Mortality Prediction of COVID-19 Patients Using Logistic Regression and Random Forest

Behzad Javaheri

Department of Computer Science, City University of London UK

Introduction: COVID-19 pandemic has affected ~65 million with ~ 1.5 million mortality worldwide [1]. Poor prediction of clinical outcome has primarily been attributed to the lack of patient-level data availability [2]. Reliable prediction of COVID-19 outcome will allow appropriate clinical management and resource allocation to reduce mortality.

Aim: To specify, compare and critically evaluate two machine learning classification methods for accurate prediction of COVID-19 mortality and survival.

Initial data exploration: patient-level data, 39 variables and 2,301,629 entries obtained from GitHub [3] and processed to address quality issues including missing values.

- Predictors/outcome: Relevant health-related variable selected as predictors: age, gender, pregnancy, obesity, diabetes, COPD (chronic obstructive pulmonary disease), asthma, immunosuppression, hypertension, cardiovascular disease, CKD (chronic kidney disease), smoking, other co-morbidities & pneumonia. Mortality/survival considered as the outcome.
- Initial analysis revealed survival of 2,134,851 with the mean age of 40 and mortality of 124,421 with the mean age of 62 years.
- Analysis revealed higher male mortality (78,835) compared to females (45,586; Fig. 1).

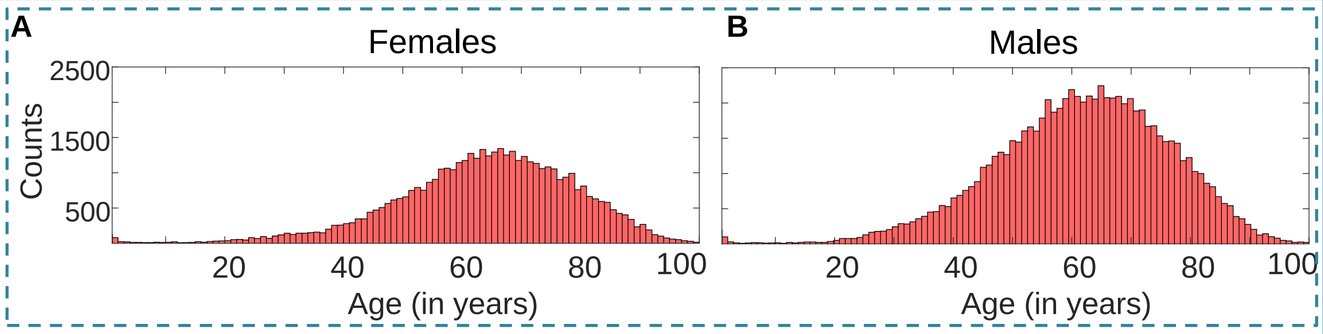

Fig 1. Distribution of mortality of female (A) and male (B) COVID-19 patients across all ages

Machine learning solutions: Supervised classification ML algorithms logistic regression (LR) and random forest (RF) selected as they are suited for binary classification.

1. LR: a white box model, highly interpretable with no parameter tuning required. It calculates regression coefficient and predicts probability of successful outcome (0-1) of resampled (e.g. bootstrapped) training data to assign a class based on a threshold (0.5) [4]. LR results in linear decision boundary and thus prone to higher bias. Successful implementation requires data modification, removal of missing values and predictor selection to reduce error [5].

2. RF: a black box model which uses bootstrap resampling (with replacement) to build large number of trees to estimate classification by majority voting aiming to reduce overfitting [6]. RF can be used when predictors are correlated and does not require data transformation or removal of missing values. Minor limitations related to complexity of implementation, computation time on large datasets and requirement for parameter tuning to achieve higher performance [6].

Hypothesis: That health background of COVID-19 patients allows prediction of mortality and survival with high accuracy by logistic regression and random forest.

1.1 LR initial models:

- Fitglm MATLAB function used to make logistic binomial models on 30 bootstrapped training set (0.8 entire data). Highest performing model selected based on accuracy, precision, recall specificity, F1, time and others (supplementary) and validated using entire training data.
- Analysis revealed that whilst this initial LR model produced 0.92 area under the curve (AUC), 0.97 F1 and predicted survival with 98.6% accuracy, the prediction for mortality was ~27% with low specificity of 0.52 (Fig.2A-C). These findings suggest that optimisations required to improve performance.

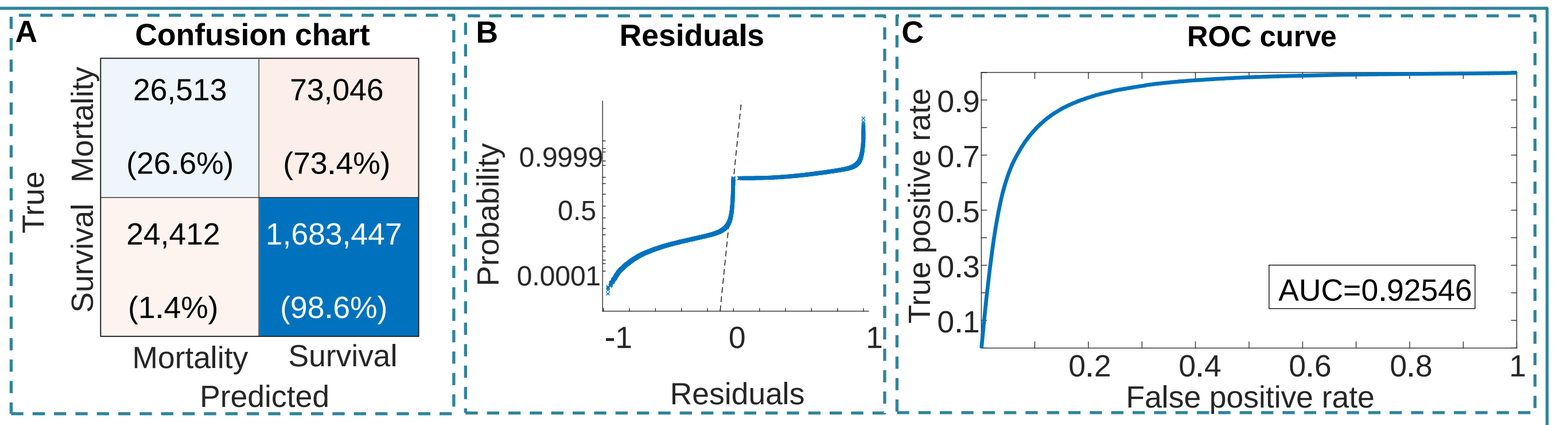

Fig 2. Confusion chart (A), residuals (B) and ROC curve (C) of initial LR model's performance on entire training data

1.2 Optimised LR models:

Data limited to over 30 years of age & split by sex to address age and sex-related imbalance [7]. Under-sampling performed to address class imbalance followed by predictor selection (Lasso).

- Performance of LR models on entire male/female training data revealed improvement in specificity (0.83 & 0.86 males/females respectively) as well as high accuracy & AUC (Fig.3A-C). Other measures (Suppl) include: recall (0.81 & 0.84: males/females), precision (0.84 & 0.87 males/females) & time (0.27 & 0.20s males/females)

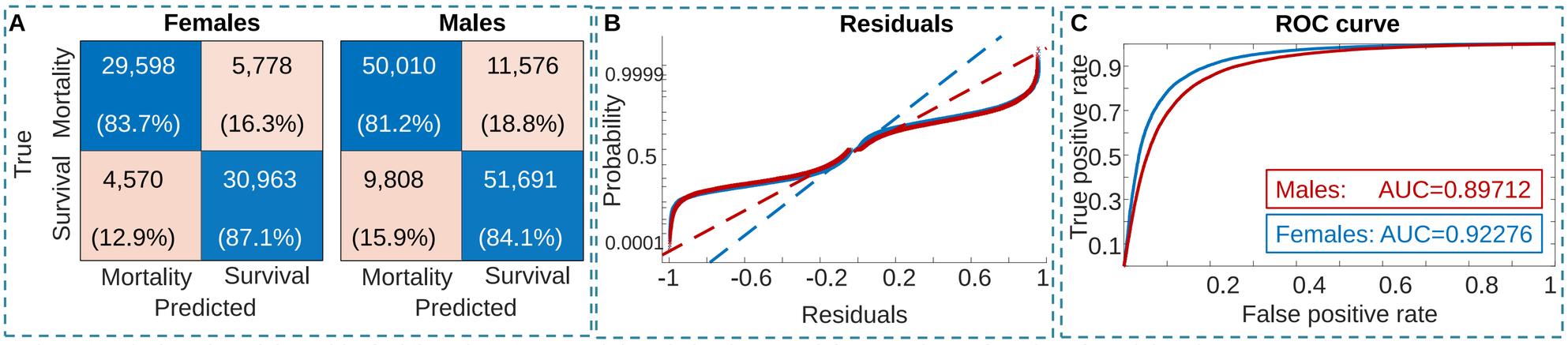

Fig 3. Confusion chart (A), residuals (B) and ROC curve (C) optimised LR models on entire male & female training data

2. Random forest

- A grid search performed for both male and female training data to obtain optimal values for number of leaf, trees, splits & predictors using TreeBagger MATLAB function. These resulted in 625 parameter combinations for each male & female datasets. Based on performance and out-of-bag error, for males 30 trees, 20 leaf, 20 splits and 9 predictors selected. For females 30 trees, 30 leaf, 20 splits and 8 predictors produced highest performance.
- Using these on the entire male & female training data revealed high accuracy and AUC (Fig.4A-B), and others (Suppl) precision (0.80 &0.83 males/females), recall (0.84 & 0.86 males/females), specificity (0.81 and 0.84 males/females).

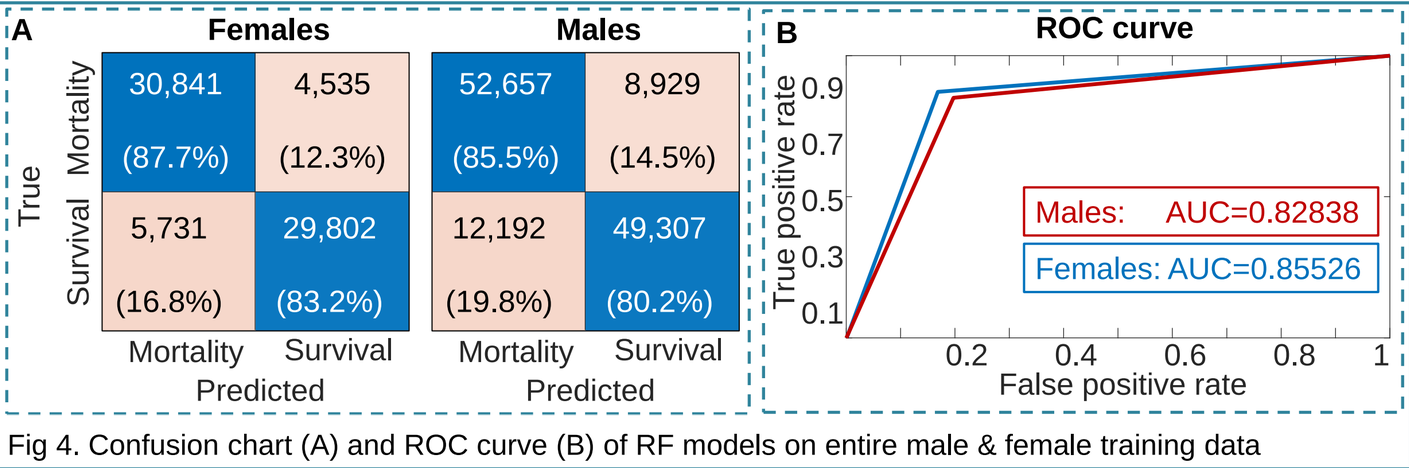

Fig 4. Confusion chart (A) and ROC curve (B) of RF models on entire male & female training data

Analysis & evaluation: Performance of optimised LR & RF models on test data (table1) indicate that optimised LR models produce higher precision & AUC, whilst recall is slightly higher in RF compared to LR; time was lower for LR (0.18 & 0.12 vs. 3.0 & 2.6 s males/females). These indicate COVID-19 outcome prediction with high accuracy based on 8/9 predictors for females/males. For females age, obesity, diabetes, immunosuppression, hypertension, CKD, other comorbidities and pneumonia. For males, in addition smoking is also included.

		Precision	Recall	Specificity	F1	Accuracy	AUC	Time
LR	Male	0.84	0.82	0.83	0.83	82.7	0.89	0.18
	Female	0.87	0.84	0.87	0.86	85.8	0.92	0.12
RF	Male	0.80	0.84	0.81	0.82	82.9	0.82	3.06
	Female	0.84	0.86	0.85	0.85	85.9	0.85	2.67

Table 1. Performance of LR and RF models on male/female testing data

- A previous study reported COVID-19 mortality were associated with age, sex, hypertension, obesity and diabetes. The authors compared LR, RF, support vector machine and XGBoost and found that XGBoost slightly outperformed LR but significantly higher than RF. Whilst in this study, XGBoost was not employed, our findings are in agreement that demographic characteristics including age and sex and health comorbidities allow prediction of mortality with high accuracy. The advantage of the present study, is the number of patients (combined male & females of 162,692) compared to 3,841 [8]. Similarly, Das *et al*., (2020) reported that LR produced higher performance than RF using a small datasets and fewer predictors [9]. The number of patients in this study is similar to a more recent study where inclusion of age, sex and patient's travel history produced mortality prediction with high accuracy (94%) and F1 score of 0.86 [10] suggesting that inclusion of fewer predictors could lead to prediction with high accuracy. This view is further supported by Yadaw *et al*.(2020).
- Taken together, our findings indicate that LR produces better performance than RF in predicting COVID-19 mortality and provides evidence to support our hypothesis in using health background to predict mortality and survival with high accuracy.

Lessons learned & future work: Evaluation of literature suggest that reduction of predictors produces prediction with high accuracy. Whilst in this study lasso regularization performed to select important predictors other methods including ridge regularization or sequential feature selection need to be explored. Furthermore, class imbalance was corrected by undersampling to avoid overfitting [11], other methods including Synthetic Minority Oversampling Technique might produces higher performance. In addition, other ML methods including XGBoost [8] may result in better prediction and thus to be further explored.

References: **[1]**. M. Roser et al., "Coronavirus pandemic (COVID-19)," *Our World in Data,* 2020. **[2]**. C. V. Cosgriff et al., "Data sharing in the era of COVID-19," *The Lancet Digital Health,* vol. 2, no. 5, p. e224, 2020. **[3]**. https://github.com/alberto-mateos-mo/covid19mx accessed on 01/Nov/2020. **[4]**. P. Ranganathan *et al*., "Common pitfalls in statistical analysis: logistic regression," *Perspectives in clinical research,* vol. 8, no. 3, p. 148, 2017. **[5]**. K. Kirasich et al., "Random Forest vs Logistic Regression" *SMU Data Science Review,* vol. 1, no. 3, p. 9, 2018. **[6]**. L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001. **[7]**. E. Barron *et al.*, "Associations of type 1 and type 2 diabetes with COVID-19-related mortality in England: a whole-population study," *The lancet Diabetes & endocrinology,* vol. 8, no. 10, pp. 813-822, 2020. **[8]**. A. S. Yadaw et al., "Clinical features of COVID-19 mortality: development and validation of a clinical prediction model," *The Lancet Digital Health,* vol. 2, no. 10, pp. e516-e525, 2020. **[9]**. A. K. Das *et al.*," *PeerJ,* vol. 8, p. e10083, 2020. **[10]**. C. Iwendi *et al.*, "COVID-19 Patient health prediction using boosted random forest algorithm," *Frontiers in public health,* vol. 8, p. 357, 2020. **[11]**. J. Wang *et al.*, *PeerJ,* vol. 8, p. e9945, 202.