OCMP5310 Project Stage 2 Report

Title: Predicting ventilator requirements for COVID-19

Author: *Timothy Creer*

Date: *28th March 2025*

Section 1: Problem

1.1 Overview:

COVID-19 which was caused by SARS-CoV-2 sparked a global health crisis that by February 2025 had infected millions and claimed over 7 million lives (World Health Organisation, 2025). Its rapid transmission placed immense strain on healthcare systems, particularly due to critical shortages of ventilators which are essential for managing severe respiratory failure. These shortages led to difficult triage decisions, excess mortality, and overwhelmed intensive care units (Fauci, Lane & Redfield, 2020; MacMillan, 2020). Individuals with underlying conditions such as diabetes, hypertension, or immunosuppression were at significantly greater risk. Early prediction of ventilator needs could enable timely interventions, optimise ICU capacity, guide resource allocation, and save lives. This project investigates whether it is possible to accurately forecast intubation requirements based on a patient's medical history and initial symptoms. Accurate predictions would allow hospitals to proactively manage ICU loads, public health agencies to plan stockpiles, manufacturers to anticipate demand, and vulnerable patients to receive prioritised care. This work aims to strengthen healthcare resilience and support evidence-based strategies for future pandemic responses.

1.2 Research Question:

Can we predict ventilator requirements for COVID-19 patients based on their medical history & initial symptoms?

1.3 Target Attribute and Success Measures

The key attribute predicted in this project is INTUBATED, a binary variable indicating whether a COVID-19 patient required mechanical ventilation. This outcome is critical for early intervention in preventing death in severe cases and efficient ventilator allocation to help alleviate hospitals during infection surges. To evaluate model performance, three metrics will be used: F1-score, recall, and precision as each are selected for their clinical relevance in high-stakes, imbalanced datasets. The F1-score provides a balanced summary of precision and recall, ensuring that both false positives and false negatives are considered. However, reporting recall and precision separately is also essential. Recall (sensitivity) ensures the model identifies most patients who truly require intubation, reducing the risk of missed critical cases. Precision on the other hand, confirms that those flagged as needing ventilation are genuinely high-risk preventing unnecessary interventions that may strain ICU resources.Although F1-score is useful for overall model comparison, recall and precision offer transparency into trade-offs, which is vital in a clinical setting. Together these metrics provide a detailed, interpretable, and ethically aligned evaluation that maximises life saving potential while ensuring responsible use of limited medical resources.

Section 2: Predictive Model

2.1 Model Description

For this project, the predictive modelling technique selected is Extreme Gradient Boosting (XGBoost), a high-performance ensemble learning method based on decision trees. XGboost was released in 2014 and since then has undergone 3 major iterations. It assumes that complex, non-linear relationships in structured data can be captured through the additive combination of weak learners with regularisation which helps to manage overfitting (Chen and Guestrin, 2016; Nielsen, 2016). This assumption aligns well with the COVID-19 dataset used in this study as it includes a mix of categorical risk factors, medical conditions, and patient demographics. XGBoost's ability to natively handle missing values and its compatibility with techniques like SMOTE makes it very suitable for the class imbalance in intubation outcomes. It also scales efficiently to the large sample size and high dimensionality present in the dataset. On the contrary, as logistic regression may be simpler, it fails to model non-linear interactions, and while random forests are robust they lack XGBoost's fine-grained regularisation and boosting

mechanism. Limitations in XGBoost include sensitivity to hyperparameters and potential overfitting if tuning isn't properly carried out. However, the integration of cross-validation and Optuna minimise these risks. Given the clinical stakes of predicting ventilator requirements, XGBoost's balance of predictive power, flexibility, and resilience to noisy, incomplete data highlights how it is one of the most suitable algorithms for this task.

2.2 Model Algorithm

The XGBoost algorithm operates within a gradient boosting framework that builds an ensemble of decision trees sequentially, correcting residual errors at each stage. It optimises a regularised objective function using both first- and second-order derivatives, a process known as Newton boosting (Chen and Guestrin, 2016; Nielsen, 2016). This second-order approximation accelerates convergence and improves robustness, particularly in complex or noisy datasets. The algorithm begins with baseline predictions, then iteratively fits trees to the gradients of the loss function. Each tree's structure and leaf weights are chosen through greedy optimisation to minimise loss. XGBoost assumes that additive models can effectively approximate the true outcome function, and uses regularisation to prevent overfitting. Key hyperparameters include max_depth, min_child_weight, and gamma (which control tree complexity), as well as eta, subsample, and colsample_bytree to manage learning rate and variance. Regularisation terms lambda and alpha help mitigate model complexity - critical in the imbalanced COVID-19 dataset used in this project.

Recent advancements such as those in XGBoost v3.0.0 have introduced GPU acceleration, native categorical handling, and improved early stopping (XGBoost, 2024). These make the model more scalable and responsive in time-sensitive healthcare scenarios. Integration with SMOTE and automated hyperparameter tuning frameworks like Optuna further enhances performance and generalisability. These qualities affirm XGBoost's continued dominance and suitability for high-stakes prediction tasks as its the most used algorithm in winning Kaggle models (Reddy, 2024). Its continued efficiency, customisability and proven track record reaffirm its applicability for this dataset to predict ventilation requirements.

2.3 Model Development

To develop the best model for intubation classification, a layered approach was used to systematically evaluate both feature engineering and model tuning strategies. Several dataset variants were created. One removed features with low correlation to the target (INTUBATED, <0.05), aiming to reduce noise. Another introduced COMORBIDITY_COUNT, a single feature aggregating the presence of comorbidities to reduce dimensionality and reflect overall health burden. Building on this, a SEVERITY_SCORE was computed by summing pneumonia, ICU admission, and comorbidity count to capture a patient's critical condition more effectively. A final version applied PCA after scaling with StandardScaler, reducing multicollinearity and compressing patterns into five components (Jolliffe and Cadima, 2016).

Each dataset was evaluated using XGBoost under four configurations: no tuning, SMOTE only, cross-validation (CV) only, and SMOTE+CV. This design isolated the individual and combined effects of resampling and validation strategies, crucial for handling the class imbalance (~21% intubated). Stratified 5-fold CV ensured the positive class was consistently represented, improving generalisation (Kuhn and Johnson, 2013). Following initial screening, each variant was passed through Optuna, a hyperparameter optimisation framework. It explored a wide range of tuning parameters (e.g., max_depth, eta, lambda, colsample_bytree, scale_pos_weight) with XGBClassifier to balance complexity and performance (Akiba et al., 2019). Objective functions were carefully defined using train_test_split, DMatrix, and early stopping for runtime efficiency and fairness.

This structured experimentation from feature transformation to automated tuning by mitigated overfitting, managed imbalance, and led to consistent, high-performing models aligned with the clinical need for reliability and interpretability.

3.1 Model Evaluation

Initial aims of model evaluation were to determine which XGBoost tuning configuration would be selected for deeper Optuna hyperparameter optimisation. Model performance was evaluated using F1-score, recall, and precision to reflect clinical priorities where early identification of critical cases is essential in saving lives for COVID. Initial testing showed models incorporating SMOTE consistently achieved the highest recall (mean: 0.5769 with CV), demonstrating the importance of oversampling in mitigating class imbalance (~21% intubated). This elevated sensitivity supports early intervention and prioritisation of patients at risk of respiratory failure. However, precision declined with SMOTE from 0.6403 (no SMOTE) to 0.3189 (SMOTE+CV) indicating an increase in false positives. However, the trade-off is acceptable in a medical context where missing a critical case is more severe than unnecessary intervention. F1-score peaked at 0.4170 in the SMOTE-only configuration, but the SMOTE+CV model (F1 = 0.4108) provided comparable performance with improved generalisation. This balance between sensitivity and robustness indicates SMOTE+CV as the optimal choice for further hyperparameter tuning.

The strongest performing models were built on feature-engineered datasets, particularly the SeverityScore and Comorbidity datasets which incorporated clinically significant variables. These datasets achieved higher recall and F1-scores under SMOTE+CV, indicating that the model was learning from medically relevant risk factors. However, while these results demonstrate promise the absolute scores remain adequate, indicating the model still requires further improvement. In hospital settings, higher precision and recall would be essential to ensure both patient safety and trust in predictions. This highlights the need for further refinement through threshold tuning, cost-sensitive learning, or maybe the data is just not rich enough.

3.2 Model Optimisation

To optimise XGBoost performance the advanced hyperparameter tuning framework Optuna was employed due to its superior efficiency compared to traditional grid and random search methods (Shahrour, 2022; Saad, 2023). Optuna uses Tree-structured Parzen Estimators (TPE) and pruning strategies to intelligently explore hyperparameter space, reducing computation time while converging on the most efficient configurations. Key hyperparameters tuned included eta (learning rate), max_depth, min_child_weight, gamma, and scale_pos_weight. These were selected to balance bias to variance trade-off, control tree complexity, and account for class imbalance (~21% intubated). This was evidenced when using scale_pos_weight which was optimised to correct the skewed class distribution, and eta was explored on a logarithmic scale to capture learning sensitivity.

Three experimental setups were tested: no validation/tuning, SMOTE-only, and SMOTE with stratified cross-validation (CV). The objective function maximised F1-score across five folds, with early stopping and Optuna's MedianPruner preventing unnecessary trials. Tuning revealed that max_depth and colsample_bytree were consistently impactful, showing that optimal tree structure and feature sampling are critical for performance. SMOTE+CV emerged as the most robust configuration despite slightly lower precision, achieving the best balance of generalisability and sensitivity. This reflects a deliberate trade-off favouring recall in high-risk clinical prediction, while acknowledging the need for further post-tuning calibration. Overall, Optuna provided an efficient, flexible, and clinically aligned optimisation approach well-suited to the complexity of ventilator prediction.

---

4. Discussion

A comprehensive evaluation of XGBoost models across multiple configurations revealed both promising outcomes and areas for improvement. Quantitatively, the top-performing model by F1-score (0.4275) was trained on the original dataset without SMOTE or cross-validation. The original, unmodified dataset struck the best balance between recall and precision because it relied on high-quality, real clinical features without introducing synthetic samples or variance from cross-validation. While not the most generalisable model, it achieved the best F1 in this specific test scenario. Key clinical features such as ICU admission, pneumonia, and comorbidity count enabled effective discrimination between intubated and non-intubated cases. Its recall of 0.5897 suggests moderate sensitivity, capturing a majority of patients who required ventilation, though precision (0.3353) indicates a notable rate of false positives. In comparison, the SMOTE+CV configuration yielded the highest recall (0.6896) but suffered reduced precision (0.2971), reinforcing the trade-off between sensitivity and alert fatigue. Qualitatively, XGBoost was well-suited to the structured nature of the dataset, handling missing values and modelling complex feature interactions. ROC AUC values (~0.70) across models suggest consistent overall discrimination. However, precision-recall curves demonstrated that as the model pushed for higher recall, precision steadily declined. This behaviour indicates the model becomes less selective and increasingly prone to flagging false positives as it attempts to capture all true positives which is an expected pattern when classifying less separable and borderline cases.

Despite some strengths it is not without its limitations. Reduced precision increases the likelihood of false alarms, which could potentially take up valuable ventilators when not needed. Furthermore, while the models rank patients by risk relatively effectively, they lack inherent interpretability which poses a barrier to clinical trust and adoption without the integration of explanation techniques. The underperformance on minority class metrics also suggests that improvements are still needed. However, the XGBoost models demonstrated the ability to identify at-risk patients using relatively adequate data but still supporting the feasibility of early intervention tools. In addition to refining thresholds and leveraging cost-sensitive learning, further focus should be placed on the quality and completeness of clinical data. External validation across diverse healthcare settings would be crucial for ensuring broader applicability and accuracy.

---

4. Conclusion

This project demonstrated the relative potential of predicting ventilator requirements for COVID-19 patients using XGBoost trained on medical history and symptom data. Among the tested models, those incorporating class balancing and clinically meaningful features achieved the best results. While the original dataset without SMOTE or CV showed the highest F1-score (0.4275), the SMOTE+CV configuration achieved the strongest recall (0.6896), a clinically prioritised metric. This emphasises the importance of recall driven optimisation in life dependent outcomes, even at the cost of precision. However, absolutely the metrics are not high enough and the models are not yet deployment-ready. Low precision remains a concern as it highlights the potential for false positives and overburdened ICU resources. The findings also suggest that data quality and feature granularity may be limiting factors. Future endeavors should focus on expanding the dataset with richer and cleaner clinical data. Model refinement should also explore cost-sensitive learning, further calibrated thresholds, and ensemble methods to boost performance and stability.

In conclusion, this study offers an interesting case for AI-assisted ventilator triage laying a solid foundation for scalable, data-driven response strategies in future pandemics. With improved data collection, advanced modelling, and external validation, this work could be leveraged in future pandemics or current triage of medical symptoms.

References

- Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2623-2631).
- Chen, T. and Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
- Fauci, A.S., Lane, H.C. and Redfield, R.R., 2020. Covid-19—navigating the uncharted. New England Journal of Medicine, 382(13), pp.1268-1269.
- Jolliffe, I.T. and Cadima, J., 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), p.20150202.
- Kuhn, M. and Johnson, K., 2013. Applied predictive modeling. Springer.
- MacMillan, C., 2020. The ventilator shortage is a symptom of a bigger problem. Scientific American. Available at: https://www.scientificamerican.com
- Nielsen, D., 2016. Tree boosting with XGBoost – why does XGBoost win every machine learning competition?. NTNU. Available at: https://arxiv.org/abs/1603.02754
- Reddy, N.L., 2024. XGBoost: The ever-relevant champion in the age of generative AI. Medium. Available at: https://medium.com/@nandinilreddy/xgboost-the-ever-relevant-champion-in-the-age-of-generative-ai-2146311fdfc6
- Saad, I.E.M., 2023. Optimizing machine learning models: Comparing grid search, randomized search, and Optuna. DEV Community. Available at: https://dev.to/iemsaad/optimizing-machine-learning-models-comparing-grid-search-randomized-search-and-optuna-4hnf
- Shahrour, A., 2022. Optuna vs GridSearch. Medium. Available at: https://medium.com/@abdalrahman_shahrour/optuna-vs-gridsearch-57227556c450
- World Health Organisation, 2025. Coronavirus disease (COVID-19) pandemic. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019
- XGBoost, 2024. XGBoost v3.0.0 Release Notes. Available at: https://xgboost.readthedocs.io/en/latest/changes/v3.0.0.html