# Introduction

The Coronavirus disease 2019  (COVID-19) has emerged as a worldwide pandemic over the last few months. The clinical spectrum of COVID-19 infection varies from asymptomatic to a severe clinical condition characterized by respiratory failure necessitating mechanical ventilation and support in an intensive care unit (ICU). To date, as infection rates continue to rise worldwide, no established method for predicting the risk for deteriorating to a critical state in COVID19 patients exists.

At the peak of the pandemic, The Sheba ARC Innovation Center has organized an international data challenge, offering data-science teams around the world access to hospitalized  COVID19 data, with the purpose of supporting medical experts in making decisions about their patients by predicting clinical deterioration and detecting non-trivial underlying patterns in the clinical data. This in turn will allow a more fine grained understanding of the dynamics of the disease, as well as give a tool for hospitals in managing and predicting patient load.

As at that time data was scarce, the disease course was unclear and clear clinical definitions for patient deterioration were not set, a novel approach for developing a clinically-usable prediction model was adopted by MAFAT, (MOD DDR&D) Israel's Ministry of Defense department for technological innovations. 5 data-science teams, formed from commercial companies and academics specializing in AI in the private sector and, has assembled by MAFAT in order to work in collaboration on the same problem - each determining its own approach and methods according to its strengths, while working closely under the guidance of Sheba clinicians.

This resulted in 5 different models, each with its strengths and weaknesses. As more and more data continues to unravel, and clinical needs are becoming clearer, clinicians will be able to choose the model most applicable for the clinical setting. Additionally, comparing the different models and analyzing their differences could potentially assist in the understanding of the disease course and treatment alternatives. One of the most significant outcomes is an AI driven pipeline that has the potential to be embedded as a support tool for COVID 19 clinical related procedures.

This report summarizes the data provided by Sheba ARC, the different models designed by the different teams and their results. As each team used different definitions and models, and subsequently received different results and conclusions, each following section will present a brief summary of each team's approach, while a detailed report can be found in the Appendix.

# Methods

## **Data description**

Sheba ARC Innovation provided access to information for 426 patients admitted to Sheba Medical Center COVID wards, including all patients with positive PCR until May 21, 2020. The data included demographic information, lab tests, medications, background diseases, diagnoses, hospitalizations, nursing actions, and more. The data is recorded for each patient from ER admission until discharge or unfortunate death. Some of the values, such as lab results, are taken multiple times during the course of disease. Therefore, each data point is provided along with a timestamp. All teams recognized 2 different types of data - 1. Static information, including demographic data, medical background, chronic medications, etc. 2. Time series information - includes measurements like HR, blood count, etc, taken at multiple times throughout the hospitalization. The 2 data types were treated differently during data preprocessing, as will be described.

## **Target definition**

The main research question, as defined by Sheba ARC's clinical team, was 'can we predict COVID 19 patient clinical deterioration'. As patient deterioration is a vague medical term, without a solid clinical definition, each team defined 'deterioration' differently. As can be seen in table 1, 2 teams focused on deterioration to Invasive Mechanical Ventilation (IMV), while the remaining 3 teams defined 'deterioration' in a broader clinical term, including deteriorations in c-EWS score, ICU admission, administration of sedation and paralysis agents, administration of inotropic and vasopressor drugs, deterioration in oxygen saturation levels,  and death.

Table 1 - target definition

| | |
|---|---|
| | 1. Invasive Mechanical Ventilation (IMV)<br>2. Death<br>3. O2-saturation value below 93% |
| Matrix | 1. Invasive Mechanical Ventilation (IMV) |
| Technion | 1. Artificial respiration2.<br>2. ICU admission,<br>3. administration of sedation and paralysis agents<br>4. administration of inotropic and vasopressor drugs<br>5. death |
| TSG | 1. Invasive Mechanical Ventilation (IMV) |
| BeyondMinds | 1. Invasive Mechanical Ventilation (IMV)<br>2. High c-EWS (covid adapted) score<br>3. Deterioration of c-EWS score |

Time windows

A crucial decision for the usability of a model is the time point during admission at which it is designed to be used for risk stratification by clinicians. 2 different approaches were taken by the different teams - the first is designing a model aimed for use during the early stages of hospitalization, and the second is designing a model that will predict deterioration throughout the entire admission duration.

For all teams, static information, such as  demographic data, medical background, etc. was calculated once for every patient and was constant in the analysis. The dynamic data, such as lab results and vital signs measurements, was calculated in the designated time window for every team's model based on the designated approach, as described in table 2.

Table 2 - time windows

| Team | Time window |
|---|---|
| TSG | Once at every hour, the algorithm defines a gap of several hours, and predicts IMV onset during a four hours window after the gap. several gaps were - 6, 12, 18, 24 hours |
| BeydonMinds | First 24  hours of admission |
| Basis | 1. Time slots at the beginning of hospitalization (12-96 hours)<br>2. 0-5 hours look back windows |
| Matrix | Time slots at the beginning of hospitalization (24-72 hours) |
| Technion | All available data starting 6 hours after admission |

Feature Extraction

Patient data included demographic information, lab tests, medications, background diseases, diagnoses, hospitalizations, nursing actions, and more. As the provided information was incredibly rich, and included information that may not be crucial for prediction, each team used a different subset of the available data. For feature engineering, all groups treated static and time-series features in a different manner, as previously mentioned. Only one group extracted information from textual data. The different feature subsets and feature extraction methods used is summarized in table 4.

Table 4 - feature extraction

| Team | Major selected features | Static Data | Time sequence data |
|---|---|---|---|
| TSG | Demographics, chronic conditions, labs, vital signs easutments | Structured data | Per window dataframe based on hourly data from last 6 hours |

| | | | |
|---|---|---|---|
| Matrix | Demographics, chronic conditions, labs, vital signs easutments | Structured data | One hot encoding for categorical features, 5 time dependant measurements for time sequence data(mean, max, trend, first, last) |
| BeyondMind | Demographics, chronic conditions, labs, vital signs easutments, free text | 200 features were extracted from the free text text fields | Described by basic statistics such as mean, std, median in each window |
| Basis | Lab tests and measurements | Structured data | the latest result of every relevant marker. |
| Technion | Demographics, chronic conditions, labs, vital signs easutments | Structured data | Survival analysis |

Model design

The teams used different model designs and algorithms based on the definitions described in previous sections. 3 teams used classic machine learning algorithms on data available at the relevant time window, 1 team used a transfer learning approach based on the MIMIC dataset, and 1 team used a survival-analysis approach. Model designs are stated in table 5

Table 5 - model designs

| Team | Model design | Algorithms |
|---|---|---|
| TSG | Transfer learning in 3 stages - Predict onset of IMV on pre-COVID-19 data (MIMIC-III), Data matching Sheba-MIMIC, domain shifting algorithm | XGBoost |
| BeyondMinds | Comparison between 4 models based on 4 model designs :<br>1. Hard labeling - Stratified K Folds<br>2. Softlabeling<br>3. SMOTE - Hard labeling - Stratified K Folds<br>4. SMOTE - Soft labeling | XGBoost |

| | | |
|---|---|---|
| Technion | Competing risks survival analysis for 2 events - release from the hospital or moving to a critical state. | L1-regularized<br>Cox regression<br>Random Survival Forests |
| Basis | 1. Predicting patients’ probability to become high risk based on data available at early admission<br>2. Predicting Onsets of o2-saturation<93% in look back windows | XGBoost |
| Matrix | Predicting patients’ probability to deteriorate to mechanical ventilation based on data available at early admission | XGBoost |

# Results

Model performance

Given the different definitions and models, results varied significantly between the teams. Results for the most important variables for the best performing models are presented in table 6. Highest prediction accuracy was achieved by teams predicting specifically IMV, using the XGBoost implementations. AUC for the different models ranged from 0.96 to 1. The competing risks analysis for the ICU or deaths events all achieved a very high concordance score, which could be paralleled to AUC, of 0.86-0.89. The event demonstrating the least predictive capabilities was o2 deterioration, achieving predictive AUC of 0.725-0.807.

Table 6 - model performance

| **Team** | **Predicted Variable** | **Result** |
|---|---|---|
| Technion | Competing risks - ICU or death | Concordance score:<br>Cox: 0.89, RSF: 0.86 |
| Basis | Ventilation or death | Accuracy: 62-78 depending on time slot |
| | O2-saturation<93% | ROC AUC 72.5 - 80.7 |

| | | |
|---|---|---|
| TSG | IMV | ROC AUC 0.96-0.98 |
| BeyondMinds | IMV | AUC 0.95-1 |
| | High c-EWS score | AUC 0.95-1 |
| | Deterioration of c-EWS score | AUC 0.98-1 |
| Matrix | IMV | AUC 0.96 - 0.98 |

Feature importance

The algorithms used by all teams, including XGBoost, Random Survival Forest and Cox all provide information about the importance of the features in the model. In the XGBoost model and Random survival Forest The importance value of each marker represents its mutual information with the response variable, measured over the forest of trees that are built as part of the training procedure. The most important features for the best model for every time team are presented in table 7.

Table 7 - Top significant factors

| **Team** | **Important features** |
|---|---|
| technion | psychiatric disorders, diabetes, obesity, autoimmune disease, hypertension,blood pressure; diuretics; blood substitutes and perfusion solutions. |
| TSG | Creatinine urine, CPK, FiO2, Arterial base excess, Calcium ionized, PTT, Albumin, CVP, Partial pressure of oxygen, CO2, Heart rate |
| BeyondMinds | O2 saturation, Blood pressure, CRP, respiration rate, RBC |
| Matrix | O2 saturation, oxyhemoglobin, lymphocytes, RBC |
| Basis | Protein levels, neutrophils, blood pressure, urea, CRP, calcium |

# Discussion

This work was a collaborative effort organized by MAFAT as part of the Sheba ARC data challenge, aimed at  predicting the future severity of COVID19 patients using their medical data and providing clinicians valuable tools in treating COVID patients. In this report we present the results of the 5 teams organised by MAFAT. All teams could predict with good accuracy which patients will deteriorate, as defined individually by each team, given age, background conditions, lab results and vital signs measurement at different time windows during the hospitalization. One team could indicate the how long a patient is expected to take until they are discharged or deteriorate, enabling better planning of patient load and possibly patient care.

The main difference between this work and many related work trying to predict severity in COVID 19 patients is that this work describes the collaborative work of several teams, working towards mutual goals in different approaches. By using this approach, we could simultaneously investigate various solutions for a clinical problem, and provide the clinical team with different models with different applications. Roughly, our models could be divided to model aimed for use at the early stages of hospitalization (Matrix, Basis, BeyondMinds), and model that monitor patients in real-time throughout the hospitalization (Technion, TSG) and offer clinicians continuous decision support.

The different approaches resulted in significant differences between the teams, with definitions of the predicted target, feature engineering and model design varying significantly. One key difference between the teams was training the model on a fixed time window at the start of the admission, vs.  training on the entire time frame. Surprisingly, results were better with the early-prediction models. This could be due to the fact that the most significant information was already available at admission, and the windows were wider (24-72 hours vs. only a few hours)

Another key difference was in the predicted target definition. While some teams chose to define clinical deterioration as a narrow term restricted to mechanical ventilation only, others chose a broader term including other clinical markers of morbidity. Notably, Results were superior for all groups when predicting mechanical ventilation alone. This could be due to the fact that IMV was a common pathway to all medical complications for patients, and that, at least at the earlier stages of the pandemic, IMV was introduced to patients in a more liberal manner.

All teams faced technical difficulties. The most prominent difficulty, affecting all teams, is the lack of data. Given less than 400 patients, and only a very limited amount of positive samples, and given the depth of available data, applying machine learning algorithms was very challenging. As the data was only available for a limited time, many technical improvements were not introduced due to time limitations, for example improved missing data imputation, balancing techniques for minority class, experimenting with other defined targets, etc.

Individual teams faced more specific difficulties - the Technion team faced difficulty of evaluating ‘zero time’, as patients reached the ED and wards in different stages in their disease, while the TSG team conducting transfer learning from the MIMIC-III dataset had many relevant clinical data missing in the Sheba data.

A possible data leakage is the fact that some labs and measurements were taken differently for patients, as more ‘alarming’ patients were taken more tests, and had richer data.

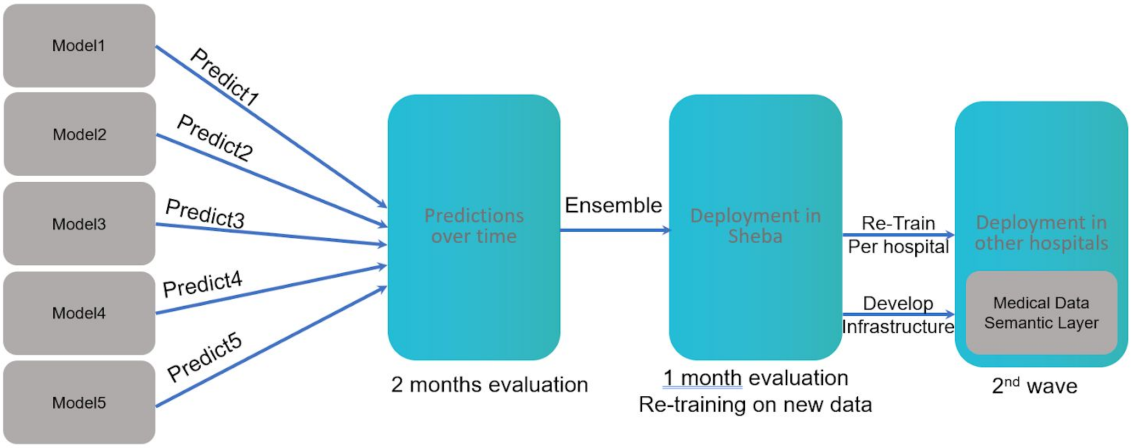
In conclusion, we have demonstrated a multi-team approach for predicting future deterioration in COVID-19 patients. Different approaches were taken by the different teams, allowing several perspectives on the same problem and demonstrating the difference in accuracy between the models.

# Future Work and Call to Action

In this work, multiple teams collaborated in a mutual effort to provide clinicians a valuable data-driven tool for treating COVID 19 patients. As the results are promising, this research is only the tip of the iceberg, and much work is yet to be done. We propose the following steps, in order to utilize our research and improve the quality of future works.

1. Hospital Implementation

Herby is a suggested architecture for a mid-term implementation and deployment of the work that has been done:



Explanations:

1. After a short alignment between the groups on the target and time interval of the prediction, the models developed should be deployed on Sheba's cloud.
2. Each day, there will be a manual run of inferences model for each of the 5 models presented in this summary.
3. The predictions (on real data) should be tested over time separately and as an ensemble model (with simple average/majority of the predictions in the beginning).  We believe a 2

month is appropriate time to evaluate the models (5 separately and ensemble one) accuracy.

4. Up to this point, doctors will get the ensemble prediction once a day for each patient, as an offline (not connected to the EMRs) report.
5. After 2 months of assessment, with some more improvements due to new labeled data, there will be an API to connect the EMRs, and operate on an hourly basis to predict patients' deterioration.
6. New hospitals can use this ensemble model (offline usage) with small adjustments by creating a generic semantic layer of medical data that is injected to the model.

### 2. Aggregate data from different hospitals and unify data structure

All teams unanimously agreed that the most significant difficulty in this work was the lack of the data. With only ~400 patients admitted to sheba and ~10% complicated patients, the ability to develop and test powerful machine-learning models was limited. A powerful solution for this problem is the use of data coming from other hospitals in Israel. A multi-center study, with hundreds of additional patients, could dramatically improve results, predictability and usability of our models. The unique structure of the Israeli health system is a unique opportunity for multicenter collaboration on a global scale, and this crisis could be a significant stepping stone for such a national-scale collaboration.

Other than the lack of data, the variability in the preprocessing of the data is a significant difficulty when trying to integrate results from different teams. A single, well established and researched data structure would decrease development time, improve interoperability, improve results and allow large scale collaboration. When considering a national-scale collaboration as previously mentioned, unifying data structure across hospitals in Israel gains even greater importance.

### 3. Establish a national, non-commercial supervising authority

Lastly, this work was only possible due to diverse and cross-industry relations and collaborations of MAFAT. MAFAT as a unifying authority allowed this collaboration, the sharing of knowledge and resources, and peer-feedback. The large scale needs previously mentioned could only be answered by a national, non-commercial authority, with the ability to allocate resources needed for such high scale and long term steps needed to improve the national response to a significant health crisis.

In conclusion, our recommendation are:

1. The ability to collaborate and share data between hospitals in Israel is critical. In further steps collaborate data from hospitals across the world.
2. A unified, well established, medical data structure is critical for future AI processes.
3. To establish an actionable effective offline pipeline to support COVID-19 clinical procedures.