnabas Alayande, Robert Riviello, Fredrick Kateera

***Abstract*— One of the greatest concerns in post-operative care is the infection of the surgical wound. Such infections are a particular concern in global health and low-resource areas, where microbial antibiotic resistance is often common. In order to help address this problem, there is a great interest in developing simple tools for early detection of surgical wounds. Motivated by this need, we describe the development of two Convolutional Neural Net (CNN) models designed to detect an infection in a surgical wound using a color image taken from a mobile device. These models were developed using image data collected from a clinical study with 572 women in Rural Rwanda, who underwent Cesarean section surgery and had photos taken approximately 10 days after surgery. Infected wounds (N=62) were diagnosed by a trained doctor through a physical exam. In our model development, we observed a trade-off between AUC accuracy and sensitivity, and we chose to optimize for sensitivity, to match its use as a screening tool. Our naïve CNN model, with a limited number of convolutions and parameters, achieved median AUC = 0.655, true positive rate sensitivity = 0.75, specificity = 0.58, classification accuracy = 0.86. The second CNN model, developed with transfer learning using the Resnet50 architecture, produced a median AUC = 0.639 sensitivity = 0.92, specificity = 0.18, and classification accuracy 0.82. We discuss the specific training and optimization methods used to compensate for significant class imbalance and maximize sensitivity.**

I. INTRODUCTION AND MOTIVATION

A. The Burden and Challenge of Surgical Infections

The process of infection produces multiplication of microorganisms in the body tissues, which can produce competitive metabolism, toxins, intracellular replication or antigen-antibody response [1]. If not treated, an infection can become systemic and spread through the body causing sepsis and leading to tissue damage, organ failure and even death.

In most developed countries and wealthier communities, approximately 2% - 5% of patients develop Surgical Site Infections (SSI), directly resulting in approximately 0.64% of hospital deaths and also causing costly readmissions [2]. In low-resource settings, however, surgical site infections (SSIs) are an even greater concern, due to limited access to medical facilities and trained health personnel. In low-resource rural areas, such as Rwanda, for example, approximately 11% of women who have Cesarean section births develop a wound infection, which often puts the mother's life at risk [3].

Fig. 1. A community health worker capturing an image of a post-Cesarean section wound during a home visit in rural Rwanda.

In many developing countries, there exist community health workers that provide home visits to families and new mothers. However, these healthy workers are low-skilled and lack tools that can help to diagnose or screen for infection.

B. Standard Practice for Detecting Wound Infection

Conventional methods for detection of infection rely on subjective clinical signs, including heat, erythema (redness), swelling, pain, fluid discharge, and odor. Several published guidelines and manuals exist from government organizations such as the Centers for Disease Control (CDC/NHSN) in the U.S. [4]. In the past 40 years, some scoring systems, such as the ASEPSIS score, have been developed [5], but require some amount of training and clinical experience.

C. Digital Tools for Surgical Infection

Over the past 20 years, a variety of computer-based tools have emerged to assist with wound care, such as *+WoundDesk Wound Care* or *WoundCheck,* that are available to help monitor and document the healing of acute and chronic wounds [6]. The value of photographs to help improve the identification of infection has been studied [7]; however, such tools do not perform any automated analysis.

Smart phone platforms, such as *Tissue Analytics* [8], and Mobile Post-Operative Wound Evaluator (*mPOWER)* [9], enable patients to transmit data and photos of their wound to their doctors and health care providers. These tools can be used to automatically measure the size of the wound, but no prediction is given regarding the infection status.

Research supported by funding from the National Institutes of Health (No R21EB022368).

R. R. Fletcher and Gabriel Schneider are with the Massachusetts Institute of Technology, 77 Massachusetts, Cambridge, MA. 02139 USA. (phone: 617-694-1428, e-mail: fletcher@media.mit.edu).

B. Hedt-Gauthier and R. Riviello are with Harvard Medical School in Boston, MA.

F. Kateera, Barnabas Alayande, and T. Nkurunziza are with Partners In Health, Kigali Rwanda.

D. Prior Work and Current Contribution

In the past year, machine learning models have begun to emerge that can be used to predict infection [10]. Based on our initial work with a simple logistic regression model [11], we discovered that overfitting can be a serious concern in these models, and we also observed that that class imbalance is an important challenge in these data sets, which needs to be addressed explicitly. In this paper, we present two deep learning models used to predict wound infection, particularly addressing the problems of class imbalance and overfitting. We also discuss how the sensitivity of the model can be tuned to meet requirements for use as a screening tool.

II. Clinical Study

A. Study Design

Image data used for model development was collected as part of a study conducted by Harvard Medical School and Partners In Health (PIH), in Kigali, Rwanda. This study focused on Cesarean section surgery at the Kirehe District Hospital in rural Rwanda. The study included mothers who were at least 18 years old and who underwent Cesarean section births between March and October 2017. Women were enrolled prior to discharge and were provided a travel voucher to return to the hospital for a special visit at 10 days (+/- 3 days) after Cesarean section. Of the 729 eligible for follow-up, 572 (78.5%) women returned for this 10 day visit.

This clinical study was approved by the Institutional Review Boards (IRBs) of Harvard Medical School, MIT, and Kirehe District Hospital.

B. Data Collection and Labelling

Wound images were captured by designated community health workers using an Android tablet (Samsung Galaxy Tab 3). Each image was also given an independent clinical examination by a general practitioner doctor to determine if the Cesarean section wound was infected or not. From this examination data, it was determined that 62 of the 572 wounds were infected.

All images were wirelessly uploaded to a central server, and all images were cross-referenced with the list of SSI diagnoses by the doctor.

III. Naïve CNN Model

A. Algorithm Design and Implementation

Using the resulting image data, comprised of 62 infected and 510 non-infected images, we first created a few-layer naïve convolutional neural net (CNN) model. The benefit of using a naive model with limited number of simple layers is that it enables the use of all pixels in an input image, which avoids the need to resize or crop the image and potentially lose information.

For development, we used the Keras Tensorflow library (https://keras.io/) to compile and train these CNN models in Python. This model architecture was comprised of 3 convolutional layers ending with a dense layer to divide images into two classes. We used ReLU activation between layers, followed by a pooling layer. Finally, we added a fully-connected sigmoid activation layer (as opposed to SoftMax) with dropout at a rate of 0.5 to reduce overfitting, resulting in the final output label "0" for non-infected or "1" for infected.

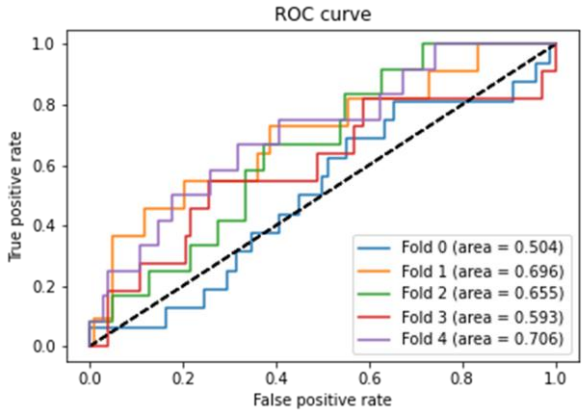

Fig. 2. The ROC curve for a trial of the naïve CNN model with median AUC = 0.655, showing range of specificity (false positive rate) and sensitivity (true positive rate).

B. Addressing Class Imbalance and Sensitivity

Given that the number of non-infected images (majority class) was 9 times larger than the number of infected images (minority class), it was necessary to properly compensate for this class imbalance. While it would be easy to achieve high classification accuracy by simply predicting all images as non-infected, the fundamental purpose of our project is to identify infected wounds. Thus, for this context, the *sensitivity* was more important than the *classification accuracy*, and we needed to choose an operating point for the model that would minimize *false negatives*.

In order to improve the model training to enable the model to correctly detect members of the minority class, we applied several standard methods described below:

Data Synthesis: As a standard method for growing the size of the minority class (infected wounds) we implemented the popular SMOTE method [12], which synthesizes new data that is statistically similar to the minority class. While this method is very effective, we limited the amount of data synthesis to a factor of 2 in order to avoid overfitting.

Class Weights: In order to compensate for class imbalance during training, the class weights were also modified within the Keras machine learning library. We ran trials of the model with different values of relative weights, from 0 to 20, and the best performance was obtained for relative weight of 9:1 (infected to non-infected), which was roughly inversely proportional to the prevalence rate of infected to non-infected images.

Creating a custom cost/loss function: A variety of custom loss functions were also explored to adjust the penalty for misclassified images during training. While this method did indeed enable the model to correctly more detect infected images (i.e. increase the true positive rate), it also significantly prolonged our convergence time during training. For our final models we decided to use instead a more conventional binary cross-entropy loss function, otherwise known as "log loss," which yielded comparable results while still allowing moderate run times.

C. Model Training

In addition to class imbalance, we addressed the general issues of generalizability and overfitting by modifying our data pre-processing to include a random flip, a random rotation, and batch normalization. Given the limited number of positive infected images, we also implemented k-fold cross validation, 1/k and (k-1)/k, for validation and training sets, respectively, with data splits of k = 5 and 10, corresponding to 13 and 6 infected images, respectively.

For model training, we also tested three standard optimizer algorithms (Stochastic Gradient Descent (SGD), Adam, and RMSprop) and two different amounts of training (30 epochs and 50 epochs). We did not extend beyond 50 epochs, since the convergence was adequate beyond 25 epochs and we wanted to avoid overfitting.

We ran over one hundred iterations of different combinations of weights, optimizer functions, training epochs, while also testing different the custom loss functions and drop-out rates for the final layer. For each combination of parameters, we ran 5-fold and 10-fold cross validation to generate a set of Receiver Operating Characteristic (ROC) curves and to calculate the median AUC value, sensitivity, and specificity. In order to better track the amount of false negatives, we also calculated the Matthews correlation coefficient (MCC), which is a more thorough metric for binary classification [13]. As we expected, 5 folds performed better, generalizing to more data and producing smoother ROC curves, while 10 splits resulted in more overfitting, despite exhibiting a higher AUC accuracy score.

D. Final Prediction Results

Our final Naïve CNN model results are shown in Figure 4. This model produced a median AUC = 65.5%, with an interquartile range, IQR = 0.27. The model yielded a true positive rate (sensitivity) = 75% and false positive rate (specificity) = 58%. As previously mentioned, higher values of AUC were possible, but we rejected models with sensitivity values < 75%.

IV. Transfer Learning Model

A. Algorithm Design and Implementation

As a possible improvement over our naïve CNN model, we also a *transfer learning* approach, which makes use of a pre-trained, highly sensitive neural network used widely for image classification. This approach uses a more complex architecture, with many convolutions and transformations. The hope was that this model would be able to find better detail and distinctions between the different image classes in order to perform binary classification. A main difference against the naïve model is that image preprocessing must be applied to our dataset in order to match the specification of the images on which the original model was trained. This preprocessing involves reducing the image pixel size to 224 x 224, which reduces detail in the image, in order to reduce the required computational training time. Given the contrast with the large image-size naïve CNN model, it was useful to explore and contrast these two different approaches.

For transfer learning, we chose the ResNet50 model, which is a 50-layer deep convolutional neural net model that has excellent demonstrated performance for image classification tasks (Figure 2). In order to customize to our application, we added some additional layers to the output of the ResNet50 model. This addition included a dense layer that outputs images into two classes, with a sigmoid activation, and using a node dropout of 20%. As in the previous model, we also added a random flip, random rotation, and batch normalization in order to improve generalizability.

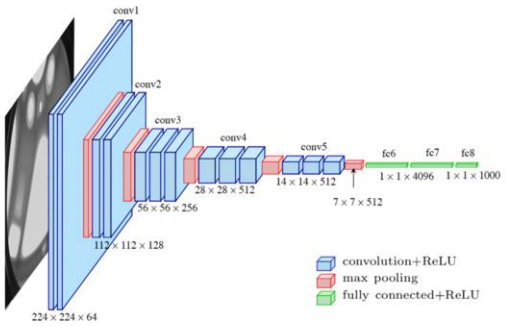

Fig. 3. The architecture of the ResNet50 neural net. Note large number of convolutions and fully-connected layers.

B. Model Training

As in the previous model, we used k-fold validation to find the training fold most representative of the remaining validation data, given our limited data size. Many trials were run with differing parameters, including number of epochs, optimizers, class weights, and splits.

C. Addressing Class Imbalance

For this transfer learning model, we addressed the class imbalance using the same methods as described previously. The most effective of these methods in the transfer learning model was moderate use of the SMOTE method to avoid overfitting, as well as adjustment of the class weights, which allowed the model to become more sensitive to the infected images.

D. Prediction Results from Transfer Learning

The resulting ROC curve for the transfer learning CNN model is shown in Figure 4. As can be seen from the figure, the median AUC of 63.9% is comparable to the naïve CNN model, with sensitivity = 93% and specificity = 18%. Although the overall performance was somewhat better than the naïve CNN, this model requires greater computational time to run, which needs to be considered for actual deployment.

V. Discussion

The summarized results are shown in Table 1. For comparison, Table 1 also shows our optimized logistic regression model that was been corrected for overfitting and class imbalance. As can be seen from these results, the neural net models perform better than logistic regression. However, the performance is only moderate.

Comparing the two CNN models, both models had similar performance, within the margin of error. However, in our training, we observed that the naïve CNN model has lower

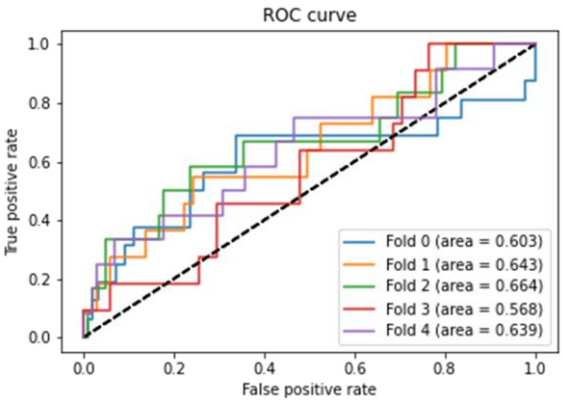

Fig. 4. The ROC curves for our final transfer learning CNN model with median AUC = 0.639.

variability across the different folds compared to the transfer learning model.

While the sensitivity of both CNN models is greater than 80%, (with classification accuracy > 80%), the AUC accuracy is fairly moderate at 65%. While these results show some promise, we are currently exploring additional color calibration and image processing methods to help improve performance. Further data collection is also needed to test the generalizability of these results.

While the ability to detect infection with a mobile phone image represents a useful contribution to global health, it is important to consider possible limitations of a color based model such as the ambient lighting and the dependence on patient skin color. Although we expect that the patient skin color is an important variable, it should be noted that in many of the communities encountered in global health, including this study, the skin color is fairly homogenous. Thus, it would be possible (and feasible) to develop several different models, depending on the skin color of the local population.

Given that neural net models have poor interpretability, we are also currently undergoing work to investigate specific properties of the image that contribute to the infection prediction, and it is known that CNN models are highly sensitive to both color and texture differences. Preliminary findings indicate that the wound color near the edge of the wound incision is particularly important. Additional data collection has also been recently conducted that will enable further validation of these algorithms with a larger number of images.

VI. Conclusion

We have developed two Deep Learning CNN algorithms to predict surgical site infection using images collected from mobile devices. Our results from our study with 572 patients demonstrate some ability to predict infection from a color image. The CNN models demonstrated sensitivity>80% which is sufficient as a screening tool; however, the overall median AUC accuracy of 64% is moderate. Nevertheless, given the widespread problem of surgical wound infection, the ability to predict an infection based on mobile phone images alone represents a promising new paradigm that should be further explored for preventing and treating infection. This technology represents an important advance for global health applications in low-resource regions as well as for outpatient care in wealthier developed countries.

Model	Sensitivity	Specificity	Median AUC	MCC
Logistic Regression	~ 60%	~ 70%	~ 0.563	0.14
Naïve CNN	~ 75%	~ 58%	~ 0.655	0.35
Transfer Learning CNN	~ 93%	~ 18%	~ 0.639	0.41

Table 1. Summary of overall results comparing performance of model prediction from questionnaire data and image data.

VII. Acknowledgements

We would like to acknowledge the great work of the field staff in Rwanda that enabled the collection of patient data, and also thank the patients themselves.

References

[1] Thomson, P.D. and Smith Jr, D.J., 1994. What is infection?.*The American journal of surgery*, *167*(1), pp.S7-S11.

[2] de Lissovoy, G., et al., "Surgical site infection: Incidence and impact on hospital utilization and treatment costs". Am J Infect Control, 37(5): (2009): 387-97.

[3] Nkurunziza, T., Kateera, F., Sonderman, K., Gruendl, M., et al., 2019. Prevalence and predictors of surgical-site infection after caesarean section at a rural district hospital in Rwanda. *BJS,106*(2), pp.e121-e128.

[4] Horan, T.C., Andrus, M. and Dudeck, M.A., 2008. CDC/NHSN surveillance definition of health care–associated infection and criteria for specific types of infections in the acute care setting.*American journal of infection control*, *36*(5), pp.309-332.

[5] Wilson, A.P.R., Weavill, C., Burridge, J. and Kelsey, M.C., 1990. The use of the wound scoring method 'ASEPSIS'in postoperative wound surveillance. *Journal of Hospital Infection*,*16*(4), pp.297-309.

[6] Gunter, R., Fernandes-Taylor, S., Mahnke, A., Awoyinka, L., Schroeder, C., Wiseman, J., Sullivan, S., Bennett, K., Greenberg, C. and Kent, K.C., 2016. Evaluating patient usability of an image-based mobile health platform for postoperative wound monitoring. *JMIR mHealth and uHealth*,*4*(3).

[7] Broman, K.K., Gaskill, C.E., Faqih, A., Feng, M., Phillips, S.E., Lober, W.B., Pierce, R.A., Holzman, M.D., Evans, H.L. and Poulose, B.K., 2019. Evaluation of Wound Photography for Remote Postoperative Assessment of Surgical Site Infections. *JAMA Surgery*.

[8] https://www.tissue-analytics.com

[9] Sood, R.F., Wright, A.S., Nilsen, H., Whitney, J.D., Lober, W.B. and Evans, H.L., 2017. Use of the Mobile Post-Operative Wound Evaluator in the Management of Deep Surgical Site Infection after Abdominal Wall Reconstruction. *Surgical Infections Case Reports*, *2*(1), pp.80-84.

[10] Wu, J.M., Tsai, C.J., Ho, T.W., Lai, F., Tai, H.C. and Lin, M.T., 2020. A Unified Framework for Automatic Detection of Wound Infection with Artificial Intelligence. *Applied Sciences*,*10*(15), p.5353.

[11] Fletcher, R.R., Olubeko, O., Sonthalia, H., Kateera, F., Nkurunziza, T., Ashby, J.L., Riviello, R. and Hedt-Gauthier, B., 2019, July. Application of machine learning to prediction of surgical site infection. *IEEE Engineering in Medicine and Biology Society Conference (EMBC) 2019*.

[12] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." Journal of Artificial Intelligence Research 16 (2002).

[13] Jurman, G., " The Advantages of the Matthews Correlation Coefficient (MCC) over F1 score and Accuracy in Binary Classification Evaluation," BMC Genomics, 2020.