We divided the task into two categories, as suggested in the problem statement:

1. Subtask 1+2 ---> classification
2. Subtask 3 ---> regression

Before getting into the models some work needed to be done on the data itself.

Firstly we looked summed up the 12 hours measurements of each patient into one single line by taking the mean of the values. Obviously this could lead to loss of time dependency like trends in the data. Therefore we tried to include other statistical measurements like variance and standard deviation but this resulted in no improvement in final score so we stuck to only mean in order to reduce features dimensionality.

This thesis was also supported by trying to use sklearn iterative imputer aiming for better data imputation but also this did not give significative improvements.

Via manual inspection we realized that some columns had great negative impact on the algorithm performance. We concluded this to be a clear sing of corrupted data.

Therefore the final solution has been one line for each patient with the mean of its hourly values + simple imputation within the data to fill the nan and feature selection using various statistics to get rid of misleading data.

After this we normalized the data to obtain faster and more stabile training.

1. For classification we firstly opted for a svm but it showed its limitations the given data. We outperformed it using a NN classifier but the real step up was using Gradient Boosting classifier. We trained one model for each medical test.
2. For regression we followed a similar path by starting out with normal ridge regression ending up using Gradient Boosting regression.