Predicting Covid-19 mortality from biomarker data: a comparison of different machine learning models*

Primary Topic: Data Mining, Secondary Topic: Data Processing & Visualisation
Course: 2020-1B – Group: 38 – Submission Date: 2020-01-30

Alex Bujorianu
University of Twente
a.bujorianu@student.utwente.nl

Evangelos Gougoulis Dimitriadis
University of Twente
e.gougoulisdimitriadis@student.utwente.nl

ABSTRACT

This paper aims to investigate whether mortality from Covid-19 in a hospital setting can be predicted using biological test data. Additionally, this paper aims to determine how many biological indicators are needed for an accurate prediction. In order to tackle the first research question, six different machine learning models were trained on a dataset of 375 patients from a hospital in Wuhan. All of the models had accuracy greater than 92%. The authors recommend support vector machines due to their easy interpretability; high discrimination ability measured by Area Under Receiver Operating Curve (AUC); and suitability for medical data. Regarding the second research question, principal component analysis indicated that 90% of the information was contained in 5 variables, with 85% of the variance explained by just 3 variables. The clinical implications of the results are assessed in the context of the pandemic response.

KEYWORDS

COVID, SARS-CoV-2, Support Vector Machine, Logistic Regression, Serum Chloride, ACE2, C-reactive protein, Cytokine Storm, Lactate Dehydrogenase

1 INTRODUCTION

The 2019 novel coronavirus (2019-nCoV, later changed to SARS-CoV-2) has been extremely challenging for health professionals to diagnose and treat, in part because of uncertainty regarding risk factors and disease progression. In this paper, aim to predict the COVID-19 patient outcome (deceased or discharged) by constructing and comparing various machine learning models. Additionally, we define the biomarkers with the largest explained variance on predicting COVID-19 mortality. We use PCA (principal component analysis) to determine which variables are strong predictors. In doing so, we can aid health professionals by identifying at-risk patients, allowing hospital resources to be focused on the patients which require the most care.

Moreover, the results are also relevant for informing vaccination campaigns, such as the rollout of the Pfizer/BioNTech and Oxford/AstraZeneca vaccines in the EU, US, UK and other countries. This is because some of these data are already commonly available in patient databases, for example, serum chloride.

*Adapted from the ACM SigConf Template. More information about the template can be found at https://www.acm.org/publications/proceedings-template

2 BACKGROUND

2.1 Zoonotic Origin

SARS-CoV-2 is a zoonotic β-coronavirus in the same family as MERS (Middle Eastern Respiratory Syndrome) and SARS (Sudden Acute Respiratory Syndrome) to which it is closely related. [17] The β-coronavirus family also includes many human coronaviruses responsible for the common cold, such as HCoV-229E. [1] Interestingly, a number of human coronaviruses are also capable of infecting bats, from which SARS-CoV-2 is believed to have originated, perhaps via civets. (ibid.) This suggests strong inter-species transmission and gene exchange. Bats may be to coronaviruses what pigs are to influenza viruses: mixing vessels. [12]

2.2 Infiltration of Human Cells

The virus binds to Angiotensin-converting-enzyme-2 (ACE2) receptors in human cells via the spike glycoprotein. [17] The upregulation of the ACE2 receptor is one of the reasons why Covid-19 leads to much higher fatality in geriatrics compared to the general population. [2]

2.3 Biomarkers: The Role of Inflammation

Inflammation is a complex process governed by a variety of different cells, and influenced by multiple environmental factors. It is necessary for life, and yet inflammation is increasingly associated with a variety of chronic, autoimmune diseases like lupus erythematosus [10]; certain cancers; metabolic illnesses like heart disease and diabetes; as well as Covid-19 outcomes. [11] Inflammation has received a lot more attention from researchers in the past 10 years or so, yet our understanding remains limited.

In the case of CovidD-19 specifically, a number of important pathways have been identified, which determine how the virus interacts with the immune system and the severity of the infection:

ACE2: The ACE2 receptor is not just the entryway for the virus; it plays an important role in immune system regulation and cardiovascular function as well. Liu et al. [11] discuss the mechanisms by which the virus affects ACE2 production. Notably, ACE2 has a vasodilating effect and prevents heart failure; the elderly tend to have lower levels of cellular ACE2 present (and higher levels in the blood, which is why serum ACE2 is marker for heart disease and diabetes). By damaging this cell, the virus increases the likelihood of catastrophic heart failure. Moreover, ACE2 inhibits angiotensin-2, which is strongly pro-inflammatory [9]

Cytokines, most notably Interleukin 1-β, IL-6 and C-reactive protein: Cytokines are proteins that regulate the immune system. When cells are damaged during infection (either directly by the

virus, as a result of byproducts from virus-killed cells, or through autoimmune effects) they emit pro-inflammatory cytokines. Normally, this is not harmful, and in fact promotes wound healing: consider the example of a scraped knee. Traumatised cells secrete cytokines which tell lymphocytes to kill them (they are basically eaten). This allows the body to easily clear away the dead cells and start producing new cells and new tissue.

However, in the case of a viral infection, this process is global rather than localised—it affects tissues in numerous organs, including the circulatory system. The production of cytokines leads to more inflammation, more cells dying, which leads to even more cytokines being produced, resulting ultimately in multiple organ failure and death. This positive feedback loop is known as a "cytokine storm" in the literature. This effect is not limited to Covid-19—a similar phenomenon was observed with influenza patients during the 1918 pandemic. [16]

Interferon production: Interferons are important because they block the replication of viral cells, primarily by signalling natural killer cells and other T-cells to attack virus particles and infected cells. Interferons have been found to affect MERS survival, another coronavirus-disease [11] and they have been used to treat Covid-19. One early study in Wuhan found interferon 2αeffective in reducing viral load. [20] More promising, randomised, double-blinded and controlled studies have found interferon β-2b to reduce time-to-recovery, improve hospital discharge rate, and reduce mortality. [14]

Additionally, Liu et al. [11] discuss how SARS-CoV-2 seems to reduce CD4+ lymphocyte count, which gives COVID-19 some AIDS-like characteristics.

3 DATASETS

3.1 Preprocessing

Data preprocessing was undertaken in the R statistical computing language. [13] The data was courtesy of Li et al. [18]. Two datasets were processed: dataset 110, which is an Excel file containing the complete data for 110 patients with the variables lymphocytes, lactate dehydrogenase and C-reactive protein; and dataset 375, which is a much larger dataset containing several variables. Both the code and the data are open-source and can be found on our Github repository: https://github.com/Alex-Bujorianu/DS-project

The main challenge with our dataset was the amount of missing data, either for patients (in which whole rows were missing) or variables, such as TNF-α, which was missing over 40% of its values. We had four possible options, all of which came with their own set of tradeoffs:

(1) Removal: simply clean out all the missing cells until the remaining rows contain real data for every variable. A full removal was obviously impractical because it would have made the dataset very small—a dataset that would further be split into training and validation sets. This would have resulted in large p-values and statistically insignificant results.
(2) Mean imputation: the main advantage of this method is that it is relatively straightforward to perform. However, it has a number of important disadvantages [5]. Notably:
    (a) Missing-at-Random (MAR) assumption. This method assumes that values are missing at random, otherwise imputation introduces additional biases. To test this assumption, I made a separate dataset containing the patients who were missing all data, to see if there were any unusual characteristics (high mortality rate, age, etc.) The only thing those patients had in common was that they spent a small amount of time in hospital—which explains why they didn't have any test results. They were not associated with unusually high or low mortalities, nor were there any demographic differences in terms of age or gender.
    (b) Parameter biasing: If the purpose of a study is to compare means, mean imputation is OK, assuming the MAR assumption is satisfied. However, we are interested in the relationship between variables, namely the direct effect of a predictor variable (like lymphocytes) on the boolean outcome variable, survival. In addition, the interaction effects are also very important. Mean imputation can artificially lower (or increase) correlations between variables.
    (c) Underestimating standard errors. Mean imputation, regardless of whether the values are missing at random, always leads to artificially low p-values. This is because we are assuming real data in our model when in fact an imputed value is just a "guess". More technically, it's the result of having reduced variance as the imputations have 0 variance.
(3) Multiple Imputation: This method generates multiple datasets with pseudo-random data for missing items. The datasets have to be modelled separately and the model is then combined. This has the advantage of not underestimating standard error, however, parameters will be biased if the data is not Missing-At-Random. The problem with this approach is that it is complicated and difficult to perform. In particular, it assumes that the imputation model is compatible with the analysis model. [8]

Ultimately, we decided on a combination of removal and mean imputation: we removed patients with no data (which cut down the observations from 375 to 350) as well as variables missing > 6% of their values (which dropped us to 48 variables). The small number of missing NAs remaining were replaced with means. Our justification for this approach is that, if the number of imputed values is small, the effect on standard errors would also be small. As for dropping variables, the rationale is that a) variables with > 40% of values missing cannot be used for making trustworthy predictions anyway (ibid.) and b) it is unlikely that a model would need > 48 variables to make accurate predictions. This intuition proved correct, as our results on principal component analysis show.

A special note regarding TNF-αand IL-6: we were forced to drop these variables from our analysis because of high missingness scores (greater than 40%). However, we know that these biomarkers are likely to be important predictors of mortality, based on the literature. The reason so many values are missing is because these tests are non-trivial to do. We could have looked at the ĩ90 patients that did have data instead. But since we already had a number of performant

models, it was not really necessary. This is something we would investigate in a future study.

Multiple imputation was too complex for our dataset, therefore we dismissed it.

3.2 Training and Test Sets

We decided on splitting the larger dataset (with 375 patients) into training and test sets rather than using the smaller dataset of 110 patients. The main justification for this was that after performing the Principal Component Analysis, it was evident that serum chloride was a better predictor than lymphocytes (which is what the smaller dataset contained).

3.3 Summarising Values by Latest instead of Mean

We considered whether it would be best to summarise the test results for each patient by the mean of all test results, or whether only the most recent value should be considered. The rationale is that more recent test results are a more valid indication of the patient's final outcome, e.g. a sudden rise in C-reactive protein might signal an impending cytokine storm. However, summarising values by mean reduces the variance and makes the data more closely resemble a normal distribution (because of Central Limit Theorem).

In the end, there was no meaningful difference between the two datasets, which is why we do not present the results of the second dataset (the reader may inspect them on Github). In practical terms, we understand the difference can be quite significant, however: a hospital would prefer to avoid having to perform multiple tests on a schedule. The practical implication then, is that hospitals can take only one test if they don't have time for more, as the prediction will still be quite accurate. However, a time series analysis would definitely be an avenue for interesting future research.

3.4 Critical Reflection on the Datasets

Datasets can be characterised by the four Vs approach developed by IBM, namely: volume, variety, velocity and veracity. We were fortunate in that our dataset was mostly homogenous (nearly every variable was a continuous positive number). We did have some trouble with the dates, as R would throw up type errors, misinterpret the date variable in the grouping, and some library functions wouldn't accept dates in the dataset. We mostly worked around these issues by selectively applying functions to the numerical columns.

Velocity was not an issue, and while the datasets are quite small, they large enough to train models and obtain significant results. We should note that data imbalance was not a problem in our dataset.

Veracity is perhaps the most difficult problem for any dataset, especially in medicine. In many cases, hospitals still rely on health care workers to enter data manually. Transcription errors are quite common. [6] Nevertheless, there is little we can do about this, except caveat our results.

4 RELATED WORKS

Zhou et al. [19] used logistic regression modelling on the 110 patient dataset, and had very good prediction accuracy (over 90%). Therefore, our first goal was to see if we could obtain similar results using logistic regression.

Our next goal was to compare several machine learning models based on a set of criteria: discrimination, which is the ability of the model to separate patients who will die from patients who will survive; and calibration, which is the agreement between observed and predicted mortality. A model with high discrimination but low calibration is overfitting—that is, the model is training on noise in the data rather than real features. Conversely, a model with a high agreement score is much more reliable but probably not very useful if the discrimination ability is low. Generally, there is a trade-off to be made.

Das et al. (2020) [3] trained a variety of different machine learning algorithms—including logistic regression but also gradient boost, random forest and support vector machines (SVM)—and found logistic regression to be the best model.

5 METHODOLOGY

Model training and parameter tuning was undertaken in Python. The outcome variable was mortality, coded as 0 (discharge) and 1 (death). We did not attempt any time series analysis. We ended up comparing six models, some also used by Das et al. [3]: Logistic Regression, Decision Tree, AdaBoost, Support Vector Machine, Neural Networks and Naive Bayes. Our parameters for each model chosen are presented below.

Logistic Regression: the most obvious approach for predicting a binary outcome variable. The main issue with logistic regression is that it assumes predictor variables have a linear effect on the outcome variable. It does take into account interaction terms, which Zhou [19] emphasises as being very important. In our testing we acquired better results by using the limited-memory Broyden–Fletcher–Goldfarb–Shanno (lbfgs) solver which approximates the second derivative matrix updates with gradient evaluations; it is relatively slow but our dataset was small. A faster but slightly less accurate approach was the use of the library for Large Linear Classification (liblinear) solver, which is based on minimizing a multivariate function by solving univariate optimization problems in a loop. This is equivalent to descending to the optimal point one direction at a time.

Decision Tree: Decision trees are better suited for discrete data rather than continuous data, but they can sometimes perform well even in the case of continuous data, in which case the data has to be grouped according to ranges. We constructed a decision tree with a maximum depth of three. In our testing there was no added benefits on building a larger tree. In addition, using entropy as a node splitting criterion resulted in slightly better performance than the default Gini criterion.

AdaBoost: AdaBoost classifier makes a forest of trees with maximum depth one (stumps). These trees are being constructed with the tree classifier used in our previous decision tree model. All these weak predictors are combined to make a stronger model.

Support Vector Machine (SVM): SVM is another type of classification algorithm that uses the maximum margin classifier to label observations. The basic idea is that an observation is classified based on how close it is one group or the other. Choosing the

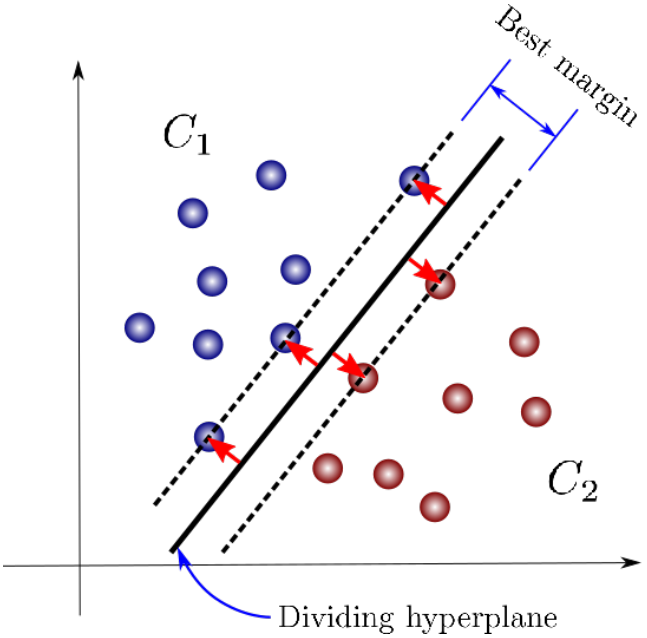

Figure 1: An example of a support vector machine.

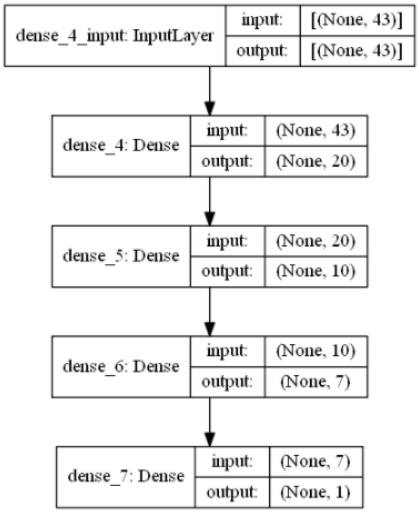

Figure 2: Neural Network architecture

threshold to be the midpoint between groups maximises the margin. The problem, of course, is that there are always some outliers in a dataset, and SVMs are extremely sensitive to this—they have low bias but high variance. Therefore, a trade-off has to be made between low variance – which means the model is reliable – and high precision (low bias). So we tuned the parameters accordingly. We set the regularization parameter (C) at 100. This dictates how much our model will avoid misclassifying each training example. The larger the value of C, the smaller the margin of the chosen hyperplane. For very tiny values of C, you should get misclassified examples, often even linearly separable dataset. However, the Covid-19 dataset was not linearly spreadable; a linear kernel was not the ideal solution. It resulted in 90% accuracy with precision and recall values that were below 90%. Using the radial basis function (rbf) kernel—which maps training data to a vector of higher dimensionality—gave us greater performance. We achieved up to 98% accuracy with ROC under the curve score of 0.98%.

Neural Networks: This approach is very different from the other models we have used, which employ quite simple and deterministic "formulas" for classification. In short, neural networks take in a whole bunch of data as input, and output a number of best guesses as the classification. The exact way it does this is complex and requires a lot of fine-tuning; it relies on breaking down the input into sub components (e.g. loops in an 8) and trying to find specific patterns of neuron activation. Neural networks can be extremely powerful but they generally require a lot of training data. Furthermore, neural networks are often "black boxes": although it is entirely possible to know how the hidden layers work mathematically—it is basically a sigmoid function with a weighted sum and a bias term—the sheer number of parameters makes it impractical. We thought it was worth a try because it is easier for a neural network to predict a small number of classifications than a large number – and our problem has only two classifications.

Parameter Tuning: In our Neural Network approach we did the following:

- Input 43 values one for each of our attributes.
- Constructed three hidden layers where each neuron had a rectified linear activation function (ReLU), which is a simple intuitive activation function that outputs the input value if it is bigger than zero. ReLU activation functions are becoming increasing popular both for their good performance and their small computational burden [4]. The first hidden layer contains twenty neurons, the second hidden layer contains ten neurons and the third hidden layer just seven.
- The output layer is a single neuron with a sigmoid activation function. With this architecture we have a binary outcome of 0 (discharge) and 1 (death).
- The optimizer of choice was the stochastic gradient descent (SGD) optimizer. It performed slightly better that RMSprop and Adam (an SGD variant) [7].
- The selected loss function was binary cross-entropy which computes the cross-entropy loss between true (test) labels and predicted labels.

Overall, after optimizing the aforementioned architecture we manage to achieve up to 94% accuracy and 94% ROC area under the curve. In Figure 1 the used is also graphically depicted.

Naive Bayes: In this case a different generative approach was taken. This model classifies observations based on their probabilities, so for example "if the sky is overcast today, there is a 75% chance it will rain". The main problem with naive Bayes is that it assumes events are independent of each other. [15] This is not always the case, but the model performs surprisingly well. Using a Gaussian distribution to model our attributes resulted to 0.94% accuracy and 0.96%ROC area under the curve.

6 RESULTS & ANALYSIS

6.1 Executive Summary

A summary of the model results is presented in the table below. In general, all the models achieve comparable performance. The decision tree and Neural Network models have average performance, while SVM and Naive Bayes are the best performers. The metrics were obtained using 5-fold cross-validation.

Model	F1 score	ROC AUC	Accuracy
Logistic Regression	0.92	0.97	0.93
Decision Tree	0.93	0.95	0.93
AdaBoost	0.93	0.98	0.93
SVM	0.94	0.98	0.95
Neural Network	0.94	0.94	0.94
Naive Bayes	0.94	0.96	0.94

Table 1: Summary of model performance.

6.2 Explanation of Results

Accuracy is the most intuitive of the three metrics: it is simply the ratio of correctly predicted observations (either positive or negative) to the total number of observations. Mathematically,

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

To understand the F1 score, it is necessary to understand recall and precision. Precision is the ratio of true positive predictions relative to the sum of true and false positives. Recall is the ratio of true positive predictions relative to the sum of true positives and false *negatives*.

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

The F1 score is a weighted mean of precision and recall. It is not as intuitive as accuracy, but it is a better measure of model performance when there is class imbalance, i.e. if false positives have a different cost compared to false negatives. Accuracy is only appropriate when we want a symmetrical trade-off. In our case, false positives are about as harmful as false negatives, so the accuracy score is a reasonable measure.

The Receiver Operating Curve & Area Under Curve (ROC AUC) is a measure of model performance at different thresholds. The ROC plots the true positive rate on the Y-axis against the false positive rate on the X-axis. Once again, there is a trade off to be made: a lower false positive rate is desirable, but if you set the threshold so low that you have almost 0 false positive rate, you will fail to diagnose patients that *are* in danger. In other worlds the larger the area under the curve, the better our model predictions are. Conversely, if the false positive rate is very high, limited hospital resources will be wasted on patients who are not actually at risk.

Moreover, therapeutic interventions for patients at risk of COVID-related mortality can be extremely invasive—mechanical ventilation, for example, requires the patient to be sedated unconscious. This introduces additional health risks. So not only does a high false positive rate lead to competition for scarce resources, but it can also worsen patient outcomes.

By calculating the area under the curve, it is possible to obtain a measure of how good the model separability is, i.e. how capable it is in distinguishing patients who are at risk from those who are not. A value of 1 indicates high separability. A value of 0 indicates that the model is classifying the groups completely backwards. A value of 0.5 is the worst; it indicates very poor separability and a high probability of both type I (false positive) and type II (false negative) errors. Fortunately, all the models have an AUC greater than 0.9.

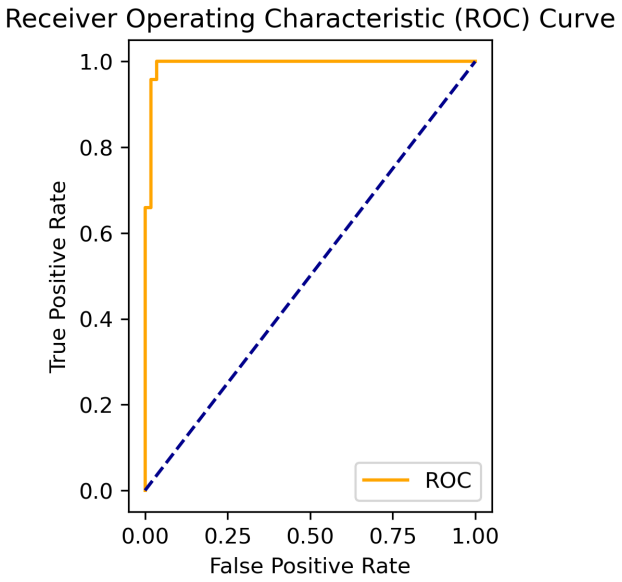

Figure 3: The ROC Curve for the Logistic Regression model

6.3 Clinical Discussion: Do we need a different outcome variable?

A quandary that arises for care givers is whether patients who are at high risk of mortality should receive maximum care, or whether that care is better allocated towards patients who have a higher chance of survival. The ethical questions aside, the technical problem is that we cannot answer this question based on a binary model. What we really need is a third "threshold outcome": we want to identify patients who are at high risk of mortality but which can also benefit from therapeutic interventions. This is not something you can really do with logistic regression, but the other models can make this classification. We would be interested in investigating this problem in a future study.

6.4 Principal Component Analysis (PCA)

PCA analysis shows that about 90% of the information can be acquired by 5 variables. Furthermore, 85% of the explained variance can be deduced from just 3 variables: lactate dehydrogenase, serum chloride, and hypersensitive C-reactive protein. These results are important, as it means hospitals can don't have to perform a wide

variety of (potentially very expensive and time-consuming) tests. Fewer resources spent on biological tests generally means more resources available for care.

An interesting discussion question that arises is: "Why these 3 variables in particular?" Serum chloride is an important electrolyte that is responsible for correct heart function, and given that cardiovascular failure is one of the main cause of COVID-related deaths (see Background section) this is not surprising. High levels of *serum* lactate dehydrogenase indicates that the body has suffered extensive internal trauma, as lactate dehydrogenase is a common enzyme found in many tissues. Finally, C-reactive protein is a marker of inflammation, which is concordant with the previously-discussed role of inflammation in disease progression.

7 CONCLUSION

At the beginning of this paper, we posed two research questions: firstly, can we predict COVID-19 mortality from biomarker data, and secondly, how many biomarkers do we need to make an accurate prediction?

The first research question can be answered "yes". Of all the different classification models tested, Support Vector Machine and Naive Bayes performed the best. Nevertheless, the gap between model performances was small. If we had to choose one, we would choose the SVM model with a non-linear kernel since the results are also easily interpretable. Naive Bayes would be our second choice. We recommend against neural networks as our dataset is too limited to train them sufficiently, and moreover, they are also "black boxes": it is very hard to untangle the chain of neurons as there are so many parameters.

Regarding the second research question, the principal component analysis indicates that 5 variables are sufficient to generate an accurate model, and if the hospital is pressed for time and resources, even 3 biological tests are enough to make a good prediction. The importance of immune system markers has already been highlighted, but this paper also finds evidence for the importance of heart disease markers as predictors of COVID mortality.

For future research, we would investigate how changes over time affect outcomes. A more difficult problem would be to predict time to hospital discharge (either as a result of death or improvement in symptoms) which could help hospitals plan their treatments more effectively. Another avenue of future research is the classification of patients into three groups, with a third "borderline" group indicating which patients should receive maximum treatment.

REFERENCES

[1] Hossam M Ashour, Walid F Elkhatib, Md Rahman, Hatem A Elshabrawy, et al. 2020. Insights into the recent 2019 novel coronavirus (sars-cov-2) in light of past human coronavirus outbreaks. *Pathogens*, 9, 3, 186.

[2] Massimiliano Bonafè, Francesco Prattichizzo, Angelica Giuliani, Gianluca Storci, Jacopo Sabbatinelli, and Fabiola Olivieri. 2020. Inflamm-aging: why older men are the most susceptible to sars-cov-2 complicated outcomes. *Cytokine & Growth Factor Reviews*.

[3] Ashis Kumar Das, Shiba Mishra, and Saji Saraswathy Gopalan. 2020. Predicting covid-19 community mortality risk using machine learning and development of an online prognostic tool. *PeerJ*, 8, e10083.

[4] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 315–323.

[5] Karen Grace-Martin. 2020. Missing data: two big problems with mean imputation. https://www.theanalysisfactor.com/mean-imputation/.

[6] Matthew KH Hong, Henry HI Yao, John S Pedersen, Justin S Peters, Anthony J Costello, Declan G Murphy, Christopher M Hovens, and Niall M Corcoran. 2013. Error rates in a clinical data repository: lessons from the transition to electronic data transfer—a descriptive study. *BMJ open*, 3, 5, e002406.

[7] Imran Khan Mohd Jais, Amelia Ritahani Ismail, and Syed Qamrun Nisa. 2019. Adam optimization algorithm for wide and deep neural network. *Knowl. Eng. Data Sci*, 2, 1, 41–46.

[8] Janus Christian Jakobsen, Christian Gluud, Jørn Wetterslev, and Per Winkel. 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC medical research methodology*, 17, 1, 162.

[9] Wang Kaiming, Gheblawi Mahmoud, and Y Oudit Gavin. [n. d.] Angiotensin converting enzyme 2: a double-edged sword. *Circulationn. d.*

[10] Aaron Lerner, Patricia Jeremias, and Torsten Matthias. 2015. The world incidence and prevalence of autoimmune diseases is increasing. *Int J Celiac Dis*, 3, 4, 151–5.

[11] Peter P Liu, Alice Blet, David Smyth, and Hongliang Li. 2020. The science underlying covid-19: implications for the cardiovascular system. *Circulation*.

[12] Wenjun Ma, Robert E Kahn, and Juergen A Richt. 2009. The pig as a mixing vessel for influenza viruses: human and veterinary implications. *Journal of molecular and genetic medicine: an international journal of biomedical research*, 3, 1, 158.

[13] R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

[14] Hamid Rahmani, Effat Davoudi-Monfared, Anahid Nourian, Hossein Khalili, Nooshin Hajizadeh, Narjes Zarei Jalalabadi, Mohammad Reza Fazeli, Monireh Ghazaeian, and Mir Saeed Yekaninejad. 2020. Interferon β-1b in treatment of severe covid-19: a randomized clinical trial. *International immunopharmacology*, 88, 106903.

[15] Irina Rish et al. 2001. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* number 22. Volume 3, 41–46.

[16] Jennifer R Tisoncik, Marcus J Korth, Cameron P Simmons, Jeremy Farrar, Thomas R Martin, and Michael G Katze. 2012. Into the eye of the cytokine storm. *Microbiology and Molecular Biology Reviews*, 76, 1, 16–32.

[17] Alexandra C Walls, Young-Jun Park, M Alejandra Tortorici, Abigail Wall, Andrew T McGuire, and David Veesler. 2020. Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*.

[18] Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, et al. 2020. An interpretable mortality prediction model for covid-19 patients. *Nature Machine Intelligence*, 1–6.

[19] Feng Zhou, Tao Chen, and Baiying Lei. 2020. Do not forget interaction: predicting fatality of covid-19 patients using logistic regression. *arXiv preprint arXiv:2006.16942*.

[20] Qiong Zhou, Virginia Chen, Casey P Shannon, Xiao-Shan Wei, Xuan Xiang, Xu Wang, Zi-Hao Wang, Scott J Tebbutt, Tobias R Kollmann, and Eleanor N Fish. 2020. Interferon-α2b treatment for covid-19. *Frontiers in Immunology*, 11, 1061.