# Anleitung zur Verwendung des Novelty-Detection-Analysis-Systems

## Einführung

Der ASIC Use Case im SMTIH-Projekt beschäftigt sich mit der Frage, wie Messwerte und Daten, die von Patienten erhoben wurden, welche auf Intensivstationen behandelt wurden, dazu verwendet werden können, eine algorithmische Früherkennung von ARDS zu ermöglichen. Außerdem soll die automatisierte Untersuchung der Patientendaten den Arzt bei der Behandlung unterstützen.

Bei der Untersuchung solcher großen Datenmengen ist jedoch zu beachten, dass hin und wieder Fehler in den Daten auftreten können, wie beispielsweise Sensorfehler. Um eine hohe Qualität der Datenanalyse zu gewährleisten, müssen solche Fehler erkannt werden. Dafür existieren verschiedene Algorithmen, die nach einer vorhergehenden Annotation einiger Datenpunkte Fehler erkennen können.

In seiner Masterarbeit entwickelte Jan van Essen eine Software, welche eine grafische Oberfläche bereitstellt, in der Datenpunkte entsprechend annotiert werden können, die Ausführung verschiedener solcher Algorithmen möglich ist und das Ergebnis visualisiert wird. Im Folgenden wird die Bedienung der Software kurz erläutert.

## Start des Programms

Wenn Sie die Software lediglich benutzen und nicht weiterentwickeln wollen, können Sie sie einfach starten, indem sie die Datei „ndas.exe“ im Ordner NDAS ausführen.

Wenn Sie das Programm weiterentwickeln wollen, müssen Sie entweder auf ihrem PC die korrekte Python-Version sowie alle notwendigen Python-Pakete installieren, oder Sie müssen die bereitgestellte Anaconda-Umgebung (ndas\_environment.yml) aktivieren. Zur automatisierten Installation der Pakete kann das Skript *install\_dependencies.bat* ausgeführt werden. Anschließend kann man den Programmcode ausführen, in dem man das Python-Skript ndas.py ausführt.

Detailliertere Informationen finden sich in der Entwickler-Anleitung „Manual for developers.pdf“.

## Grundlegendes zur Bedienung

Nach dem Programmstart befindet man sich in der Hauptansicht. Zentraler Bestandteil ist ein Koordinatensystem, in der geladene Messwerte und Daten im zeitlichen Verlauf dargestellt werden.

Am oberen Bildschirmrand befindet sich eine Menüleiste, über welchen man Patientendaten importieren und exportieren kann sowie der aktuelle Zustand gespeichert und der Plot in verschiedene Formate exportiert werden kann. Unter der Menüleiste befinden sich einige Registerkarten, mit denen zwischen den verschiedenen Modulen der Software umgeschaltet werden kann.

## Import von Patientendaten

Über den ersten Punkt der Menüleiste, „Import“, können Patientendaten in die Annotationsansicht geladen werden. Dafür stehen mehrere Möglichkeiten zur Auswahl:

* Laden von CSV-Dateien
* Laden von Waveform-Dateien
* Import von Daten aus den intensivmedizinischen Datenbanken, die zurzeit dem i11 zur Verfügung stehen.

Die CSV-Dateien müssen das folgende Format haben:

* Die erste Spalte gibt die Zeitachse (beispielsweise Minuten ab Aufnahme in die Intensivstation) an, also die Werte, die im geladenen Plot als X-Werte für alle Punkte verwendet werden. Der Spaltenname ist irrelevant.
* Die restlichen Spalten enthalten die Zeitreihendaten, die zu der entsprechenden Zeitachse gehören, also die Werte, die im geladenen Plot als Y-Werte für alle Punkte verwendet werden. Der Spaltenname ist der Titel des entsprechenden Plots. Ausgenommen hiervon sind Spalten, deren Name bestimmte Substrings enthalten. Per Default gehören folgende Strings zu diesen Ausnahmefällen: „ID“, „gender“, „age“, „ethnicity“, „height“, „weight“ and „ICD“. Diese können jedoch über das Konfigurationsmenü (erreichbar über Settings/Configure Options) angepasst werden. Spalten, deren Name einen dieser Strings enthalten, werden ignoriert.
* Optional können einige Metadaten sowie Daten zu einigen ICD-9-Diagnosen mit eingefügt werden, die in der Imputationsansicht mit angezeigt werden. Dies kann durch die folgenden Spaltennamen geschehen:
  + **ID:** ID des Patienten (beliebige Zahl)
  + **gender(n m f):** Geschlecht des Patienten. Mögliche Werte sind 1 (für männlich) und 2 (für weiblich). Alle anderen Werte werden als „nicht spezifiziert“ interpretiert.
  + **age(90= >89):** Alter des Patienten. Ein beliebiger Integer kann eingetragen werden – Werte über 89 werden als „>89“ angezeigt, Werte unter 0 als „unkown“.
  + **ethnicity(n cauc asia hisp afram natam):** Ethnie des Patienten. Folgende Werte sind möglich: 1 für „Caucasian“, 2 für „Asian“, 3 für „Hispanic“, 4 für „African American“ und 5 für „Native American“. Alle anderen Werte werden als „Not Specified / Other“ angezeigt.
  + **height(cm):** Körpergröße des Patienten (beliebige Dezimalzahl)
  + **weight(kg):** Gewicht des Patienten (beliebige Dezimalzahl).

Sofern sowohl Gewicht als auch Körpergröße angegeben werden, wird automatisch der BMI des Patienten berechnet und mit angezeigt.

Bei allen vorangegangenen Spalten muss der entsprechende Wert jeweils in der ersten Zeile stehen; alle weiteren Zeilen werden ignoriert.

* + Es können ebenfalls **ICD-9-Codes** für Gruppierungen von Krankheitsbildern aus den Kategorien Erkrankungen des Blutes, Erkrankungen des Kreislaufsystems, Erkrankungen der Atemwege und Verletzungen angegeben werden. Der Spaltenname muss mit „ICD“ beginnen, gefolgt vom ersten und letzten Code, die zu dieser Gruppierung gehören (also beispielsweise ICD460466 für akute Atemwegsinfektionen). Die Spaltenwerte können 1 oder 0 annehmen – 1 bedeutet, dass die Diagnose zu dem gegebenen Zeitpunkt aktiv ist, und 0 bedeutet entsprechend, dass die Diagnose nicht aktiv ist.

Damit Daten aus den i11-Datenbanken geladen werden können, müssen die folgenden Voraussetzungen erfüllt sein:

* Der Nutzer muss mit dem VPN des i11 verbunden sein.
* Der Nutzer braucht einen Zugang zum „smith-interface“-Server des i11 und zur Datenbank. Beide können über die Menüpunkte „Settings/Configure SSH authentication data“ bzw. „Settings/Database authentication data“ konfiguriert werden.

### Annotation

In der Annotationsansicht, die beim Programmstart geöffnet wird, befindet sich in der Mitte ein Koordinatensystem, indem die geladenen und ausgewählten Patientendaten dargestellt werden. Mithilfe der rechten Maustaste kann man die Ansicht verschieben. Bei gedrückter STRG-Taste wird mit der rechten Maustaste die Ansicht gestreckt und gestaucht und mit der linken Maustaste die Ansicht gleichzeitig verschoben und gestreckt bzw. gestaucht. Hinein- oder herauszoomen kann man mithilfe des Mausrads.

Mithilfe eines Linksklicks auf einen Datenpunkt wird dieser Punkt selektiert. Zieht man mit der linken Maustaste ein Fenster über mehrere Punkte, werden alle diese Punkte ausgewählt. Oberhalb des Koordinatensystems befinden sich Steuerelemente, mithilfe derer ausgewählte Datenpunkte annotiert werden können. Zusätzlich können die ausgewählten Punkte manuell als Novelty markiert bzw. entmarkiert werden. Fährt man mit dem Mauszeiger über einen bestimmten Datenpunkt, zeigt ein kleiner Tooltip die genauen Koordinaten (Wert und Zeitpunkt) an.

Auf der rechten Seite, innerhalb der Box *Analysis Settings,* kann ein Novelty-Detection-Algorithmus ausgewählt und ausgeführt werden. Im Bereich *Data Slicing* kann die Menge der Datenpunkte auf ein bestimmtes Intervall eingegrenzt werden. Im Bereich *Visualization* können einige grafische Einstellungen am Plot vorgenommen werden. Zudem kann hier der darzustellende Plot ausgewählt werden, falls mehrere Parameter in den geladenen Daten verfügbar sind.

Nach dem Ausführen eines Novelty-Detection-Algorithmus werden die Datenpunkte verschiedenfarbig markiert. Die verschiedenen Farben bedeuten im Einzelnen folgendes:

* *Grau:* Vom Algorithmus nicht verwendete Datenpunkte,
* *Türkis:* Als Trainingsdaten verwendete Datenpunkte,
* *Blau:* Als normal erkannte Datenpunkte
* *Gelb* und *rot:* Als Anomalie erkannte Datenpunkte.

Ausführen der Datenimputation (nächster Tab) erzeugen folgende zusätzliche Farben:

* *Pink*: Punkte, die vorher nicht in den Daten waren und ergänzt wurden
* *Violett*: Punkte, die vorher in den Daten waren, aber entfernt wurden

### Data Imputation

Imputation bezeichnet in der Statistik eine Methode, fehlende Werte in Datensätzen zu berechnen. Mithilfe der Imputationsfunktion des NDAS können also Lücken in den Plots geschlossen werden.

In der Imputationsansicht wird ein Überblick über die vorhandenen Daten gewährt.

Auf der linken Seite werden die Zeitverläufe der im Datensatz vorhandenen Datenreihen, sowie die dazugehörigen Statistiken (Punktzahl, Wertebereich, Mittelwert, Quartile) angegeben.   
Ganz oben wird zusätzlich der Zeitverlauf von Diagnosen dargestellt, wobei Veränderungen als blaue Box dargestellt werden. Wenn die Maus über eine solche Box gehalten wird, werden die Informationen zu den Diagnosen-Veränderungen des jeweiligen Zeitpunkts angezeigt. (Die linke Seite wird erst angezeigt, wenn ein Datensatz reingeladen wurde)

Auf der rechten Seite sind die Steuerelemente und zusätzliche Informationen zu finden. In der Box *Data Imputation Settings* kann die Imputationsmethode gewählt werden. Der ermittelte Median-Zeitabstand zwischen Messungen ist dort zu finden. Darunter kann gewählt werden, ob Zwischenpunkte erzeugt werden sollen, um die Datendichte zu erhöhen (Dichte-Multiplikator). Schließlich kann per Knopfdruck eine Imputation durchgeführt werden.

In der Box *Data Visualization Settings* können Datenreihen versteckt werden, sowie der X-Betrachtungsbereich gewählt werden. Außerdem ist es hier möglich das Imputationsergebnis mit den Ursprungsdaten zu vergleichen, indem die Toggle bedient wird.

Mit Click auf *Apply* *imputation results onto loaded dataset* wird der interne Datensatz durch dir imputierten Werte ersetzt (sodass alle anderen Tabs dann mit den imputierten Daten arbeiten).

In der *Box Patient Information* werden schließlich Informationen zum aktuellen Patienten angezeigt. Neben Angaben wie ID, Geschlecht, Alter, Ethnizität, Größe, Gewicht und BMI werden hier nochmal alle während des Betrachtungszeitraums auftretenden Diagnosen hervorgehoben. Wenn die Maus über eine Diagnose gehalten wird, erscheint die Bezeichnung der Diagnosegruppe, sowie ein Zeitverlauf der Diagnose (wann wurde sie hinzugefügt bzw. entfernt)

### Mass Error-Correction

In diesem Bereich kann Novelty-Detection und Imputation automatisiert auf mehrere Patientendaten parallel durchgeführt werden. Der Bereich ist von den anderen Bereichen entkoppelt. Zunächst müssen die zu korrigierenden Patientendaten ausgewählt und ein Zielordner (sowie Namenszusatz) gewählt werden. Weiterhin kann gewählt werden ob zusätzlich eine „Maskenmatrix“ exportiert wird, die darstellt welche Werte verändert wurden.  
Schließlich werden die Detektions- und Imputations-Methoden gewählt.

Durch Drücken der ausführen Taste wird die Datenkorrektion mit den gewählten Optionen gestartet. Ein Fortschrittsbalken zeigt den allgemeinen Fortschritt sowie die errechnete übrige Zeit an. Darunter wird für die aktuell aktiven Operationen deren individueller Fortschritt gezeigt. Mittels des *Early Stop* Knopfes kann die Ausführung frühzeitig beendet werden. Hier werden die noch nicht begonnenen Datenkorrektionen abgebrochen, sodass nur noch die Laufenden Prozesse beendet werden müssen.

### Statistics

In diesem Bereich werden einige statistische Informationen über die importierten Patientendaten grafisch dargestellt.

### Data Inspector

In der *Data Inspector-*Ansicht werden die importieren Patientendaten komplett in tabellarischer Form dargestellt. Die Datensätze können durch Klick auf die entsprechenden Spaltennamen sortiert werden.

### Data Generator

Hier können Testdaten generiert werden, mithilfe derer die Funktionalitäten des Programms ausprobiert werden können. Hierbei können verschiedene Anomalien eingebaut werden. (Dieser Tab wird aktuell überarbeitet und ist daher nicht aktiv)

### Benchmark

Das Benchmarksystem baut auf dem Datengenerator auf und ermöglicht es, verschiedene Novelty-Detection-Algorithmen zu testen. Im ersten Schritt werden die Testdaten generiert, im zweiten Schritt werden die zu testenden Algorithmen ausgewählt und anschließend erhält man eine Übersicht, anhand derer zu erkennen ist, wie gut der Algorithmus Anomalien erkannt hat. Dies funktioniert logischerweise nur auf generierten Testdaten, bei denen die Anomalien bereits im Vorfeld bekannt sind.