# Manual for using the Novelty Detection Analysis System

## Introduction

The ASIC Use Case of the SMITH project investigates the question how measurements and data of intensive care patients can be used to enable an algorithmic early diagnosis of ARDS. Moreover, an automated investigation of the patient data shall support the physician in the treatment.

Since this involves a very large amount of patient data and measurements, one must consider that these data may contain some errors, e.g., due to sensor failures. To guarantee a high quality of the data analysis, these errors need to be detected. There exists various algorithm targeting this issue, which can detect such errors after the user has annotated a small part of the data manually first.

In his master thesis, Jan van Essen developed a software which provides a graphical user interface which can be used to annotate the data points accordingly, to run various novelty detection algorithms and which provides a visualization of the analysis results. In the following, the operation of the software is described.

## Starting the software

If you just want to use the software and not develop it, it suffices to run the executable file “ndas.exe” which can be found in the folder “NDAS”.

If you want to develop the program, you either need to install a compatible python version and all necessary python libraries onto your computer or you can activate the provided Anaconda environment (ndas\_environment.yml). The automated installation of all packages can be done running the script *install\_dependencies.bat.* Afterwards, the program code can be executed starting the python script “ndas.py”.

Detailed information on this can be found in the developer’s manual “Manual for developers.pdf”.

## Using the software

After starting the program, you are in the main view. On the centre, there is a coordinate system which shows the loaded data and measurements over time.

At the top of the screen, there is a menu bar which enables the user to import and export patient data as well as saving the current state and export the plot into different formats. Below the menu bar, there are some tabs which can be used to switch between the different modules of the software.

## Importing patient data

Using the first menu item, “import”, patient data can be loaded into the annotation view. There are several options to do this:

* Loading csv files
* Loading waveform files
* Loading data from the icu data bases that the i11 has locally instantiated.

The CSV files must follow the following format:

* The first column contains the time axis (e.g., minutes from ICU admission), so the values that are used as X values for plotting the data. The title of the column is irrelevant.
* The remaining columns contain the time series data for different parameters that belong to the time axis, so the values, which are used as Y values for plotting the data. The column title is the title of the corresponding plot. However, there are some columns which are explicitly excluded if they contain specific substrings. By default, the following are these strings: : „ID“, „gender“, „age“, „ethnicity“, „height“, „weight“ and „ICD“, but this can be changed inside the configuration menu (accessible over Settings/Configure options).
* Optionally, some metadata and information on ICD 9 diagnoses, which are then depicted inside the imputation view, can be added. To do this, use the following column names:
  + **ID:** ID of the patient (can be any number)
  + **gender(n m f):** Gender of the patient. Possible values are 1 (male) and 2 (female). Any other values will be interpreted as “not specified”.
  + **age(90= >89):** Age of the patient. Any integer can be used - values over 89 will be shown as “>89”, values blow 0 will be shown as “unknown”.
  + **ethnicity(n cauc asia hisp afram natam):** Ethnicity of the patient. Possible values are 1 for “Caucasian”, 2 for “Asian”, 3 for “Hispanic”, 4 for “African American” and 5 for “Native American”. All other values will be depicted as “Not specified / other”.
  + **height(cm):** Height of the patient (any decimal number).
  + **weight(kg):** Weight of the patient (any decimal number).

If weight as well as height are given, then the BMI of the patient is automatically calculated and displayed.

For all previous columns, the corresponding value must be given in the first row (after the column title) – all other columns will be ignored.

* + Additionally, some **ICD-9-Codes** for groups of diseases from the categories “Diseases of blood”, “diseases of the circulatory system”, “Diseases of the respiratory system” and “injuries” can be given. Then, the column name must start with “ICD”, followed by the first and the last code that belongs to the group. (e.g. ICD460466 for acute respiratory infections”). The column values can be 0 (which means that the disease is currently not diagnosed) and 1 (which means that the disease is diagnosed at the given time).

To load data from the i11 databases, the following requirements must be fulfilled:

* The user has to be connected to the VPN of the i11.
* The user needs access to the “smith-interface” server and to the database. The access data can be configured using the menu items “Settings/Configure SSH authentication data”, “Settings/Database authentication data” respectively.

### Annotation

In the centre of the annotation view opened at the program start, there is a coordinate system which depicts the loaded and selected patient data. Using the right mouse key, the view can be moved. Using the right mouse key while pressing the CTRL button, the view gets stretched and compressed. The view can be moved and stretched or compressed respectively simultaneously using the left mouse key. Zooming can be done using the mouse wheel.

Data points can be selected by clicking on them. Dragging a window over multiple points using the left mouse key selects all those points. Above the coordinate system, there are some control elements which can be used to annotate the selected datapoints. Additionally, the selected points can be marked or demarked respectively as novelty manually. If you move the mouse pointer over a data point, a small tooltip is shown which presents the exact coordinates (value and timestamp).

On the right side, inside the box *Analysis Settings,* you can select and execute a novelty detection algorithm. In the area *Visualization* you can change some graphical settings on the plot. Moreover, the plot to be shown can be selected here if the loaded data contain more than one parameter.

After executing a novelty detection algorithm, the data points will be marked in different colours. The different colours have the following meanings:

* *Gray:* data points which were not used by the algorithm,
* *Turquoise:* data points which were be used as training data,
* *Blue:* data points which were be marked as normal
* *Yellow* and *red:* Data points which were be detected as anomaly.

Executions of the data imputation (see next section) create the following additional colours:

* *Pink:* points, which were not in the data before and were added
* *Violet:* points which existed originally in the data, but which were removed

### Data Imputation

In statistics, imputation designates a method to calculate missing data in datasets. So, the data imputation functionality of the NDAS enables to fill gaps in the data plots.

The imputation view provides an overview on the existing data.

On the left side, the timelines of the existing data series and the associated statistics (number of points, range of values, mean value, quartiles) are shown. Additionally, the timeline of diagnoses is shown at the top, whereby changes are depicted as a blue box. If you move the mouse pointer over such a box, the information of the diagnosis changes is shown. (The left side is not shown until some data were loaded.)

The right side contains the control elements and additional information. Inside the box *Data Imputation Settings*, the imputation method can be chosen, and the determined median time interval is shown. Below that, it can be chosen whether intermediate points should be generated to increase the data density (density multiplier). Using the *Impute* button, the imputation can be executed. The checkboxes inside the box *Data Visualization Settings* can be used to hide or show the different timelines and below that, the span which should be shown on the x axis can be set. Moreover, you can switch between the imputation result and the original data. The button *Apply imputation results onto loaded dataset* can be used to replace the original data with the imputed values (which means that all other modules of the software will work with the imputed data from then on).

The box *Patient Information* contains information about the current patient. Besides information like ID, gender, age, ethnicity, height, weight and BMI, these are also diagnoses which occurs in the selected observation period. When the mouse pointer is moved over such a diagnosis, a designation of the diagnosis category and the information, when the diagnosis was added or removed respectively.

### Mass Error-Correction

This module can be used to run the novelty detection and imputation automated and parallel on different patient data. This functionality is decoupled from the other modules. First, you have to choose the patient data which are to be corrected and a destination folder (and the name suffix). Moreover, it can be chosen if a “mask matrix” should be exported, which depicts, which values were changed. Finally, you have to choose the detection and imputation methods.

After pressing the execution button, the data correction is started using the chosen settings. A progress bar shows the general progress and estimated remaining duration. Below that, the individual progress for the currently active operation is shown. Using the *Early Stop* button, the execution can be stopped early, which means, that the data corrections, which were not started yet, will be aborted, so that only the currently running processes have to finish.

### Statistics

Some statistical information about the imported patient data is graphically presented here.

### Data Inspector

Inside the *Data Inspector* view, the imported patient data are completely presented in tabular form. By clicking on the corresponding column name, the data records can be sorted.

### Data Generator

With this module, test data can be generated which can be used to test the functionalities of the program. Doing this, various anomalies can be inserted into the test data. (This module is currently under development and hence not active.)

### Benchmark

The benchmark system is based on the data generator and enables the possibility to test different novelty detection algorithms. In the first step, test data are generated, in the second step, the algorithms to be tested are selected and finally an overview is presented which shows how well the algorithm detected anomalies. Of course, this only works on test data for which the anomalies are already known before.