Using Machine Learning to Predict Patient Mortality

L. Castelli
Artificial Intelligence, Machine Learning Coursework
Submitted: May 5, 2021

The aim of this project was to compare the predictive power of three different machine learning models to estimate a patient's mortality risk on admission to hospital. We use the dataset put together by Xu et al., and the SMOTE-ENN sampling algorithm to balance the data. A random forest, a support vector machine and a neural network were trained. We find that the random forest was the best predictor with the highest accuracy of 0.9511 ± 0.0008 along with the highest precision. We note that all models trained had a low recall, and suggest that this is due to the sampling technique used and the lack of important data regarding high-risk comorbidities.

I. INTRODUCTION

The Covid-19 pandemic has infected over 150 million people, making it the largest outbreak of an infectious disease in over 100 years [1]. Over the course of a year, it is estimated that Covid-19 has claimed the lives of over 3 million people [1]. The majority of those deaths have come in the last 6 months [2].

One of the greatest dangers of a large-scale outbreak of an infectious disease is the overwhelming of hospitals. When healthcare systems are under extreme stress, the reduction in available healthcare resources can cause the mortality rate of other diseases and medical emergencies to increase [3]. In one of the most recent epidemics, the Ebola outbreak of 2014, it is estimated that the reduced resources led to an over 50% increase in malaria and tuberculosis deaths compared to pre-outbreak levels [3]. An analysis of mortality statistics in Scotland estimates that over 27% of the excess deaths during the pandemic are due to causes other than Covid-19 [4]. With the Covid-19 pandemic still at large globally, effective prioritization and usage of available resources will be vital in combating the virus.

When hospitals begin to come under stress, they are advised to triage their patients with the following criteria: save as many lives as possible [5]. They do this by prioritizing care for patients based on their probability of short-term survival [5]. Those decisions are made through evaluating the symptoms of the patient and relevant risk factors to determine who should receive care.

However, there is currently no one standardized approach through which this is done, which can lead to variability in the effectiveness of hospitals in providing treatment [6]. Multiple methods for evaluating patients have been proposed. Cho et al. developed a model to determine the number of days spent in hospital, while Shang et al. came up with a scoring system for patients based on their underlying conditions and symptoms which could be used to rate the risk of a patient dying while in hospital [7][8].

Machine learning has also been leveraged to address this issue. In 2020, 419 reports were written regarding machine learning and Covid-19, and over 19 of those publications focused on predicting disease progression [9]. A neural network was trained to identify a patient's mortality rate with an accuracy of 87% [6]. Other models focused on predicting days in either the hospital or the ICU, using a variety of different AI algorithms and approaches [9]. Each of the above was aimed at increasing the effectiveness of hospital resource management.

In this report, we attempt to investigate and compare different machine learning algorithms on their ability to predict the outcome of a patients admission to hospital, using information gathered at the point of admission.

II. METHODS

The machine learning models are trained and evaluated on retrospective data from the publicly available dataset put together by Xu et al. [10]. This dataset contains curated data compiled from municipal to national health reports, containing as much available information as possible regarding the patient. The age and gender were reported alongside location data, symptoms and comorbidities for over 2 million people who tested positive for coronavirus. A mixture of numerical and both nominal and ordinal categorical data are included.

We are primarily focused on individuals who were admitted to hospital, for which an outcome of the hospital visit is available. Individuals for which no outcome is recorded, either due to never having been admitted or due to not being inputted into the dataset are removed. The remaining outcomes were examined, and only the outcomes that were synonymous with either death or discharge were kept. It is noted that outcomes like "Stable" and "Unstable" were also removed, as it is unclear whether the patient was successful in their recovery. The outcomes were then converted into a binary column, with a 1 representing that the patient has died and 0 representing that they have recovered.

The location of a patient can be an important metric in determining individual hospital stress. Hospitals generally treat the patients in their vicinity. Hence, location can serve as a proxy for which hospital they have ended up in. Understanding how capable a given hospital is of successfully treating the patient will be important in determining the patient's mortality risk. Xu et al. used the best available estimate for their location, which was then converted into latitude and longitude [10]. However, for many of the patients, this data was not available beyond a provincial level. While this is not as ideal, this is still acceptable as hospitals will transfer patients between them, and hence if one hospital in an area is under stress, it is likely true for most of the hospitals in that area. Hence, patients were filtered further by requiring that at least province level location data was available.

Another important metric in determining hospital stress is the time of the year. The Covid-19 pandemic has come in multiple waves, with periods of extreme hospital stress occurring in response to a surge in cases [2]. While individual hospital admission statistics were not collected, by combining the time and the location this can likely serve as a reasonable replacement. Hence, wherever possible the date of admission to the hospital was included.

While some machine learning models are able to handle categorical data, for the sake of comparison, we standardize the data across all models and convert the categorical data into numerical inputs.

The symptoms were similarly cleansed as above, with typos or miscategorizations being removed. The symptoms of a patient were used to determine the severity of the disease. The different types of diseases are split into five classifications, as laid out in Wu et al.: Asymptomatic, Mild, Moderate, Severe, Critical [11]. The symptoms for the asymptomatic disease to the moderate disease are all relatively clear, and the symptom which was associated with the highest classification was used to determine the severity of the disease for a given patient. Each patient's symptoms were then assigned a number associated with their severity level, starting from 1 for asymptomatic to 5 for critical.

For severe and critical, the patients are more evaluated on oxygen levels. While this data is not available, some distinctions can still be made. Any symptoms which were consistent with a patient's organs failing or them going into shock were classified as critical, while the severe classification was reserved for a worsening or progression of moderate symptoms. For example, while pneumonia is a moderate symptom, severe pneumonia would be classified as severe.

While some symptoms were present that were not listed in Wu et al., they were infrequent, occurring less than 5 times across the whole dataset. Hence, they were ignored. Where available, the dates for when the patient started experiencing symptoms were also included, to help assist in identifying how far the disease has progressed at the point of admission.

The chronic diseases were also converted to numerical values by converting them into a series of binary classes. The comorbidities were first split into chronic diseases linked with an increased risk of a severe corona virus disease and all other type of chronic disease. The comorbidities associated with a higher risk were then split into three approximately even binary categories: whether or not they had diabetes, hypertension, and/or any other high risk chronic disease. For each patient, the above three categories and an additional binary category containing the lower risk chronic diseases were returned. The high risk comorbidities were classified according to the guidelines from the U.S. National Institute of Health [12].

All dates were converted into days since January 1st, 2019. If ranges were provided for either dates or ages, the center value of the range was taken, rounding down. Gender was encoded as a binary value, with females being encoded as a 1 and males encoded as 0s. Ages were returned as integers. Each patient was encoded into 12 numerical categories as described above. When a value was missing from a category, it was filled with a negative one value. The columns were subsequently individually normalized to between 0 and 1.

There is a significant imbalance in the dataset. Of the cleansed data remaining, only 5% of the cases result in the death of the patient. An imbalanced dataset can cause the majority class to dominate the much smaller minority class, causing the model to have significantly more error in predicting the minority class than in a more balanced dataset [13]. To combat this, it is common to use a mixture of over and under-sampling techniques.

For datasets with small numbers of positive cases, synthetic minority over-sampling combined with edited nearest neighbor (SMOTE-ENN) is one of the most effective modern techniques for balancing a dataset [13]. The SMOTE method functions through first generating additional samples via interpolation between existing minority class points. The new synthetic minority points tend to be noisy, so the ENN algorithm is applied, which removes points that differ significantly from their nearest neighbors.

The data was randomly divided according to a 70-15-15 train-validation-test split into three datasets. The train data was used to optimize the models, while the validation set was used to optimize the hyperparameters. Both datasets had the SMOTE-ENN algorithm applied to them. A final evaluation of the performance was done on the test set, which was unaltered. Each model was trained on 5 random splits of the data.

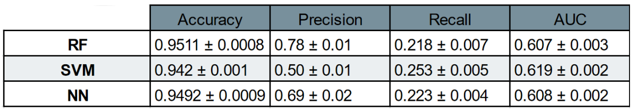
Three types of models were trained: an artificial neural network (NN), a random forest (RF) and a support vector machine (SVM). The first two types were used in the 19 publications on disease progression, and hence a comparison of the performance of the two may help future works optimize their choices for addressing this topic [9]. An SVM was also trained, as it has found success in modelling other areas of the pandemic and is a popular machine learning algorithm that is often considered in classifications problems like this [14]. Each model was optimized, and the parameters with the highest accuracy used.

The SVM uses a radial basis function kernel, with the penalty parameter C and kernel width γ optimized by using a grid search extending over a logarithmic range with 5-fold cross validation [15]. The radial basis function kernel was chosen as it tends to perform better on large datasets than other kernel choices [15]

The NN was modelled after the network used in Abdulaal et al. [6]. It contained a series of dense layers with rectified linear unit activation, with the final layer being a single sigmoid node to produce a prediction for the patient's mortality risk. Binary cross entropy was used as the loss function and the learning rate was reduced only when the validation loss plateaued. Early stopping and drop out layers were also employed to reduce overfitting. The weights and biases were initialized randomly and optimized according to the Adam algorithm. The model architecture was optimized by comparing performance on the validation set and adjusting the number of drop out layers to reduce any observed overfitting or underfitting. Each model was trained on three random states and the average taken, with the best performing architecture being chosen.

The depth of the tree, the minimum samples required to allow a split of a node in a tree, the maximum number of leaves in a tree and the minimum number of samples per leaf were optimized through using grid search and 5-fold cross validation on the train and validation datasets. We allow the maximum depth and maximum number of leaves to vary logarithmically between 10 and 1000, while we let the minimum samples per leaf and minimum samples to split a node vary between 10 and 100. The number of trees in the forest was set to 100.

After optimizing each of the parameters on the balanced train data, the predictive powers of the model were evaluated on the unaltered and unseen test data. This procedure was repeated 5 times, and the average value and uncertainty was returned. The accuracy, precision, recall and area under the curve (AUC) were all evaluated, and the recall vs precision for each of the 5 trials was plotted.



	Accuracy	Precision	Recall	AUC
RF	0.9511 ± 0.0008	0.78 ± 0.01	0.218 ± 0.007	0.607 ± 0.003
SVM	0.942 ± 0.001	0.50 ± 0.01	0.253 ± 0.005	0.619 ± 0.002
NN	0.9492 ± 0.0009	0.69 ± 0.02	0.223 ± 0.004	0.608 ± 0.002

TABLE I: The average metrics for the three different models with their uncertainty, calculate from the 5 trials.

III. RESULTS

The optimal parameters for the SVM were found to be C = 1000 and gamma = 1000. The best parameters for the RF were maximum depth= 100, maximum number of leaf nodes= 1000, minimum number of samples per leaf= 10, and minimum number of samples per node split = 17. The optimal NN was built using 4 hidden layers with nodes equal to 100, 70, 50 and 20 respectively.

The accuracy, precision, recall and AUC for the 34,257 patients evaluated was plotted in Table I. The random forest produced the highest accuracy at 0.9511 ± 0.0008, along with the highest precision of 0.78 ± 0.01. The highest recall and AUC were achieved by the SVM at 0.253 ± 0.005 and 0.619 ± 0.002 respectively.

The recall for each of the five trials is plotted against the precision for each of the different models in Figure I. There is a clear tradeoff between the high precision in the RF and the relatively high recall in the SVM. There was generally little variation between the individual trials of the model.

IV. DISCUSSION

Overall, the RF algorithm had the highest accuracy. However, given the intended use case for these algorithms, the accuracy is not the only metric which needs to be considered. While successfully identifying a patient who is likely going to die is important, it is even more vital that the algorithm does not identify a patient as likely to die if in actuality with treatment they could survive. In order for hospitals to be able to use this algorithm effectively, they will need to be able to trust that a positive result from the model is highly associated with a very high mortality risk. A failure in this regard will render the model next to useless, as the hospital will not be able to distinguish between a false positive and a true positive, and hence cannot take a risk on trusting the algorithm as it could potentially lead to unnecessary deaths.

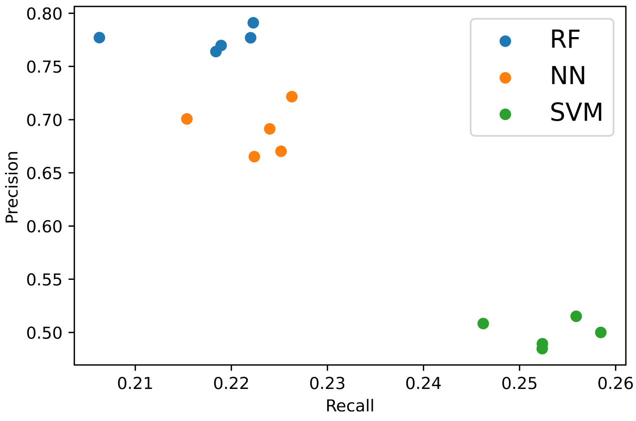


FIG. 1: The recall plotted against the precision for each of the five different trials for each model. There is a clear tradeoff between recall and precision between the RF and the SVM

Hence, precision should be the metric used alongside accuracy to determine which algorithm would be most effective. In this case, the RF model has both a higher accuracy and a higher precision, indicating that it is likely the most effective algorithm for this situation.

An inspection of the plot in Figure I reveals that each of the models tend to be consistent with their results, indicating that they are relatively stable. It is noted that the RF has a relatively large outlier for one of the trials, however a similar trend is also shown for the NN and the SVM, indicating that the random data split used in that particular trial may have been especially difficult to evaluate. It is noted that there is significant variation in the precision between each of the models while the recall remains relatively stable.

Another consideration is how limited all the models become given how poorly they performed in the recall metric. The recall metric evaluates how well a model classifies a positive case. Across all the models trained, the SVM model scored the highest in this metric, with a recall of 0.253 ± 0.005. Even the best performing classifier in this area only correctly identifies about a quarter of the positive cases. This will limit their impact, as very few cases will ever return a positive prediction and hence the hospital will still need to make many of their predictions without the assistance of the model. Hence, the models trained in this paper could only ever marginally improve the recovery rate of patients hospitalized with Covid-19. So while the models will still be useful due to their high precision, it is unlikely to drastically change the trajectory of the virus.

This trade-off between recall and precision may have been influenced by the data sampling techniques used during training. The SMOTE-ENN algorithm will generate more samples from the minority class (in this case patients who would go on to die in hospital), however they will be generated away from any borderlines as the edited nearest neighbor algorithm will typically delete data that is approaching the majority class. While this allows the generation of additional synthetic minority class samples to balance the dataset, it is usually the borderline samples which are misclassified [16]. Hence, as SMOTE-ENN fails to generate additional borderline samples, it will likely lead to continued misclassification of the minority class, as was observed in our models. This is likely also responsible for the observed little variation in the recall, as each of the models suffered from the above limitation. As this restriction was not as important for precision, there was significantly more variation. Future investigations into using machine learning models may find more success in the recall metric while using a different balancing technique which generates mores samples near the border, like Borderline-SMOTE [16].

Another potential limitation of the approach comes from the dataset used. While the dataset assembled by Xu et al. contains a wide variety of information, it lacks a few key bits of data which are vital for evaluating how the disease will progress. Previous investigations have found the obesity and smoking history are linked to an increased mortality risk, both of which would be available to a hospital at the point of admission [12][17]. However, neither of these values were included in the Xu et al. dataset. Another critical piece of information which would be collected at the hospital is oxygen saturation and body temperature, and the analysis of both of these metrics can give significant insight into the mortality risk of the patient [12][17]. The collection and distribution of a large dataset containing this information would likely lead to large improvements in all types of machine learning models.

The time and location were used as a substitute for hospital stress, however often the date of admission or date of symptom development was unavailable, and the location was usually no more specific than province level. While the classifiers were still able to perform relatively accurately, it likely was done without leveraging any information as a proxy for hospital stress. This is particularly vital as while hospitals in a similar area will face similar levels of stress, this is usually time specific. Given the range of our dataset, which only extended from the 1st of January 2020 to the 8th of June 2020, it is possible that this did not play as large of a role, as all the dates fall within the "first wave" of the virus. We suggest that the information given was likely not a workable proxy for hospital stress and suggest that the lack of information regarding this key statistic may have contributed to the poor recall observed. Hospital stress could likely have been better estimated by including statistics on the number of available intensive care units at the point of admission, or the current number of reported hospitalizations in the region or country at the time. The collection of this type of information should be a priority for future work.

The final consideration is with regards to variants. As mentioned above, the range of dates over which the dataset was collected extends only a few months into 2020. It is estimated that the first variant which was significantly more virial and deadly, B.1.1.7., mutated in mid-September [18]. Other significant variants also came into existence after this time. Hence, it is likely that there was little to no overlap in these time periods, and any affects due to the presence of variants can be ignored.

V. CONCLUSIONS

In this report, we have implemented three different types of machine learning algorithms to try and predict a patient's mortality risk as they enter a hospital. Using the dataset from [10], we are able to achieve high levels of accuracy with all three algorithms, and a particularly large precision score with the random forest model.

We note that none of the models trained achieved high recall, suggesting that they are not adept at identifying patients who have a large mortality risk, which may limit their usefulness in the field. This is likely attributed to the sampling technique used, as while SMOTE-ENN generates additional samples they are typically away from the boundary, which can allow the majority class of low mortality risk patients to dominate the minority class. Limitations in the dataset used were also identified as a possible limiting factor.

Future investigations should aim to improve the available data, which may allow the training of more effective models. The usage of different sampling techniques like Borderline SMOTE-ENN may also help increase the recall, however it is noted that false positive rate should be minimized primarily to allow hospitals to confidently use the model. The training and implementation of an effective prediction model could significantly improve a hospital's ability to treat patients, and potentially save thousands of lives.

REFERENCES

[1] Dong, Ensheng, et al. ”An interactive web-based dashboard to track COVID-19 in real time.” The Lancet, vol. 20, no. 5, 1 May 2020.

[2] Centers for Disease Control and Prevention. ”COVID Data Tracker Weekly Review.” Covid-19, U.S. Department of Health & Human Services, 30 Apr. 2021.

[3] Parpia, Alyssa S., et al. ”Effects of Response to 2014–2015 Ebola Outbreak on Deaths from Malaria, HIV/AIDS, and Tuberculosis, West Africa.” Emerging Infectious Diseases, vol. 22, no. 3, Mar. 2016.

[4] Docherty, Kieran F., et al. ”Excess deaths during the Covid-19 pandemic: An international comparison.” medRxiv, 13 May 2021.

[5] Joebges, Susanne, and Nikola Biller-Andorno. ”Ethics guidelines on COVID-19 triage—an emerging international consensus.” Critical Care, vol. 23, 6 May 2020.

[6] Abdulaal, Ahmed, et al. ”Prognostic Modeling of COVID-19 Using Artificial Intelligence in the United Kingdom: Model Development and Validation.” Journal of Medical Research, vol. 22, no. 8, 25 Aug. 2020.

[7] Cho, Sung-Yeon, et al. ”Prognosis Score System to Predict Survival for COVID-19 Cases: a Korean Nationwide Cohort Study .” Journal of Medical Research, vol. 23, no. 2, 22 Feb. 2021.

[8] Shang, Yufeng, et al. ”Scoring systems for predicting mortality for severe patients with COVID-19.” The Lancet, vol. 24, 1 July 2020.

[9] Syeda, Hafsa B., et al. ”Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review.” Journal of Medical Research, vol. 9, no. 1, 11 Jan. 2021.

[10] Xu, Bo, et al. ”Epidemiological data from the COVID-19 outbreak, real-time case information.” Scientific Data, vol. 7, 24 Mar. 2020.

[11] Wu, Z., and J. McGoogan. ”Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72,314 cases from the Chinese Center for Disease Control and Prevention.” JAMA, vol. 323, no. 13, 2020, pp. 1239-1232.

[12] National Institutes of Health. ”Overview of COVID-19.” COVID-19 Treatment Guidelines, 21 Apr. 2021.

[13] Batista, Gustavo E., et al. ”A Study of the Behavior of Several Methods for Balancing machine Learning Training Data.” ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, June 2004, pp. 20-29.

[14] Hao, Yan, et al. ”Prediction and analysis of Corona Virus Disease 2019.” PLOS ONE, vol. 15, no. 10, 5 Oct. 2020.

[15] Prajapati, Gend, and Arti Patle. ”On Performing Classification Using SVM withRadial Basis and Polynomial Kernel Functions.” 2010 3rd International Conference on Emerging Trends in Engineering and Technology, 19 Nov. 2010.

[16] Han, Hui, et al. ”Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning.” ICIC’05: Proceedings of the 2005 international conference on Advances in Intelligent Computing, vol. 1, Aug. 2005, pp. 878-887.

[17] Yadaw, Arjun, et al. ”Clinical predictors of COVID-19 mortality .” The Lancet, vol. 2, no. 10, 1 Oct. 2020.

[18] Galloway, Summer E., et al. ”Emergence of SARS-CoV-2 B.1.1.7 Lineage — United States, December 29, 2020–January 12, 2021.” Morbidity and Mortality Weekly Report, vol. 70, no. 3, 22 Jan. 2021.