Predicting the Severity of COVID-19 Respiratory Illness with Deep Learning

Connor Shorten
Florida Atlantic University
cshorten2015@fau.edu

Taghi M. Khoshgoftaar
Florida Atlantic University
khoshgof@fau.edu

Javad Hashemi
Florida Atlantic University
jhashemi@fau.edu

Safiya George Dalmida
Florida Atlantic University
sgeorge@health.fau.edu

David Newman
Florida Atlantic University
dnewma14@health.fau.edu

Debarshi Datta
Florida Atlantic University
ddatta2014@health.fau.edu

Laurie Martinez
Florida Atlantic University
lauriemartin2017@health.fau.edu

Candice Sareli
Memorial Healthcare System
csareli@mhs.net

Paula Eckard
Memorial Healthcare System
peckardt@mhs.net

Abstract

Patient care in emergency rooms can utilize urgency labeling to facilitate resource allocation. With COVID-19 care, one of the most important indicators of care urgency is the severity of respiratory illness. We present an early analysis of 5,584 patient records, of whom 5,371 (96.2%) have returned a positive COVID-19 test, to understand how well we can predict the severity of a respiratory illness given other features describing a patient using Deep Learning methods. The goal of our work is to illustrate the connection of our COVID-19 patient dataset with Deep Learning techniques, setting the stage for future work. The features in our dataset include when COVID-19 symptoms began, age, height, weight, demographics, and pre-existing conditions, to give a quick preview. We report train-test performance of a Deep Multi-Layer Perceptron (MLP) to predict the severity of respiratory analysis on a one-hot encoded scale of 5 labels. This 5-level scale is a truncation of our available labels, which we plan to extend and include in future work. We utilize a high-level of Dropout in order to avoid overfitting with our Deep Learning model. Further, we particularly study the impact of class imbalance on this dataset (Johnson and Khoshgoftaar 2019). We find that Random Oversampling (ROS) is an effective solution for decreasing minority class false negatives, as well as increasing overall accuracy. Readers will understand the performance of Deep Learning, with Dropout and ROS, to predict the severity of a COVID-19 patient's respiratory illness in which patients are described with Tabular Electronic Health Records (EHR).

Introduction

Improving patient care is one of the most important goals of applied technology. In this work, we utilize data-driven techniques to assess the severity of a COVID-19 patient's respiratory condition. The ability to predict the urgency of care facilitates resource allocation. Resource allocation is especially important for overloaded Healthcare systems, such as what was experienced in the COVID-19 pandemic. Within the umbrella of data-driven algorithms, Deep Learning has made many recent advances. We utilize 30 features describing 5,584 patients to evaluate Deep Learning models for the application of COVID-19 patient care. These features are selected because they do not require specialized hospital equipment to measure, such as a lab test to record white blood cell count and we aim to predict the initial severity of a respiratory illness before patients are admitted to the hospital.

More particularly, this work focuses on predicting the initial condition of a patient when they first enter the hospital. We utilize many features that could be documented prior to entrance. This includes when COVID-19 symptoms began, age, height, weight, race, ethnicity, and pre-existing conditions. As a prospective application for hospital care, patients, or family members of patients, could enter the information used in our models through interfaces such as mobile apps or websites before entering the hospital. Our models can then utilize this information to predict how severe a given patient's respiratory illness is. With this prediction, the hospitals can better organize resources for treating each patient and optimizing healthy outcomes. We acknowledge problems with this application such as if the patient is too sick to enter this information and does not have available family members for this. However, despite this limitation, we believe this is a great application to baseline applications of Deep Learning to patient care due to the simplicity of the problem setup. In future work, we plan to target additional modeling applications, such as forecasting the trajectory of respiratory illnesses while undergoing hospital care.

Our modeling technique builds off of recent research on applying Deep Learning to Tabular structured data. We use a Multi-Layer Perceptron (MLP) neural network architecture. Our Deep Learning model has a much higher parameter count than typical Machine Learning models. Our MLP models range from 250,000 to 1,000,000 parameters. Further, these parameters are organized in several layers of non-linear processing to capture complex interactions between variables. One of the pitfalls of using models like this is that they are very susceptible to overfitting. In our work, we combat overfitting with high levels of Dropout.

We classify the severity of COVID-19 respiratory illnesses into a categorical scale of 5 levels. Table 1 displays the frequency of each label in our dataset. After preprocessing, we find that 72.6% of our dataset is labeled as either a 3 or a 5. As we show in our experiments, this issue of class imbalance biases our Deep Learning model to predict 3 or 5, making frequent errors with patients labeled as a 4.

Severity	Frequency
1	97
2	250
3	1824
4	1183
5	2230

Table 1: A description of the frequency each categorical label appears in our dataset. These labels indicate the severity of a COVID-19 respiratory illness; 1 and 5 respectively represent the most and least severe illnesses.

Correctly distinguishing patients in this interval is crucial for the resource allocation application we are targeting. We alleviate the problem of imbalanced labels with the use of Random Oversampling (ROS), a common technique studied to remedy class imbalance in Machine Learning. ROS describes randomly selecting minority class instances to be duplicated in a training loop. Our experiments present the results of searching for the optimal percentage for this duplication.

In summary, our contributions are as follows:

- We present an analysis of Deep Learning techniques on a novel dataset of Electronic Health Records (EHR) describing COVID-19 patients.
- We find significant improvements on model performance with the use of ROS to handle imbalanced labels.

Related Work

Machine Learning for Healthcare

Machine Learning for Healthcare is one of the most important goals of Artificial Intelligence (Rajpurkar et al. 2022). There have been many recent advances in this research. The Medical Information Mart for Intensive Care (MIMIC-III) dataset of ICU septic patients has been an extremely useful benchmark to document the progress of Machine Learning systems (Johnson et al. 2016). The MIMIC-III dataset includes Demographics, Clinical Measurements, Interventions, Billing, Medical History, Pharmacotherapy, Clinical Lab Tests, and Medical Data. Compared to MIMIC-III, we mostly utilize Demographics and Medical History. We describe future work to integrate more data sources in our Discussion.

In tandem with Machine Learning for Healthcare and Deep Learning for Tabular data, our work is closely related to the more specific study of Deep Learning with EHRs. Xiao et al. (Xiao et al. 2018) have composed a systematic review of the opportunities and challenges in Deep Learning with EHRs. Rajmoker et al. (Rajmokar et al. 2018) present an analysis of modeling 216,221 adult patients described as EHR records. In this study, Rajmoker et al. predict in-hospital mortality, prolonged length of stay, and all of a patient's final discharge diagnoses. More closely related to our survey, Gong et al. (Gong et al. 2021) report the performance of Deep Learning techniques processing both Computerized Tomography (CT) images and EHRs. Gong et al. predict patient prognosis, whereas our experiments predict the initial severity of COVID-19 respiratory illnesses before patients have been admitted to hospital care.

Tabular Deep Learning

Our work builds off recent research on applying Deep Learning techniques to Tabular structured data. Tabular structured data is one of the most common data domains for studying Machine Learning techniques such as Logistic Regression, Decision Trees, Support Vector Machines, k-Nearest Neighbors, Naive Bayes, and XGBoost, to name a few. However, Deep Learning neural network architectures have mostly been developed for image and text data processing. Borisov et al. (Borisov et al. 2021) presented a survey of Deep Neural Networks and Tabular Data. Fiedler (Fiedler 2021) presented the results of a suite of training techniques such as Dropout, Activation Functions, and Batch Normalization, to name a few. Kadra et al. (Kadra et al. 2021) present a series of "regularization cocktails" showing that well-regularized MLPs can outperform specialized neural network architectures and the XGBoost algorithm on Tabular datasets. Similarly to these findings, we believe Deep Learning techniques are worth exploring further because of the control enabled in the optimization loop. For example, we achieve a massive boost in performance by modifying the traditional Stochastic Gradient Descent (SGD) update with a Random Oversampling (ROS) sampling technique.

Methodology

We present the results of predicting the severity of respiratory illness when a patient enters the hospital. Our dataset is sourced from the Memorial Healthcare System. We have access to 5,584 patient records for this analysis. We use 30 features to predict the severity of respiratory illness recorded at Day 0 in the hospital. These features include how long ago symptoms began, the results of a COVID-19 laboratory test, and demographic information such as age, height, and weight, to name a few. We note that the feature describing how long ago symptoms began can be subjective and varies on perspective from patient to patient. We utilize the Dropout regularization technique to avoid overfitting to spurious correlations of noisy features. We intend to explore additional regularization methods based on Data Augmentation in future work (Shorten and Khoshgoftaar 2019, Shorten et al. 2021, Shorten and Khoshgoftaar 2021). These features additionally contain pre-existing conditions such as cancer, hypertension, and chronic obstructive pulmonary disease (COPD), to give a few examples. We preprocess all features by integrating binary variables, categorical values mapped to a numeric scale, and continuous values. We fill in missing values with the median derived from each feature category. We plan to explore more algorithms for filling missing values in future work. We predict the categorical label of a patient's respiratory illness on a scale of 1 to 5. Our experiments are conducted by randomly assigning 80% of the dataset to train the models and 20% to a test set for evaluation.

Table 2 presents an initial comparison of a Deep Learning-based Multi-Layer Perceptron (MLP) architecture

Modeling Technique	Test Accuracy
MLP	73.6%
XGBoost	72.9%
Decision Tree	**73.9%**
kNN	43.4%

Table 2: A comparison of a Deep Learning MLP model with Machine Learning techniques.

with Machine Learning models such as XGBoost, Decision Tree, and k-Nearest Neighbors (kNN). MLP, XGBoost, and Decision Tree perform roughly the same, averaging 73.5% test accuracy. The kNN model performed very poorly at 43.4% test accuracy. Our MLP model contains 3 hidden layers with 256, 512, and 256 units. This is a fairly standard small-sized neural network design. Further, we utilize a large dropout rate of 70% between layers (including the input) to regularize the Deep Learning model and prevent overfitting. Dropout describes randomly assigning either inputs or intermediate activations, depending on where the Dropout layer is placed, to zero. With a Dropout parameter of 70%, either 70% of the input or 70% of the intermediate representation is assigned to zero. This forces the model to distribute representations of data and avoid overfitting. We find a strong result of Dropout to align increases in training accuracy with the held-out test set. Our models do not significantly increase in performance after 100 epochs, so we limit our experimentation to 200 epochs of training.

In Table 2, we use the total accuracy to report the performance of these models on predicting the severity of COVID-19 respiratory illness. We analyze the model's errors in a confusion matrix and reveal that it is biased towards predicting the majority class labels. For example, patients are 72.6% likely to have either a rating of 3 or 5 on our categorical label scale. This issue of class imbalance is a commonly studied problem in the Machine Learning literature. In this work, we experiment with the ROS technique to avoid majority class bias. In our initial model without any ROS, we find 152 false negatives with respect to the 4 label. In our experiments with oversampling, we find the most interesting performance changes to be around this region of the 3 and 4 labels. However, we do not see the same entanglement with the 4 and 5 labels. Thus, we highlight the 3 and 4 region for studying the ROS technique, reporting on the increases in false negatives in which true 3s are classified as 4s and vice versa, in addition to the increase in 4s correctly classified as 4s. Table 3 presents the initial confusion matrix of our Deep MLP trained without any ROS. As mentioned previously, our model makes 152 mistakes predicting 3 when the true label is 4. This error makes up 152 out of 282 (53.9%) of the total misclassifications on the held-out test set.

ROS is a data sampling technique that avoids majority class bias by duplicating instances belonging to the minority class. For example, an image classification dataset containing 80 dogs and 40 cats performs twice as many gradient updates on the dog images. ROS duplicates cat images such that the model makes a more equal number of gradient updates for each class. In our case, it is less common for a

Predicted	1	2	3	4	5	Total
1	0	0	1	0	0	1
2	0	0	0	0	0	0
3	13	51	315	152	5	536
4	1	2	27	58	1	89
5	2	0	16	14	421	453
Total	16	53	359	224	427	1079

Table 3: A confusion matrix of test predictions trained with the original training set.

Predicted	1	2	3	4	5	Total
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	9	40	243	106	3	401
4	5	12	100	106	9	232
5	2	1	16	12	415	446
Total	16	53	359	224	427	1079

Table 4: A confusion matrix of test predictions trained with 100% ROS.

COVID-19 patient to have a respiratory illness rated as a 4 on our categorical scale from 1 to 5 indicating severity. We randomly select instances to duplicate to remedy the bias of making majority class updates. We begin at 100% duplication, which duplicates every instance exactly twice. We then perform ROS by randomly selecting 50% and 25% of instances to duplicate at each training epoch. The results of these experiments are presented in Tables 4, 5, and 6.

We begin with 100% ROS. This eliminates the random component of the algorithm, deterministically doubling every instance in the training set. With this modification, our model goes from 89 total predictions of the 4 label to 232 total predictions on the held-out test set. Although this increases the correct predictions of the 4 label from 58 to 106, it increases false negatives predicting 4 when the ground truth is 3 from 27 to 100 (roughly a 370% increase). Overall, 100% ROS decreases the total performance severely from 73.6% to 70.8%. This result is fairly intuitive. By doubling the minority instances, we push the model to overfit to the particular details of these minority instances. This is why a ROS approach that utilizes randomness of selection is a better solution to alleviate class imbalance and simultaneously facilitate regularization.

Predicted	1	2	3	4	5	Total
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	10	49	300	127	1	487
4	4	3	44	83	1	135
5	2	1	15	14	425	457
Total	16	53	359	224	427	1079

Table 5: A confusion matrix of test predictions trained with 50% ROS. This is the best overall model produced in this report, as further described in Table 7.

Random Oversampling %	Total Accuracy
0%	73.6%
100%	70.8%
50%	**74.9%**
25%	72.7%

Table 6: The overall accuracy of the models trained with different levels of ROS.

As shown in Table 7, we find the best results from a 50% ROS sampling parameter. This describes duplicating half the minority instances in the training set at each epoch. However, the samples that make up the oversampled half change at each step, further providing a regularization effect in oversampling. Table 5 illustrates the confusion matrix of 50% ROS, increasing correct 4 predictions from 58 to 83 and decreasing false negatives of 3s when the ground truth is 4 from 152 to 127. We do not find much improvement with 25% ROS. We believe further performance could be squeezed out of a fine-grained ROS parameter search between 50% and 80%, which we leave to future work. However, it does not look like the optimal ROS parameter falls in the range of 0 to 50%. We further intend to explore a curriculum of gradually altering this parameter throughout training.

Discussion

We are working on developing Precision Medicine systems in collaboration with the Memorial Healthcare System with Deep Learning techniques (Shorten et al. 2021). Precision Medicine describes the utilization of data-driven algorithms to facilitate personalization of patient care and treatment planning. Our experiments illustrate how predictive analytics can be used to predict the severity of a patient's initial condition when entering the hospital. These predictions can be used to guide resource allocation and maximize healthy outcomes. More particularly, we have demonstrated how these datasets suffer from class imbalance and how we can remedy this problem with ROS.

We additionally intend to utilize the time-series structure of a patient's respiratory illness to forecast illness trajectories. In addition to the initial severity of a patient's respiratory illness, we have access to a time series rollout of 42 days of hospital stay. Further, we have information about Convalescent Plasma (CP) therapy as a treatment intervention. In future work, we intend to explore Offline Reinforcement Learning algorithms (Levine et al. 2020) to model patient trajectories based on CP therapy decisions.

Conclusion

In conclusion, we illustrate the effectiveness of predicting the severity of COVID-19 respiratory illness with Deep Learning based on initial data. We began with a comparison of Deep Learning techniques with Machine Learning models such as Decision Tree, XGBoost, and k-Nearest Neighbors. We further demonstrated how Random Oversampling can enable further control over the Deep Learning model and achieve the best performance overall. We have highlighted many directions for future work to develop Deep Learning models for COVID-19 patient care. For future work, we will consider different approaches to improve the results presented in this paper.

Acknowledgments

The authors wish to thank Memorial Health System for their helpful assistance with this project. We thank Maria Deane from the IT Department in Memorial Health for assistance creating the data set. We would also like to show our gratitude to Nithya Sundararaman from the Memorial Health Clinical Research Office for their support on the IRB submission and approvals process.

References

Borisov V., Leemann T., Sebler K., Haug J., Pawelczyk M., and Kasneci G. 2021. Deep Neural Networks and Tabular Data: A Survey. ArXiv:2110.01889.

Fiedler J. 2021. Simple Modifications to Improve Tabular Neural Networks. ArXiv:2108.03214. In Neural Information Processing Systems 35.

Gong K., Wu D., et al. 2021. A multi-center study of COVID-19 patient prognosis using deep learning-based CT image analysis and electronic health records. In European Journal of Radiology, Volume 139.

Johnson A. E. W., Pollard T. J., Shen L., et al. 2016. MIMIC-III, a freely accessible critical care database. In nature scientific data 3.

Johnson J. M. and Khoshgoftaar T. M. 2019. Survey on deep learning with class imbalance. In Journal of Big Data, volume 6, Article number 27.

Kadra A., Lindauer M., Hutter F., and Grabocka J. 2021. Well-tuned Simple Nets Excel on Tabular Datasets.

Levine S., Kumar A., Tucker G., and Fu J. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. In arXiv:2005.01643.

Rajmokar A., Oren E., Chen K., et al. 2018. Scalable and accurate deep learning with electronic health records. In npj Digital Medicine 1, Article number: 18.

Rapurkar P., Chen E., Banerjee O., and Topol E. J. 2022. AI in health and medicine. In nature medicine 28, pages 31-38.

Shorten C., Khoshgoftaar M. T., and Furht B. 2021. Deep Learning applications for COVID-19. In Journal of Big Data.

Shorten C. and Khoshgoftaar M. T. 2019. A survey on Image Data Augmentation for deep learning. In Journal of Big Data.

Shorten C., Khoshogftaar M. T., and Furht B. 2021. Text Data Augmentation for Deep Learning. In Journal of Big Data.

Shorten C. and Khoshgoftaar M. T. 2021. Investigating the Generalization of Image Classifiers with Augmented Test Sets. In Journal of Big Data.

Xiao C., Choi E., and Sun J. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. In Journal of the American Medical Informatics Association, Volume 25.