S1 Protocol. Data Cleanup In a small number of cases laboratory parameters are not specified with absolute numeric values, but rather by denoting that the value fell above or below a specific threshold (based on guidelines from each institution). To obtain numeric values we converted these using the protocol shown in Supporting Information S1 Table. Any values of Potassium, Lactate Dehydrogenase, and Magnesium that were considered as an outlier by the institutions were removed.

S1 Table. Data Cleanup Data cleanup protocol thresholds and replacement values are provided in Table 4.

Table 4. Data cleanup protocol thresholds and replacement values.

| Parameter               | Specified as    | Converted to    |
|-------------------------|-----------------|-----------------|
| C-reactive protein      | < 3 mg/L        | 2 mg/L          |
| D-Dimer                 | > 4 mg/L        | 4.01 mg/L       |
| D-Dimer                 | < 0.17 mg/L     | 0.16 mg/L       |
| Gamma-Glutamyl Transferase | < 5 U/L       | 5 U/L           |
| Procalcitonin           | < 0.03 ng/ml    | 0.01 ng/ml      |
| Procalcitonin           | > 75 ng/ml      | 76 ng/ml        |

Value ranges of the 27 laboratory test results and chest x-ray score, separated by institute and by positive/negative RT-PCR test results. JHH stands for Jeroen Bosch hospital and this data is used as the test dataset. BHH stands for Bihorel Hospital and the data from this institution is used as the training dataset in our study. Positive cases are indicated by 1.0 and Negative by 0.0. It should be noted that the chest x-ray scoring model was trained using data from JHH leading to a difference in the scores produced between institutes.
**S1 Fig. Data Distribution** The input data distribution per hospital and per RT-PCR test outcome is provided.

**S1 Text. Training Settings**

**RT-PCR test prediction from CXRs**

**Training settings**

We train Resnet-18 using a cyclic learning rate \([1]\) between 0.001 and 0.01 and 2.5 epoch step size. We use stochastic gradient descent with Nesterov momentum of 0.95. We minimize the cross-entropy loss between the softmax activations and binary labels. We perform all experiments with a batch size of 16.

When fine-tuning the last layer on RT-PCR test results, we use a heavy label smoothing regularization \([2]\) of 0.2 to prevent overfitting. We apply a weight of 0.625 to the positive samples, to account for the class imbalance. This number is calculated by dividing the number of negative cases in the training dataset by the number of positive cases. After training, we restore the model weights that have achieved the best validation cross-entropy loss.

**Image preprocessing**

We convert the DICOM files to 8-bit PNG files by clipping the values above the 99th percentile and scaling the resulting values between 0 and 255. We resize the resulting images to 512 by 512. Before we feed those images to the convolutional neural network, we apply data augmentation to prevent overfitting. We augment the images by cropping a width and height in the range of \([409, 512]\) randomly. We resize the resulting image to 448 by 448. We randomly scale (by \([0.75, 1.25]\)) and shift (\(\pm 64\)) the pixel values and clip the values above 255. We randomly flip the image from left to right. We standardize the resulting pixel values using the ImageNet dataset means and variances.
Model Selection: Validation dataset experiments

Training settings

Similar to the previous experiment, we train DAE, SDAE, FCN, and DFCN using a cyclic learning rate \([1]\) between 0.001 and 0.01 and 2.5 epoch step size. We use stochastic gradient descent with Nesterov momentum of 0.95. We minimize the cross-entropy loss between the softmax activations and binary labels. For all models we use ReLU activations for intermediate layers. We train our models until there is no improvement in the validation loss for 10 epochs. We perform all experiments with a batch size of 16. When reconstruction regularization is used, the loss is calculated using the sum of squared errors of the prediction differences, applied with a coefficient of 0.03. We normalize each of the 27 laboratory parameters to unit mean and variance with respect to the training dataset statistics. After normalization, we set the masked and missing values to zero. We calculate the reconstruction loss only using the known values that are not missing in the training dataset. We apply a weight of 0.625 to the positive samples to account for the class imbalance as described in Supporting information Training settings.

Random Forest Settings

We use the Random Forest implementation of Scikit-learn (v0.22.1) [3]. This implementation does not allow partial training or mini batching. Hence, we apply input masking by repeating the dataset 100 times and applying the input masking on this set. For comparability with the other methods, we impute the missing values using the mean. We use the default model parameters because they have performed the best in our preliminary experiments.

S2 Table. Robustness validation experiments expanded results The results of Table 2 is truncated for simplicity. Here in Table 5, we provide the full robustness validation experiment results.

Table 5. All robustness evaluation AUC results of experiment Model Selection: Validation dataset experiments.

| IMP | DAE | SDAE | FCN | DFCN | RF  |
|-----|-----|------|-----|------|-----|
| NIM | 0.0  | 0.746| 0.799| 0.763| 0.796|
| 0.1  | 0.798| 0.806| 0.810| 0.812| 0.816|
| 0.2  | 0.815| 0.810| 0.817| 0.822| 0.785|
| 0.3  | 0.810| 0.818| 0.822| 0.823| 0.792|
| 0.4  | 0.783| 0.834| 0.834| 0.832| 0.799|
| 0.5  | 0.795| 0.834| 0.838| 0.836| 0.812|
| 0.6  | 0.800| 0.828| **0.838**| **0.843**| 0.830|
| 0.7  | 0.789| 0.827| 0.834| 0.827| **0.831**|
| 0.8  | 0.783| 0.730| 0.833| 0.808| 0.816|
| 0.9  | 0.681| 0.497| 0.822| 0.809| 0.797|

S3 Table. All p values for significance tests in results Clinically relevant subsets A and B The statistical significances mentioned in results Clinically relevant subsets A and B are provided in Tables 6, 7, and 8.
Table 6. p values from comparing AUC results (DeLong’s test) of experiment Clinically relevant subsets A and B on All test data. Models with statistical significance and significant p-values are presented in bold.

| Model 1 | Model 1 Training | Model 2 | Model 2 Training | Test Dataset | p-value |
|---------|------------------|---------|------------------|--------------|---------|
| DAE     | All              | DFCN    | All              | All          | 0.00681 |
| DAE     | All              | FCN     | All              | All          | 0.58261 |
| DAE     | All              | RF      | All              | All          | 0.02875 |
| DFCN    | All              | DAE     | All              | All          | 0.00681 |
| DFCN    | All              | FCN     | All              | All          | 0.03849 |
| DFCN    | All              | RF      | All              | All          | < 0.0000 |
| DFCN    | All              | SDAE    | All              | All          | 0.23376 |
| FCN     | All              | DAE     | All              | All          | 0.58261 |
| FCN     | All              | DFCN    | All              | All          | 0.03849 |
| FCN     | All              | RF      | All              | All          | 0.00134 |
| FCN     | All              | SDAE    | All              | All          | 0.40531 |
| RF      | All              | DAE     | All              | All          | 0.02875 |
| RF      | All              | DFCN    | All              | All          | < 0.0000 |
| RF      | All              | FCN     | All              | All          | 0.00134 |
| RF      | All              | SDAE    | All              | All          | 0.00024 |
| SDAE    | All              | DAE     | All              | All          | 0.16355 |
| SDAE    | All              | DFCN    | All              | All          | 0.23376 |
| SDAE    | All              | FCN     | All              | All          | 0.40531 |
| SDAE    | All              | RF      | All              | All          | 0.00024 |
Table 7. p values from comparing AUC results (DeLong’s test) of experiment statistical significance and significant p-values are presented in bold.

| Model 1          | Model 2          | Model 3          | Test Dataset | p-value |
|------------------|------------------|------------------|--------------|---------|
| DAE              | DFCN             | All              | Subset A     | 0.01578 |
| DAE              | DFCN             | Subset A         | All          | 0.01424 |
| DAE              | RF               | Subset A         | All          | 0.01578 |
| DAE              | RF               | Subset A         | All          | 0.01578 |
| DAE              | SDAE             | Subset A         | All          | 0.01578 |
| DAE              | SDAE             | Subset A         | All          | 0.01578 |
| DAE              | RF               | Subset A         | All          | 0.01578 |
| DAE              | RF               | Subset A         | All          | 0.01578 |
| DAE              | SDAE             | Subset A         | All          | 0.01578 |
| DAE              | SDAE             | Subset A         | All          | 0.01578 |
| DAE              | RF               | Subset A         | All          | 0.01578 |
| DAE              | RF               | Subset A         | All          | 0.01578 |
| DAE              | SDAE             | Subset A         | All          | 0.01578 |
| DAE              | SDAE             | Subset A         | All          | 0.01578 |
| DAE              | RF               | Subset A         | All          | 0.01578 |
| DAE              | RF               | Subset A         | All          | 0.01578 |
| DAE              | SDAE             | Subset A         | All          | 0.01578 |
| DAE              | SDAE             | Subset A         | All          | 0.01578 |
| DAE              | RF               | Subset A         | All          | 0.01578 |
| DAE              | RF               | Subset A         | All          | 0.01578 |
Table 8. p values from comparing AUC results (DeLong’s test) of experiment
Clinically relevant subsets A and B on Subset B labels of the test data. Models with
statistical significance and significant p-values are presented in bold.

| Model 1 | Model 1 Training | Model 2 | Model 2 Training | Test Dataset | p-value |
|---------|------------------|---------|------------------|--------------|---------|
| DAE     | DFCN             | All     | Subset B         | Subset B     | 0.00876 |
| DAE     | DFCN             | Subset B| Subset B         | Subset B     | 0.00876 |
| DAE     | FCN              | Subset B| Subset B         | Subset B     | 0.00108 |
| DAE     | RF               | Subset B| Subset B         | Subset B     | 0.00521 |
| DAE     | RF               | Subset B| Subset B         | Subset B     | 0.00109 |
| DAE     | SDAE             | Subset B| Subset B         | Subset B     | 0.04032 |
| DAE     | DFCN             | All     | Subset B         | Subset B     | 0.00002 |
| DAE     | DFCN             | Subset B| Subset B         | Subset B     | 0.00032 |
| DAE     | FCN              | Subset B| Subset B         | Subset B     | 0.00019 |
| DAE     | RF               | Subset B| Subset B         | Subset B     | 0.00114 |
| DAE     | SDAE             | Subset B| Subset B         | Subset B     | 0.00596 |
| DAE     | SDAE             | Subset B| Subset B         | Subset B     | 0.04280 |
| DAE     | DFCN             | All     | Subset B         | Subset B     | 0.00021 |
| DAE     | DFCN             | Subset B| Subset B         | Subset B     | 0.00019 |
| DAE     | FCN              | Subset B| Subset B         | Subset B     | 0.00019 |
| DAE     | RF               | Subset B| Subset B         | Subset B     | 0.00021 |
| DAE     | SDAE             | Subset B| Subset B         | Subset B     | 0.04803 |
| DAE     | DFCN             | All     | Subset B         | Subset B     | 0.98772 |
| DAE     | DFCN             | Subset B| Subset B         | Subset B     | 0.65732 |
| DAE     | FCN              | Subset B| Subset B         | Subset B     | 0.00521 |
| DAE     | RF               | Subset B| Subset B         | Subset B     | 0.00108 |
| DAE     | SDAE             | Subset B| Subset B         | Subset B     | 0.04032 |
| DAE     | DFCN             | All     | Subset B         | Subset B     | 0.00002 |
| DAE     | DFCN             | Subset B| Subset B         | Subset B     | 0.00032 |
| DAE     | FCN              | Subset B| Subset B         | Subset B     | 0.00019 |
| DAE     | RF               | Subset B| Subset B         | Subset B     | 0.00114 |
| DAE     | SDAE             | Subset B| Subset B         | Subset B     | 0.00596 |
| DAE     | SDAE             | Subset B| Subset B         | Subset B     | 0.04280 |
| DAE     | DFCN             | All     | Subset B         | Subset B     | 0.00021 |
| DAE     | DFCN             | Subset B| Subset B         | Subset B     | 0.00019 |
| DAE     | FCN              | Subset B| Subset B         | Subset B     | 0.00019 |
| DAE     | RF               | Subset B| Subset B         | Subset B     | 0.00021 |
| DAE     | SDAE             | Subset B| Subset B         | Subset B     | 0.04803 |
| DAE     | DFCN             | All     | Subset B         | Subset B     | 0.98772 |
| DAE     | DFCN             | Subset B| Subset B         | Subset B     | 0.65732 |
| DAE     | FCN              | Subset B| Subset B         | Subset B     | 0.00521 |
| DAE     | RF               | Subset B| Subset B         | Subset B     | 0.00108 |
| DAE     | SDAE             | Subset B| Subset B         | Subset B     | 0.04032 |
| DAE     | DFCN             | All     | Subset B         | Subset B     | 0.00002 |
| DAE     | DFCN             | Subset B| Subset B         | Subset B     | 0.00032 |
| DAE     | FCN              | Subset B| Subset B         | Subset B     | 0.00019 |
| DAE     | RF               | Subset B| Subset B         | Subset B     | 0.00114 |
| DAE     | SDAE             | Subset B| Subset B         | Subset B     | 0.00596 |
| DAE     | SDAE             | Subset B| Subset B         | Subset B     | 0.04280 |
| DAE     | DFCN             | All     | Subset B         | Subset B     | 0.00021 |
| DAE     | DFCN             | Subset B| Subset B         | Subset B     | 0.00019 |
| DAE     | FCN              | Subset B| Subset B         | Subset B     | 0.00019 |
| DAE     | RF               | Subset B| Subset B         | Subset B     | 0.00021 |
| DAE     | SDAE             | Subset B| Subset B         | Subset B     | 0.04803 |
| DAE     | DFCN             | All     | Subset B         | Subset B     | 0.00021 |
| DAE     | DFCN             | Subset B| Subset B         | Subset B     | 0.00019 |
| DAE     | FCN              | Subset B| Subset B         | Subset B     | 0.00019 |
| DAE     | RF               | Subset B| Subset B         | Subset B     | 0.00021 |
| DAE     | SDAE             | Subset B| Subset B         | Subset B     | 0.04803 |
| DAE     | DFCN             | All     | Subset B         | Subset B     | 0.00021 |
| DAE     | DFCN             | Subset B| Subset B         | Subset B     | 0.00019 |
| DAE     | FCN              | Subset B| Subset B         | Subset B     | 0.00019 |
| DAE     | RF               | Subset B| Subset B         | Subset B     | 0.00021 |
| DAE     | SDAE             | Subset B| Subset B         | Subset B     | 0.04803 |

References

1. Smith LN. Cyclical Learning Rates for Training Neural Networks. arXiv:150601186 [cs]. 2017.
2. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. arXiv:151200567 [cs]. 2015.
3. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(85):2825–2830.