BACHELOR THESIS

SUBPOPULATION PROCESS MINING IN HEALTH

YOALINA KALCHEVA

CREATIVE TECHNOLOGY

SUPERVISOR: FAIZA BUKHSH

CRITICAL OBSERVER: NACIR BOUALI

DATE 03.02.2023

Department of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, Enschede, the Netherlands

UNIVERSITY OF TWENTE.

Abstract

Process mining is a technique used for analyzing processes based on event logs. In this study, process mining techniques will be applied in the health care sector for evaluation of cancer and breast cancer subpopulations. Therefore, a publicly available online MIMIC-III dataset will be used for data extraction as it contains data from the intensive care units of a large hospital. In addition, this research is a continuation of previous studies on this topic and will be extended in terms of data quality and machine learning. For improving the quality of selected data for this research project, the methodology of CRISP-DM (CRoss Industry Standard Process for Data Mining) model is followed. Moreover, process mining techniques are applied on the extracted cancer- and breast cancer- related data which helped for finding inconsistencies in the datasets. In addition, the results showed that the combination of machine learning techniques can work together with process mining. However, the data need to be carefully selected so that the results can be applicable. This project can be elaborated in the future by using real-world databases. Furthermore, it can be extended by doing research on more machine learning methods which can be used with process mining techniques.

Key terms: **process mining; process comparison; data quality; machine learning; breast cancer; MIMIC database**

Table of Contents

Table of Figures

1. Introduction

Process mining is a part of process science [1]. It is built upon process model-driven approaches and data mining. Its goal is to improve processes and support authentic insights by using techniques and analyses executed in the form of event logs [2]. Event log consists of a trace identifier, event or activity that is executed, and a timestamp informing about what time the activity has happened [1]. Data quality has a critical role in the health care sector and accordingly it is needed for reliable construction of model performance for machine learning applications [3,4]. In regard, the improvements of the processes and visualizations of event logs will lead to more accurate results of the researched data which will contribute to the outcome of this study – making a comparison between subpopulations regarding data quality and machine learning [5]. The objective of this study is to investigate commonalities and differences of the care paths of subpopulations of patients suffering from cancer and more specifically, breast cancer. Consequently, the research will be extended in terms of data quality and machine learning as it is a continuation of previous work. In addition, an evaluation of data extracted from publicly available MIMIC-III Critical Care database related to cancer and breast cancer will be made. Moreover, MIMIC-III database (Medical Information Mart for Intensive Care) [7] is publicly available data related to adult patients admitted to critical care units between 2001 and 2012. It contains data of approximately 53,000 distinct hospital admissions with comprehensive information about individual patient care.
The purpose of this research project is to evaluate how combination of machine learning and process mining work together for bigger dataset on complex data analytics problems. If these two techniques work together, the results would have an impact in business and in data science domains. Moreover, researchers and people in the healthcare sector will benefit from them.

Figure 1 shows the overall process of this research in five steps described in this report. The first step was to define and understand the objective of the project (Section 1). After which, a search strategy was created and relevant articles to this research topic were selected based on the search methods and inclusion and exclusion criteria (Section 2). After that, the research questions were defined and research on a related work was made. The next step was to understand and follow the methodology of Cross Industry Standard Process for Data Mining (CRISP-DM) model (Section 3) for improving data quality. After which, MIMIC-III Critical Care database was analyzed, relevant data was selected, and prepared for evaluation (Section 4). For the model evaluation process mining techniques were observed (Section 5) and applied to the selected cancer and breast cancer data in ProM Software (Section 6). The processed models were analyzed and evaluated. The outcomes contributed to the answers to the research questions. As a result, a conclusion was conducted (Section 7) and discussion about limitations, recommendations and future work was made (Section 8).

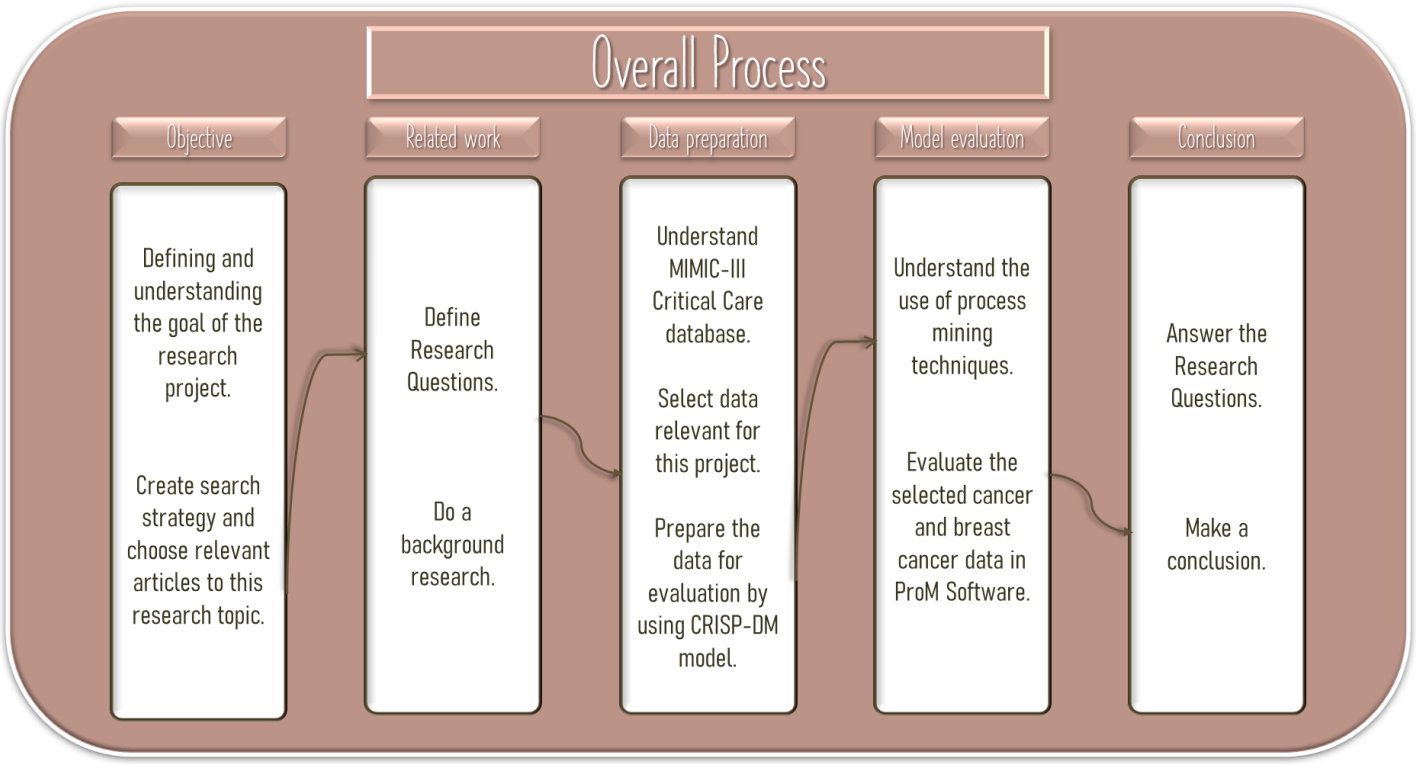


Figure 1. Overall process of the research project.

2. Research Methods

This section will provide an overview of the approach strategies referring to this research which are based on the methodology principles of Barbara Kitchenham et al. [6] following "Evidence-based Software Engineering" (EBSE) method. They claim that using systematic literature review helps researchers to understand the specifics of evidence-based guidelines. The goal of learning what has already been explored and found, is to help practitioners to develop innovative software engineering solutions to existing problems. Furthermore, Kitchenham's et al. observation states that researchers who use the approach of systematic literature review improve with practice, which makes the set baselines more qualitative to build upon.

The following sub-sections are made in a systematic way and describe the research questions, search strategy, inclusion and exclusion criteria, and execution of synthesis strategy. In addition, previous work related to process mining techniques and MIMIC dataset is explained. Consequently, partial answers to the research questions are provided.

2.1. Research questions

The main research question (**RQ**) aims to evaluate the quality of cancer and breast cancer data extracted from MIMIC-III Critical Care database by applying machine learning techniques: *How can we assure quality evaluation of data extracted from public, provided online healthcare dataset - MIMIC-III Critical Care database, related to cancer and breast cancer?*

There are three sub-research questions which are necessary for contribution to the answer of the main RQ:

- **RQ1:** *To what extent is the data, related to cancer and specifically breast cancer from MIMIC-III dataset, qualitative using machine learning (via process mining techniques) regarding previous research on this topic?*
  Literature regarding the MIMIC-III database will be explored. Therefore, it will be clearly made what disease it contains, and which ones are cancer- and breast cancer- related. Based on this, the question will be answered withing the scope of the literature review.

- **RQ2:** *What are the similarities and differences observed from the data analysis (using process mining techniques) concerning cancer- and breast cancer- related data extracted from MIMIC-III dataset regarding data quality?*
  Literature related to data quality will be explored and used to analyze if the evaluated data from MIMIC-III dataset is qualitative following various steps for cleaning and checking data.

- **RQ3:** *What are the similarities and differences observed from the data analysis (using process mining techniques) concerning cancer- and breast cancer- related data extracted from MIMIC-III dataset regarding machine learning?*
  The most common machine learning methods will be extracted from the selected literature. Therefore, they will be applied on cancer-related data from MIMIC-III dataset, and the results will be evaluated. After which a comparison with previous studies on the same topic will be made.

2.2. Search strategy

For this research there were used three digital libraries: Scopus, ScienceDirect and FindUT, as they are ones of the most popular libraries within the scope of University of Twente.

The search query was: MIMIC database AND process mining AND process comparison AND data quality AND machine learning AND breast cancer. It consists of all key terms of this study connected with an AND operator as the main purpose was to find articles which are very similar to the topic of this study. Moreover, if an OR operator is used there were found thousands of results as they contain at least one of the key words. Therefore, this was considered as the worst search method.

2.2.1. Inclusion and exclusion criteria

The articles that are needed for this study should be scientific, easily accessible to future researchers and relevant to the topic. Therefore, they will be selected based on inclusion and exclusion criteria.

The inclusion criteria for this study were:

- **IC1:** Include scientific papers only. This refers to research articles.
- **IC2:** The papers relate to the topic of the research. This means papers related to process mining, machine learning and data quality.
- **IC3:** Reviewed papers, research articles, book chapters, and conference papers.
- **IC4:** Papers are in English.
- **IC5:** Papers have open access and archive.

The exclusion criteria were:

- **EC1:** All nonscientific papers.
- **EC2:** Papers which do not relate to the topic.
- **EC3:** Articles which are not research papers, reviewed, book chapters, or conference abstracts.
- **EC4:** Papers which are not in English.

- **EC5:** Papers with closed access.

After the inclusion and exclusion criteria were applied, there were 100 results left. The title and abstract of each result were evaluated so that the irrelevant papers can be found and removed. It could be considered that most of the articles were related to machine learning methods, some for different cancer types or comparison between datasets. However, with this search strategy there were not many articles related to data quality. Therefore, a snowballing effect was used for gathering more information about data quality and MIMIC-III dataset.

2.2.2. Execution of synthesis strategy

In *Figure 1*, the process of article selection is provided. After executing the search query to the three digital libraries, 362 results were obtained in total. Consequently, the inclusion and exclusion criteria were applied, and the results became 100 in total. Therefore, after closer observation of all these articles, duplicates were removed, as well as irrelevant papers, and some papers were added via snowballing search, which resulted in total of 18 selected articles. These articles were selected by following Barbara Kitchenham's et al. [6] methodology for systematic literature review and will be used for the background research of this project.

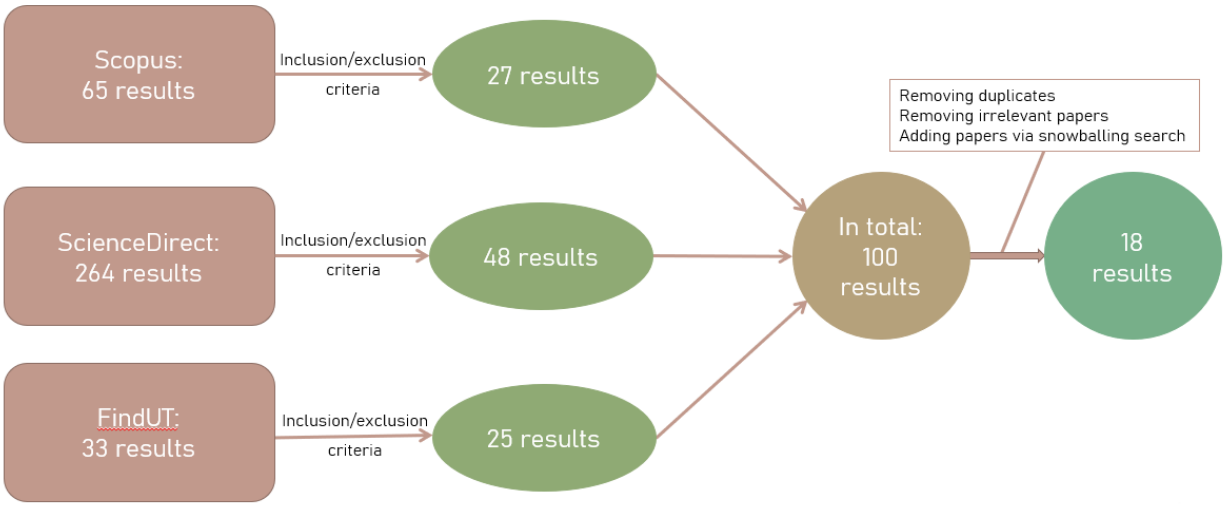


Figure 2. Execution of synthesis strategy.

2.2.3. Answer to the main research question (RQ)

In this part of Section 2, related work regarding the main research question (RQ) is presented. Data extracted from the literature review concerning this section can be found in Appendix A.

There are various newly developed machine learning methods that can be used for comparison of process models for different patient populations. Firstly, Marazza et al. [11] present an automatic comparison of processes for different breast cancer patient populations. The process models extracted from event of electronic health records (EHR) result in smart indication for process similarity. For drawing these conclusions, the data is visualized via process mining techniques using cross-log conformance checking and standard graph similarity measures. In addition to these approaches, the method proposed by Zhao and Weng [17], is weighted Bayesian Network Inference (BNI) model to predict pancreatic cancer. This model is using Bayes' theorem for decrement of risk prediction by being

able to use joint probability distribution which provides analyses of relation of the combination of text-based knowledge with a substantial clinical database. The study shows that weighted BNI model is significantly more accurate than conventional BNI. Similarly, another research provides an accurate prediction of mortality risk for patients in the health care system which can improve care quality and reduce costs [12]. It applies gradient, boosted trees and deep neural networks as machine learning methods on MIMIC III dataset. In addition, neural network classifiers are used to define a methodology for designing data-driven medical diagnostic tools [19]. Furthermore, Peissig et al. [13] conduct an evaluation of relational machine learning (ML) using inductive logic programming (ILP) over electronic health record-driven phenotyping. The goal of this method is to establish and classify information for risk patients from electronic health records (EHRs) in a qualitative way. ILP systems deal with qualitative rules, and thus, with the basics of the learning process.

Slater et al. [16] show that a similarity between patient-patient and patient-disease is classified aiming to set differential diagnosis of common disease based on a technique using unextracted text phenotypes from patients' profiles. MIMIC-III dataset is used for collection of patients' visits and data regarding their diagnoses per visit. The method used for deriving similarity scores is called pairwise similarity. Thus, the Semantic Measures Library using Resnik measure is applied for the calculation of semantic similarity score, as well as HPO codes associated with sampled patient visits and all disease phenotypes are registered. Therefore, to measure the similarity scores, Area Under the receiver operating characteristic Curve (AUC) and Mean Reciprocal Rank (MRR) are used. This approach results in a high-ranking model which differentiates diagnosis of common diseases by showing a powerful performance regarding text-derived patient phenotype profiles which improves the text classification in a clinical setting. Additionally, there are eminent relationships of the patient-patient comparisons. Besides, MIMIC dataset is more useful for differentiation of secondary diagnoses than the primary ones.

Additionally, SVM has been evaluated with two different extensions - support vector machine with linear kernel (linear SVM), support vector machine with radial kernel (radial SVM) [10, 25]. During the conduction of the study three more machine learning classifiers were measured: linear regression, gradient boosting, and naïve Bayes. Intending to measure the best performance of all classifiers, typical confusion matrix evaluation scores have been used. Generally, all algorithms resulted in correct prediction of survivor sepsis patient rate. The goal of the study was to show prediction of survival within patient suffering from sepsis by evaluating the provided machine learning algorithms which were using three medical characteristics to make correct computations: sex, age, and septic episode number. Therefore, Chicco and Jurman [10] evaluated that naïve Bayes has the most accurate prediction of survival rate regarding the test evaluations (primary and study cohort) within sepsis patients. Moreover, by using the same ML methods and *rpart*, another study aims to develop efficient FASTCORMICS RNA-seq workflow to build 10,005 high-resolution metabolic models from the TCGA dataset to capture metabolic rewiring strategies in cancer cells [15]. However, other studies make use of logistic regression (LR), decision tree, and random forest to develop and evaluate a FHIR-based EHR phenotyping framework for identification of patients with obesity and multiple comorbidities using semi-structured discharge summaries [14], as well as to establish a classification of breast cancer survival patterns and offer a treatment decision-making reference [18]. In addition, linear discriminant analysis (LDA) and k-nearest neighbors (KNN) significantly improve the functionality of the initial intelligent remote patient monitoring (IRPM) framework by building three machine learning models for readmission, abnormality, and next-day vital sign measurements [22]. Moreover, Wu et al. [9] evaluated random forest (RF), SVM, deep learning (DL), and gradient boosting decision tree (GBDT). The results showed that GBDT-based

model outperformed the others in both internal and external validation of datasets which is the second widely spread ML method in this review. After which, RF performed the next-best results for both datasets. The reason for this outcome is that in GBDT model the decision trees are iteratively trained which means that each decision tree is fixing the inconsistency of all its previous decision trees. Whereas, in RF model, all decision trees are grown in parallel. In addition, RF is also evaluated as one of the best performing ML algorithms in terms of best AUPRC (Area Under the Precision Recall Curve) values, together with Backward Elimination (BE) by Poucke et al. [8]. Moreover, the combination of RF with Gini Selection and Forward Selection showed higher AUPRC values. The objective of the article was to evaluate different predictive techniques for analyzing the stimulation of Predictive, Preventive and Personalized Medicine (PPPM) in terms of care quality and cost.

In addition, there are more ML techniques used in various specific situations. For example, regarding health care data, Peissig et al. [13] show that ILP + BP (borderline positive) rules provide very high accuracy in analyzing patients' existential biases based on which injuries or diseases can be found, and therefore the phenotype model can be improved. This machine learning approach outperforms the others conducted in the research project such as ILP-1, J48, WEKA, JRIP,PART, by using receiver operator characteristic (ROC). Furthermore, according to data quality and improvement, ensemble methods need to be considered. Moreover, another research aiming to develop a new quantitative image feature analysis scheme and investigate its role along with two genomic biomarkers uses techniques such as Naïve Bayesian network-based classifier, simple multilayer perceptron-based classifier, WEKA data mining software package, SMOTE synthetic data generation algorithm, leave-one-case-out validation method, ROCKIT program [23]. In another study Faster R-CNN, Non-Maximum Suppression (NMS), and Region Proposal Network (RPN) are used for breast mass detection [20]. Therefore, the impact of noise on the training of object detection networks for the medical domain is researched, as well as how it can be mitigated by improving the training procedure. Analogously, a paper presenting new ways for visualization of nonlinear classifiers and improvement of the interpretability of results while maintaining high prediction accuracy, uses different specific combinations of ML techniques such as Localized radial basis function (LRBF) kernel, RBF, ReliefF Method, recursive feature elimination (RFE) with an SVM (SVM-RFE), Sensitivity Analysis [24]. Additionally, there is research which uses multitask MKL method to discriminate early-stage and late-stage cancers using genomic data and gene sets and compare this algorithm against two other algorithms - Cutting-Plane Algorithm and BDForest Algorithm [21].

As a conclusion of this subsection, all the process model approaches are evaluated and proved as considerably more effective and qualitative than other previously used methods. Furthermore, every machine learning technique shows improvement in the quality of data of the respective dataset. However, every approach is used differently based on the case and the provided data which can be extracted and analyzed. In some cases, a combination of techniques is necessary to find the best performing model regarding data quality. In the following subsection, a bar chart visualizing the most used machine learning techniques will be present.

From the data extracted in Appendix A, a relation between the year of publication of the selected articles and number of citations could be created. Therefore, a visualization is presented in *Figure 3*. It can be noticed that the older the papers are, the more citations they have which causes descending linear relation. However, there are more papers published around 2020 which means that the topic of process mining and machine learning is becoming more popular and explored. This statement gives a hint for more future elaboration in this direction.

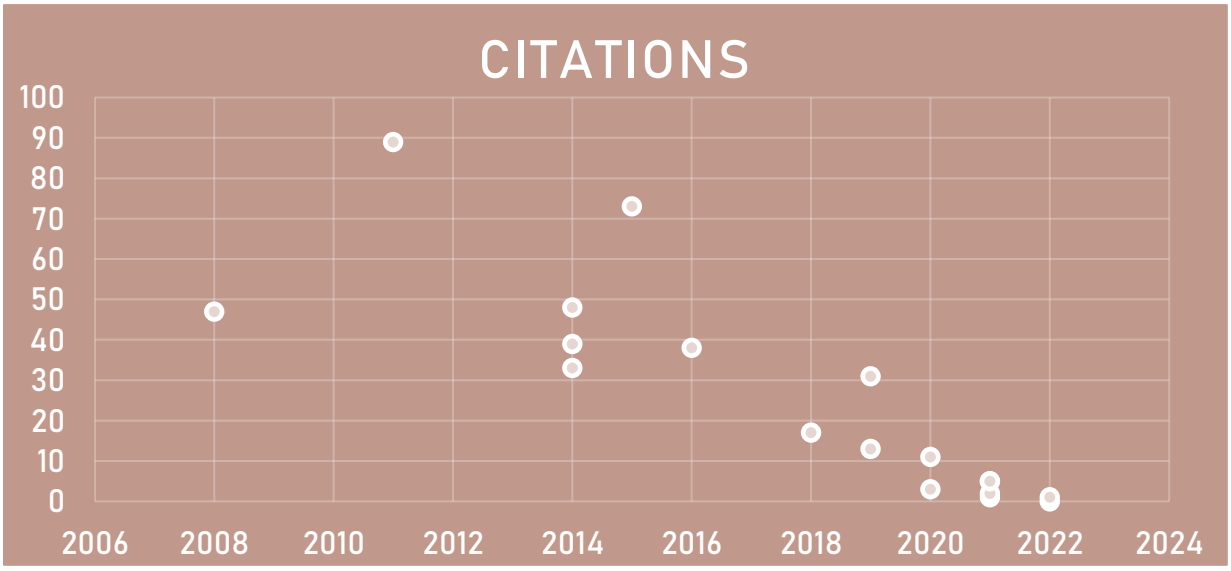


Figure 3. Relation between Year of publication and Citations.

2.2.4. Answer to research questions 1, 2, 3

Data extraction of the selected literature was made regarding the three sub-questions. It can be found in Appendix B. In this sub-section an answer to RQ1 will be provided as it is fully covered withing the literature review. Moreover, partial answers and insights into the other sub-questions will be given.

Machine learning is a process based on artificial intelligence which is learning from experience [1]. Machine learning techniques are evolving without being programmed to do certain things. Therefore, the most popular machine learning methods used among the selected articles have been investigated and can be seen in *Figure 4*. This is Support Vector Machine (SVM), and it is used in 10 out of 18 articles. After which, Naïve Bayes is the second most used one in 5 articles. However, it is considered as the most accurate one [10].

In *Figure 5* the most used database among the relevant articles to this research is visualized. This is MIMIC database – in 7 out of 18 articles. It is followed by Electronic Health Records (EHR) which is a database used in 4 scientific papers in total. Moreover, The Cancer Genome Atlas (TCGA) occurs twice in the selected literature.

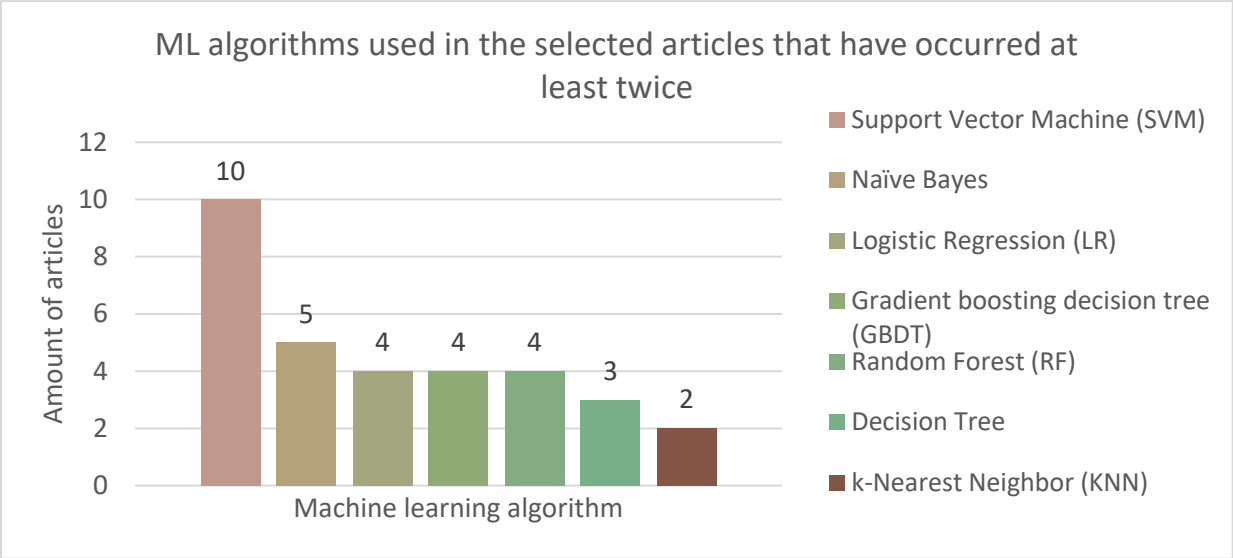


Figure 4. Most used machine learning methods.

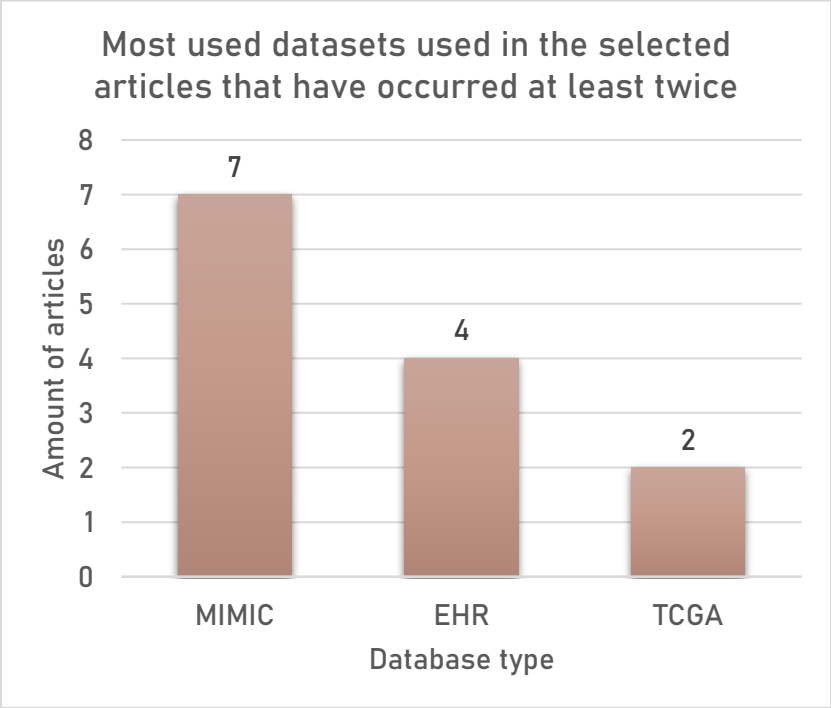


Figure 5. Most used databases extracted from the selected literature.

In *Figure 6* data extraction from the selected literature can be seen in terms of use of supervised and unsupervised machine learning algorithms, breast cancer and comparison between datasets.

Supervised machine learning is a technique used for data classification and prediction of outcomes. Whereas unsupervised ML is analyzing and clustering unbiased datasets [11]. In the bar chart, it can be observed that 16 of the articles use supervised ML algorithms, one uses unsupervised ML techniques, and one does not contain any ML methods.

Breast cancer is a disease in which the cells are growing rapidly in different parts of the breast [18, 21]. Therefore, there are various breast cancer types. It is mentioned in 4 out of 18 scientific papers related to the research topic. In addition, 6 articles contain comparisons between different datasets.

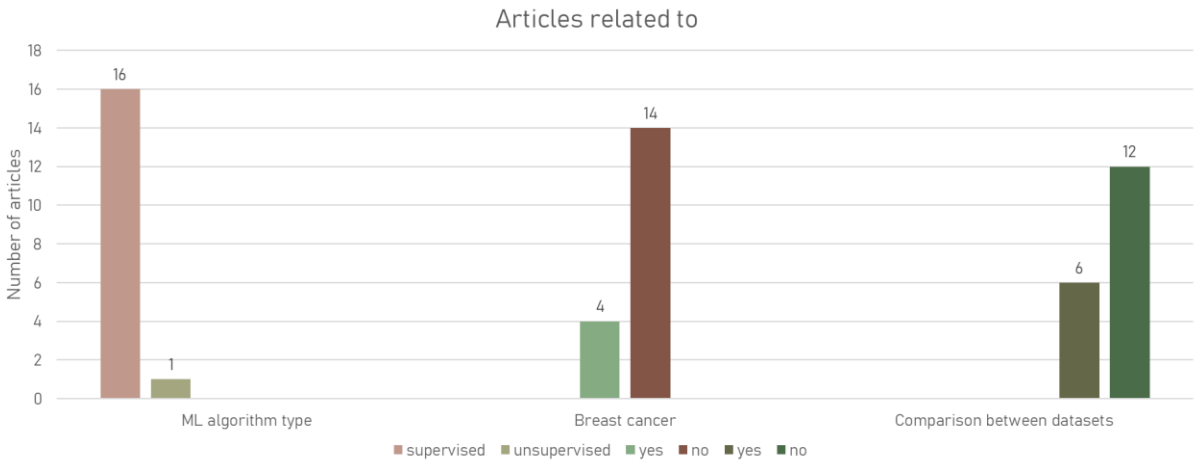


Figure 6. Data Extraction from the selected literature.

Based on the selected literature review there are four main open research directions: optimization and development of phenotyping algorithms; improvement of machine learning model; discovery of a disease; and process model comparison with qualitative investigation. This research is focused on comparing the performance of different modeling techniques, and data quality evaluation.

Overall, regarding the selected literature as a related work for this research topic, it can be concluded that it is appropriate, applicable, and constructive for this research which answers the first research question (RQ1). A substantial part of it refers to different types of machine learning techniques. They are used for improving the accuracy and predictions of the researched results. Additionally, breast cancer and MIMIC database are used in some of the papers, as well as comparison of different datasets.

3. Data Analysis Methodology

This section refers to a data mining model used in this project for improving the quality of the data. In addition, it is shown that following this pattern helps extract subsets of the dataset necessary for the proper and accurate evaluation of the research outcomes.

CRISP-DM (CRoss Industry Standard Process for Data Mining) is a process model used as a framework for data mining projects [27]. Its goal is to ensure quality of knowledge discovery based on the results, as well as optimizing cost and time of the process by reducing the required skills [28]. CRISP-DM model consists of a reference model and a user guide [27, 28]. Reference model is a hierarchical and abstract description of phases, generic tasks, specialized tasks, and process instances. There are six CRISP-DM phases at the top level of the model: business understanding, data understanding, data preparation, modeling, evaluation, deployment. Each phase consists of general tasks which are formed to be stable and complete so that they can cover the whole process of data mining and its applications, as well as future model developments. Business understanding is about understanding the objective of the project and its requirements from a business standpoint and creating a plan for achieving the goal based on the defined data mining problem. The general tasks of this phase are to determine the objectives, assess the situation, define a data mining goal, and produce a project plan. Data understanding consists of collecting,

describing, and exploring the data to establish data quality problems or subsets of data, after which ensuring the quality. In the following phase - data preparation, the data is selected based on inclusion and exclusion criteria, cleaned, constructed, and reformatted so that it can be used in the modelling tool. The general tasks of the modeling phase are selection of a modeling technique for design evaluation, model building and assessment. While generating the models, it is frequently observed that the researcher concocts new ideas about the data selection or finds errors that need to be removed from the data. Afterwards, the tasks of the evaluation process are to assess the results, approve the model, make a review of the process, and determine subsequent possible actions. Finally, the last step consists of deployment plan construction, monitoring and maintenance, finalization of the report and review of project. This step is necessary for the future researchers who are going to use the report or the implemented data mining technique. The final process instance level of the reference model refers to the actual actions that have happened during the data mining processes which are not generalized but specific and different. It is built upon all previous tasks, decisions, and results.

While reference model represents an overview of the different phases, their general tasks and outcomes, the user guide describes in detail recommendations for each phase and tasks and gives an insight on how to do a data mining project.

These methodologies are applicable to this research project as they can be used as guidance on how to process the data in a qualitative way. Therefore, in terms of **business understanding** the objective of this project is to evaluate and compare cancer and breast cancer datasets by applying data quality techniques and machine learning algorithms. The resource from which the data will be extracted is the publicly provided online MIMIC-III dataset. The plan is to select appropriate data for this project such as patients admitted to the hospital with cancer diagnosis, after which create a new dataset with patients suffering from breast cancer. The cancer-related data will be evaluated in ProM Software with machine learning plug-ins. The results will show comparison between datasets in terms of data quality and how the machine learning techniques affect the process model. Regarding **data understanding**, MIMIC-III Critical Care dataset contains data related to clinical care which is credentialed [7]. It is collected from patients who have visited a large hospital. This dataset was chosen as it has been used in previous research on this topic. After gaining access to the dataset, each of the tables (26 in total) was reviewed and analyzed. The selected data for this project included tables containing columns with data for diagnoses, identifiers of the subjects, time of admission to the hospital, and discharge time. This information was in three different tables which can be linked based on the identifier columns such as ROW_ID and SUBJECT_ID. In the following phase of **data preparation**, a final dataset was made including columns with row and subject identifier, short and long name of the disease, time of admission to the hospital and time of discharge. From all deceases the data regarding cancer was selected, after which another table containing only data with breast cancer patients was created. It was noticed that the dates for admission to and discharge from the hospital contained extra values for the years which made the data inconsistent and faulty. Therefore, these extra digits were removed from both datasets, and the data was cleaned. Consequently, the data was **modeled** in ProM Software by using different process mining techniques. Some of these techniques are related to machine learning algorithms so that the goal of the research can be reached. While processing one of the first models, it was noticed that there are unclarities with the column of admission date. That could be understood because the program could not recognize the date pattern as it did for the column of discharged date. For that reason, it was necessary to return to the previous state of data preparation and repair or clean the incorrect values. One mistaken admission date was found. Hence, as it was unclear on which date the patient has been admitted to the hospital, this error could not be fixed.

Therefore, cleaning the patient's record was considered as a solution to this issue. Afterwards, the data was modeled again, and no errors were noticed during the process. However, when inductive visual minor showing service times and paths was applied (Section 6.2 and 6.3), more inconsistencies in the data were found. The admission date of one patient was after their discharge data, therefore the service time (the time of stay in the hospital) was not calculated and visualized in the model. For this reason, the data was prepared once more. All rows containing similar invalid dates were removed, after which the data was integrated. Then, the fixed data was processed another time and there were no divergences found. Following the next phase of evaluation, based on the applied techniques the data was **evaluated** and the results were reviewed before making conclusions. Most of the applied techniques were relevant and useful for the outcome of this research. The ones that did not provide sufficient results or were inappropriate were put in the appendix of the final report. In the last step of **deployment**, the findings of the project were finalized and documented. They might be used for future elaboration on this topic.

4. Data preparation

In this section, data collection, selection and preparation for this research will be described.

Firstly, as this project is a continuation of previous research on this topic, the dataset chosen is the same - MIMIC-III Critical Care database. The data is freely provided online. However, access is restricted and there are requirements for people who would like to use the data. The first step is to create an account at PhysioNet where explanation of the project is necessary as well as details about the university and supervisors. After receiving confirmation for registration and having an official account, a request for access to MIMIC-III database needs to be made by following the steps on the PhysioNet website. This request will be assessed only after completion of CITI program (Collaborative Institutional Training Initiative) which is related to human research. The program is consisted of nine mandatory courses: Belmont Report and Its Principles, History and Ethics of Human Subjects Research, Basic Institutional Review Board (IRB) Regulations and Review Process, Records-Based Research, Genetic Research in Human Populations, Populations in Research Requiring Additional Considerations and/or Protections, Research and HIPAA Privacy Protections, Conflicts of Interest in Human Subjects Research, Massachusetts Institute of Technology. The courses must be passed with a minimum 90% of the grade after which the learner receives a certificate which has to be sent to PhysioNet for confirmation. Moreover, the researcher must sign a usage agreement on using the data securely and appropriately without putting effort into patient identification [7].

Secondly, MIMIC-III consists of 26 tables which are linked by identifiers [7]. The data includes five fields – design, measurements (for example, demographics, medical history), technology (such as medical record and electronic billing system), factor and sample characteristics. For this research, the tables ADMISSIONS, DIAGNOSES_ICD and D_ICD_DIAGNOSES were used as they contain information about diseases of the patients admitted to the hospital. In table D_ICD_DIAGNOSES the diagnoses are described with their short and long name, as well as with their associated code (ICD9_CODE). The table ADMISSIONS contains admission and discharge time of each patient. DIAGNOSES_ICD table is filled in with identification codes and is used to link the other two tables based on ROW_ID from D_ICD_DIAGNOSES and SUBJECT_ID from ADMISSIONS. Therefore, a new table was created containing ROW_ID, SUBJECT_ID, long and short title of the disease, admission, and discharge time.

However, in the newly created table there were found mismatches between the subject identifier and their relative diseases. The reason for that could be because each table had a different number of rows.

To check where this error occurs, a column from the DIAGNOSES_ICD table was added to the other two tables. For example, the column containing ICD9_CODE in DIAGNOSES_ICD was added to D_ICD_DIAGNOSES which also contains ICD9_CODE. For the addition of a column from one table to another, the VLOOKUP() function was used, where the columns were linked based on the row identifier - ROW_ID. The goal of this action was to check if the columns containing ICD-9 codes of both tables are concurring. It was found that the rows of these columns do not dovetail. Similarly, a column with SUBJECT_ID was added, and the ICD-9 codes did not match too. Besides, the table D_ICD_DIAGNOSES had the least number of rows because of which it was decided that columns from the other tables will be linked to the columns of this table as the chances for exact matches are higher. Therefore, the table ADMISSIONS and the table D_ICD_DIAGNOSES were connected based on their ROW_ID columns. The results were exact matches between patient identifiers and their diagnosis. Consequently, the table ICD_DIAGNOSES was considered superfluous, and it was not used.

Thirdly, for this research only cancer data would be appropriate to be evaluated. As there are different cancer types, the 4-digit ICD-9 code will be used for selection. The ICD-9 code for oncology related diseases varies between 140 to 239 [11, 26]. After creating a table with cancer data, data with patients admitted for breast cancer can be extracted in another table.

In conclusion, there are two tables which are going to be used for analysis and evaluation in this research. The first one contains data with patients who have been admitted to the hospital with any cancer type, after which they have been discharged. The second one consists of breast cancer patients and includes the same columns in terms of information as the first table.

5. Process mining

Process mining is a technique used for analyzing processes based on event logs [1]. Event log is an input which presents a certain perspective on the applicable event data. Process mining describes the life cycle of business process management (BPM) by creating a relation between event logs, event data and process models. It includes process discovery using event data, conformance checking using process model, and process enhancement to improve the process model by following the previous methods. The first discovery technique builds a model upon event log, without using any specific information. It can generate a Petri net which visualizes the activities of the given model. Secondly, the conformance technique provides a comparison between a real process model with its discovered event logs. This method can be used to check if the behaviour of an existing reference model correlates to the processed event log. The goal is to measure the fidelity and find possible inconsistencies between the model and the reality, so that they can be removed. Thirdly, there are two types of enhancement: repair and extension. Repair refers to adjusting the processed model so that it correlates more accurately to the existing reference model. Additionally, extension is related to cross-correlation between the process model and the event log, aiming to provide a new aspect of the model. For example, a model can be extended in terms of frequencies, diagnostics, information regarding resources or control-flow.

A sequence of events is called *trace* [1]. Event log consists of a trace identifier, event or activity that is executed, and a timestamp providing information about what time the activity has happened. Process mining is used for discovering and visualizing the models of the event logs. Based on the data and required outcomes, different process models can be generated. The most common process discovery methods are alpha mining, heuristic mining algorithm, fuzzy miner, inductive miner.

Alpha minor connects the selected data with process model discovery [29,30]. It creates process models based on the input event logs by showing relations between the different steps of the processes. For example, the research data selected for this project consists of subject identifier as a trace identifier of the event log, a cancer or breast cancer decease as an event and admission and discharge time as a timestamp. Therefore, it visualizes the activities in order by starting at one point (admission time) and ending in another (discharge time). If there is a patient who has been accepted to the hospital more than once it is presented as a sequence of events. Alpha minor generates a petri net. An example of a petri net with breast cancer data can be seen in *Figure 16*.

The heuristic mining algorithm considers the order of the events based on the timestamps and constructs a dependency graph [1]. It is used to operate with noisy data by reckoning with the frequency. This approach makes the process more robust than the others. In addition, it searches for relations between long-term dependencies.

Fuzzy minor can deal with spaghetti processes which are unstructured processes and have a lot of variability in terms of activities [1]. This process minor constructs hierarchical models. The activities that have lower frequencies are considered as subprocesses. It is used for structuring roadmaps or traffic connections, and provides information about the paths, activities, and frequency.

Inductive mining techniques can maneuver infrequent behavior of extensive incomplete event logs [1]. It can handle congestion in the data analysis that can cause slower process. Moreover, the inductive minor supports large and noisy event logs. It can ensure formal correctness criteria like rediscovering the initial model due to flexibility and scalability.

The activities related to process mining start with an initialization step which consists of planning the idea and extraction of data, after which analysis iterations need to be constructed [1]. The data for this research is processed by using the CRISP-DM model user guide. Once the data is cleaned, different process mining techniques are applied so that it can be properly evaluated. Moreover, some of the techniques are machine learning extensions as the goal of this project is to extend previous work in terms of data quality and machine learning. If the evaluation of the data is sufficient and represents accurate and appropriate visualizations, the results are summarized. However, if this is not the case because for example, some inconsistencies are found, the data is processed again. The last activity of process mining is implementation done by process improvement. Consequently, the models can be analyzed and evaluated. For the evaluation of the performance of various process mining techniques in this research, ProM Software is used. ProM Software is an extensible open-source framework that supports a variety of process mining techniques [1]. It visualizes the input data so that the user can analyze and interpret it. XES, MXML, and CSV files can be loaded in the software. To prepare the model, the data is imported as CSV file in the platform, after which it can be converted to XES file. Then, the process mining techniques can be applied, and the visualizations of the modeled data can be evaluated.

6. Results

In this section, process mining techniques in the software ProM will be applied on cancer and breast cancer data, after which analysis and evaluation of the data outcomes will be made.

6.1. Data Formatting

In both tables for cancer and breast cancer, the admission and discharge dates were inconsistent. After each year there was additional information in the form of two extra digits which was not relevant to be

used in the event log. Therefore, these inappropriate digits have been removed to avoid faulty process models. After this process, it was noticed that one admission date was incompatible - 29th February 2021. Consequently, the row with the subject admitted on this date has also been removed. After applying modeling techniques on the final version of the data, there were found more inconsistencies in the dates. It was noticed that there were patients admitted to the hospital at some time and being discharged earlier than the admission time. In total, there were 15 similar cases which were cleaned from the dataset. Afterwards, the data was integrated and modelled for further evaluation. There were no more fallacies discovered.

Referring to the methodology of CRISP-DM model and data preparation techniques, an answer to RQ2 can be provided. It is important to ensure that the data is accurate, reliable, and unbiased so that the results are correctly evaluated without any misleading paths. The observations from the data analysis of cancer- and breast cancer- related data extracted from MIMIC-III database are helping data quality improvement by following the CRISP-DM model.

6.2. Inductive Visual Minor for Cancer Data

Once the data was checked for consistency, it was imported in ProM as a CSV file. Consequently, it was converted to XES file where the case column was Subject_ID and the even column was the Short_Title of the disease so that it can be used by the software. The timestamps that are imported in ProM are admission time as a starting time and discharge time as completion time in and from the hospital.

In *Figures 7* a visualization of Visual Inductive Minor of the event log of cancer dataset is presented. The yellow circles represent the patients who are admitted to the hospital. Each patient visits their path based on the cancer disease they suffer from, after which they leave the hospital. In addition, it can be observed that 445 patients have been admitted to the hospital, and all of them discharged. There are no patients who have left the hospital in the middle of their treatment such as deciding to continue it in another hospital or deciding to stop it.

The little red dotted loops at the source and sink places of the process model are log moves that are present in the event log but according to the discovered model, they are redundant. Log moves are appearing because *Figure 7* is showing the abstracted version of the event log using a miner fraction of activities (by using activities and paths sliders) that are not perceivable in the model but are present in the log and the model is rejecting those. Moreover, the activities slider sets the fraction of included activities from the event log on which a discovery algorithm is applied. Activity slider allows abstraction of the event log from 0 to 1. Where 0 gives maximum abstraction and 1 gives none. For example, if 0.44 is selected on the activity slider, all events corresponding to the activities that occur more than 0.4 times the occurrence of the most-occurring activity, would be included. Additionally, the paths slider controls the amount of noise filtration. If it is set to 1, no noise filtering is applied, while if it is set to 0, maximum noise filtering is applied. The default is 0.8, which corresponds to 1-0.8 = 0:2 noise. By selecting different activities slider settings, the paths selector is also changed which is affecting the results. The top line represents all 428 activities which are not visualized because the activities slider is set to 0.028.

After evaluating the model in *Figure 7*, data inconsistencies were found. Consequently, the data was cleaned after which it was modeled one more time. The model in *Figure 8* shows an application of inductive visual minor on a prepared cancer dataset. The activities slider is set to 0.028 in both figures so that a comparison can be made. Firstly, it can be noticed that the patients have decreased by 15 in *Figure 8* which means that they are 430. As the activities slider is set to an exceptionally low value, the top line

represents 414 cancer diseases which are not visualized. Moreover, the visualized disease types are 15, while in *Figure 7* they are 16 due to the higher number of hospital admissions. Another observation is that the little red dotted loops at the source and sink places of the process model are missing in the model of the cleaned data. The reason could be that in *Figure 8* only paths are visualized while in *Figure 7* there are also deviations.

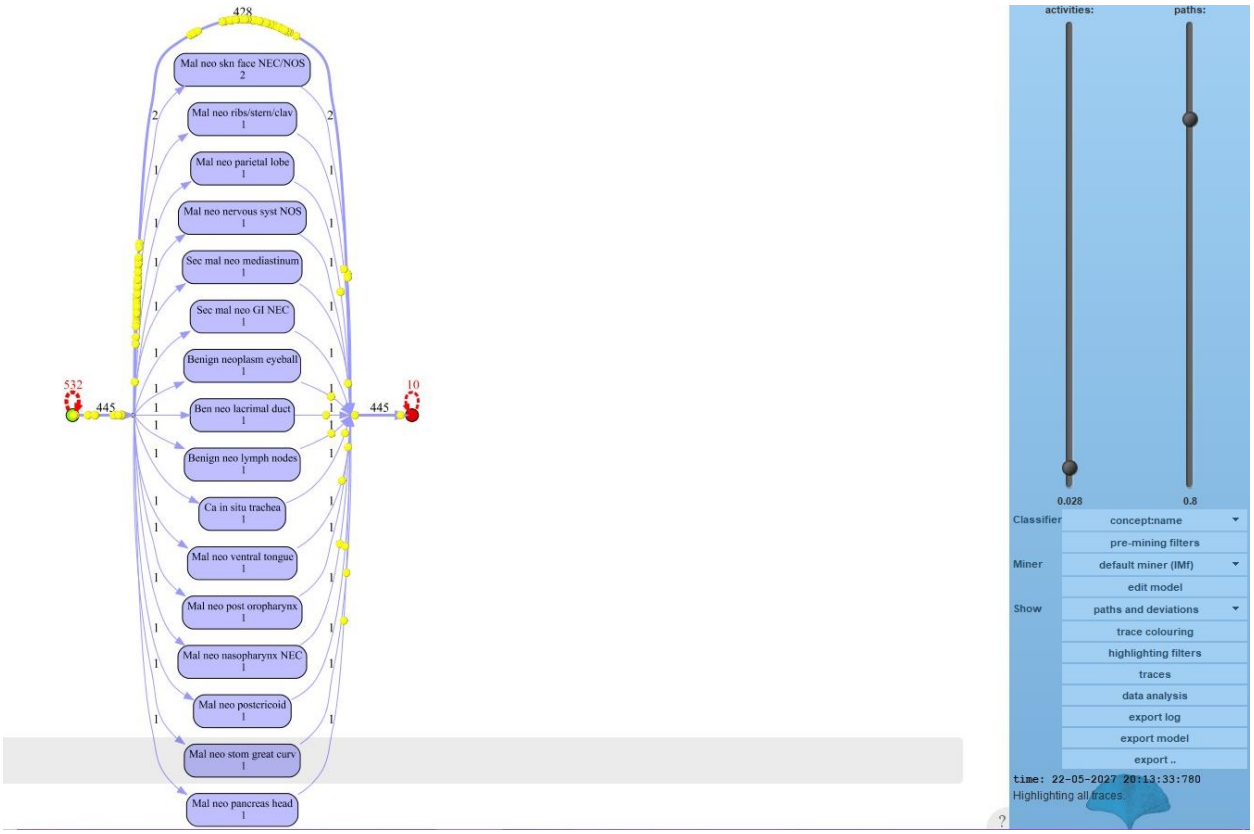


Figure 7. Inductive Visual Minor – Cancer dataset with invalid data.

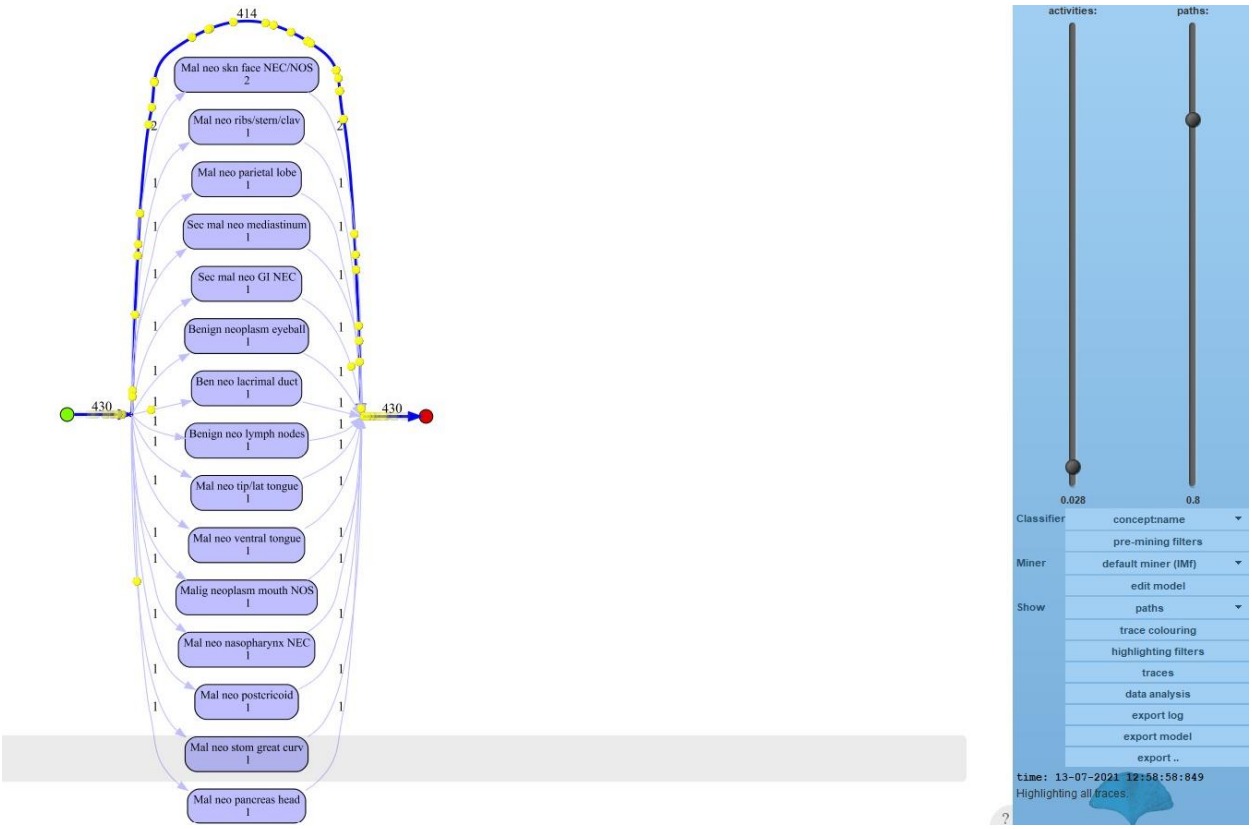


Figure 8. Inductive Visual Minor – Cancer dataset with cleaned data.

In *Figure 9*, a closer screenshot of the inductive visual miner with activities slider set to 1, is represented. Overall, this means that all patients and cancer disease types are visualized. The full diagram of inductive visual minor of invalid cancer dataset can be seen in *Figure 22* in Appendix C, as well as a model with activities slider set to 0.127 in *Figure 21*. Similarly, there can be found the full diagram of inductive visual minor of cleaned cancer dataset with activities slider set to 0.127 and to 1, in *Figure 23* and *Figure 24*, respectively.

In *Figure 9*, it can be noticed that there is a complex patient who has been admitted to the hospital for eleven different cancer types. The reason could be that cancer is growing rapidly and affecting other areas of their body. Moreover, it can be observed an overlap of patient's admissions. While they are staying in the hospital for one cancer type - secondary malignant neoplasm of kidney from 7th February 2021 until 21st March 2021, they have been registered for a second one - secondary malignant neoplasm of brain and spinal cord, on 1st of March 2021 until 5th of March 2021.

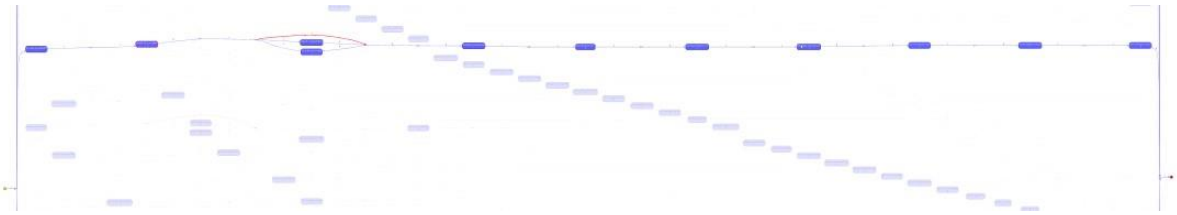


Figure 9. Patient with extreme amount of hospital admissions and discharges.

In addition, the data analysis of the inductive visual minor of cleaned cancer data are shown in *Figure 10*. It provides details about the number of completion events. It shows that the maximum number of cancer diseases because of which a patient has visited the hospital in the selected period is eleven. The reason is that this patient has made multiple doctor appointments or has been emergently admitted to the hospital regarding the different cancer types. On the contrary, the minimum number of diseases per patient is one, and it occurs in most of the patients which is the reason why the median is equal to 1. Additionally, there are 430 traces containing a cancer disease attribute in total.

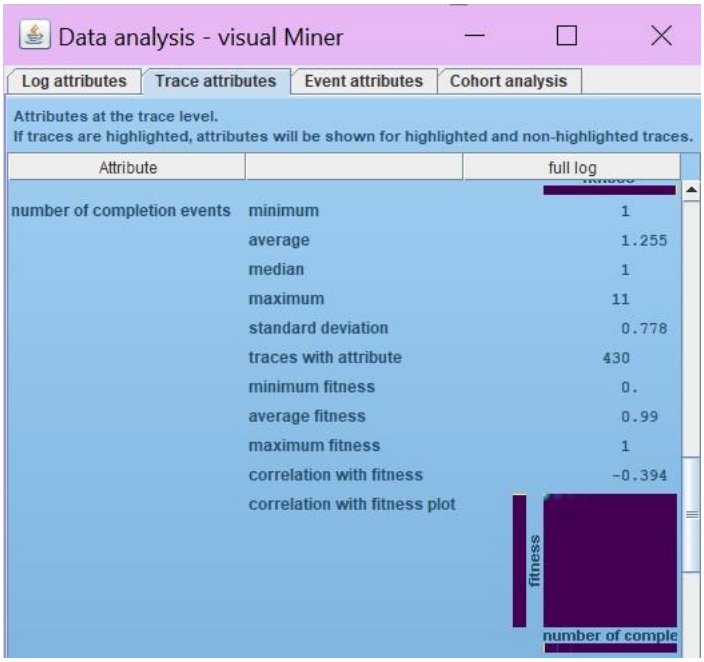


Figure 10. Data analysis of number of completion events.

Whereas *Figure 7* represents the paths and deviations of the models, in *Figure 11* and *Figure 12*, paths and service times can be observed. Service time represents the amount of time a patient has stayed in the hospital. On the visualization the more saturated the red is, the more time the patients have stayed at the hospital.

In *Figure 11*, there is one excessively light box with a disease in which instead of service time, a dash is written. This means that the time of stay cannot be calculated. Therefore, the model shows that the data is invalid. The reason for this could be because the admission time is after the discharge time. A closer view of the model with activities slider set to 1 can be seen in *Figure 25* in Appendix C.

After removing the inconsistent data, a model was processed in *Figure 12*. It can be noticed that the data is valid as there are no boxes without service time. In addition, models of the invalid and cleaned datasets with activities slider set to 0.101 can be seen in *Figure 26* and *Figure 27* in Appendix C, respectively.

These models were useful for this type of data input as they clearly represent where the data is faulty and needs to be fixed. Therefore, the process mining technique - Inductive Visual Minor, helps for quality improvement of the data by doing model evaluation.

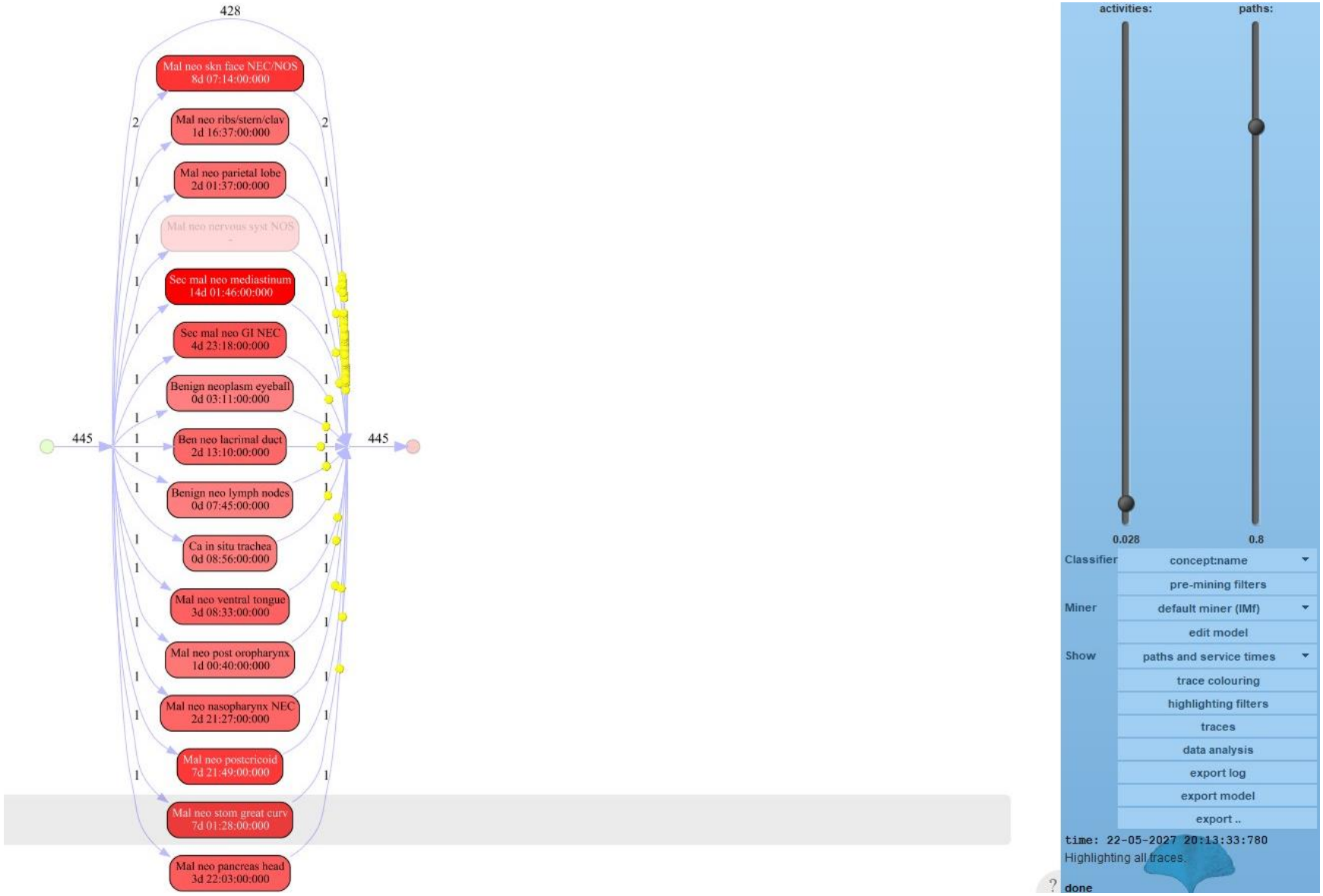


Figure 11. Inductive Visual Minor – Cancer dataset with invalid data - visualizing paths and service times.

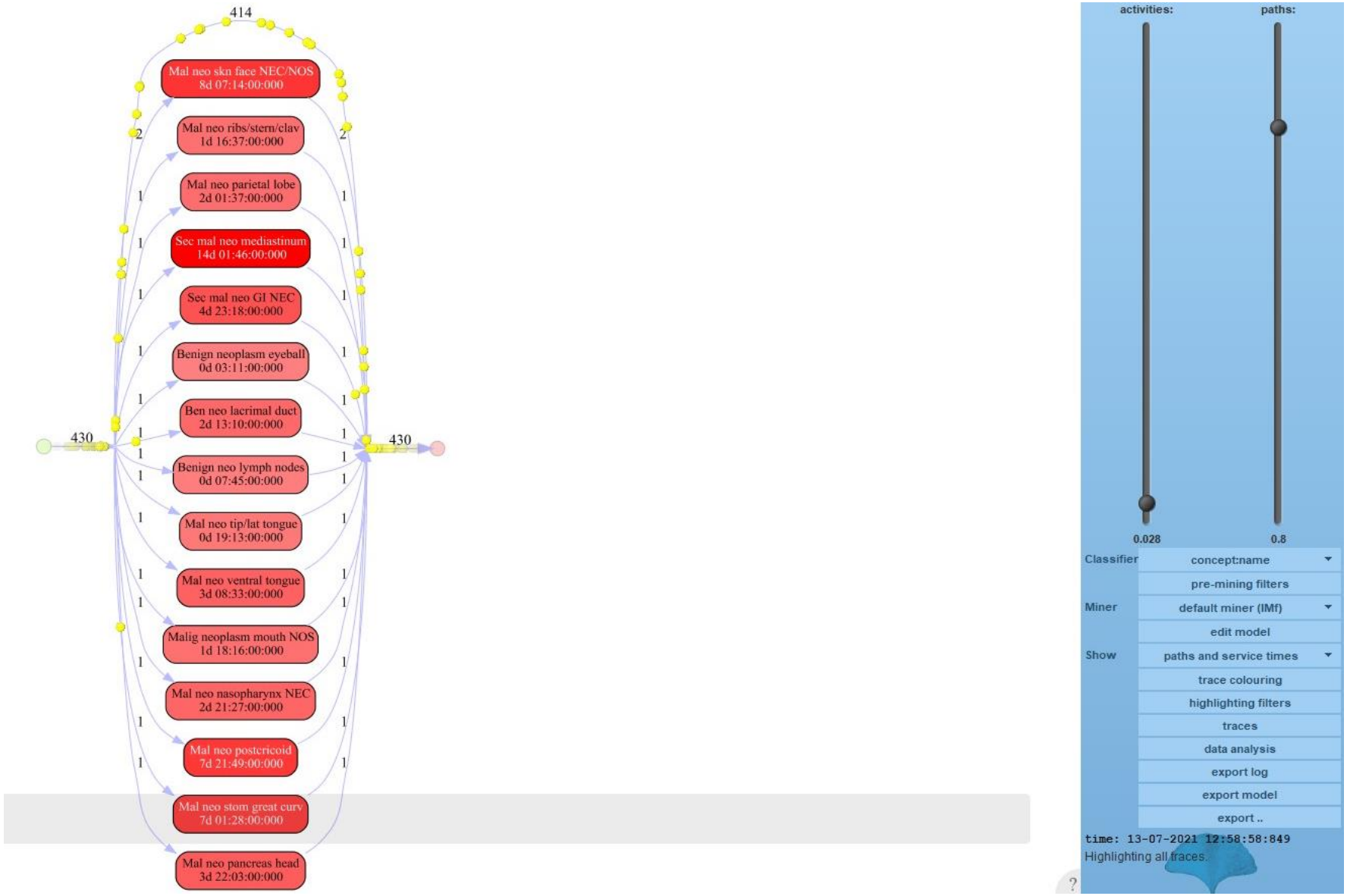


Figure 12. Inductive Visual Minor – Cancer dataset with cleaned data - visualizing paths and service times.

6.3. Inductive Visual Minor for Breast Cancer Data

In this subsection an inductive visual minor will be applied on breast cancer dataset which is extracted from the cancer dataset.

As the data is not large, the activities slider is set to 1. In *Figure 13* the model shows that there are 15 instances of breast cancer disease while there are 13 patients. There is one high risk patient who has been admitted to the hospital three times for different types of the disease. Correspondingly to the model with cancer data, it can be observed that all patients who have entered the hospital have been discharged. In *Figure 14* the cleaned data is loaded as an event log for the visual inductive miner. The only difference between the invalid and cleaned breast cancer data is the number of patients admitted to the hospital.

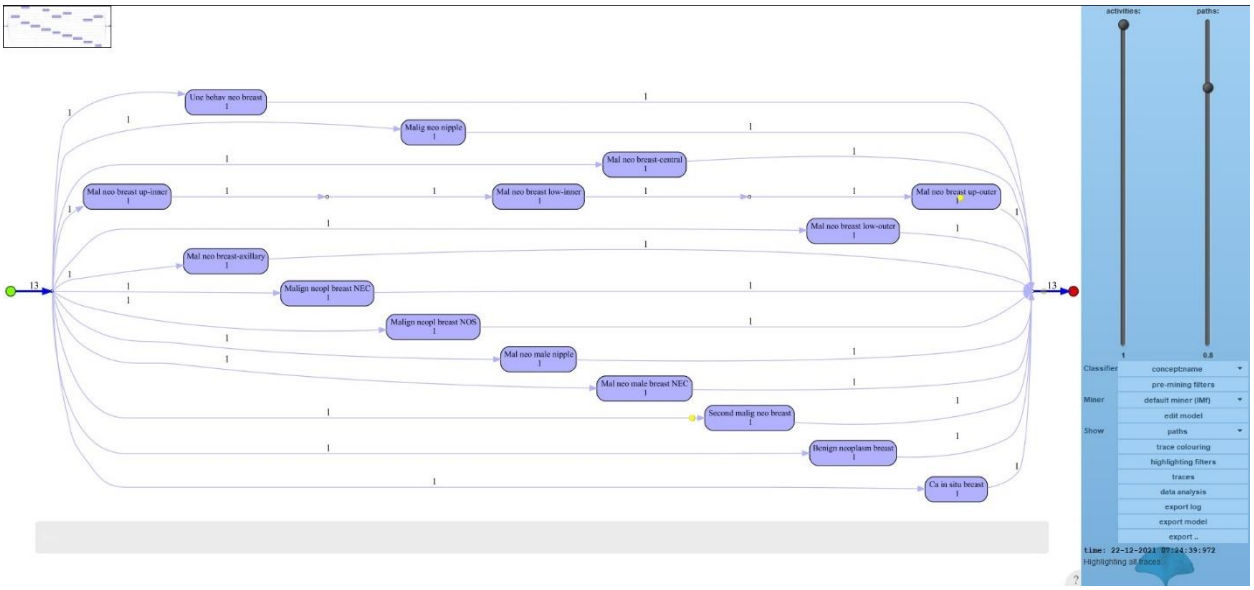


Figure 13. Inductive Visual Minor with invalid data - Breast Cancer Data.

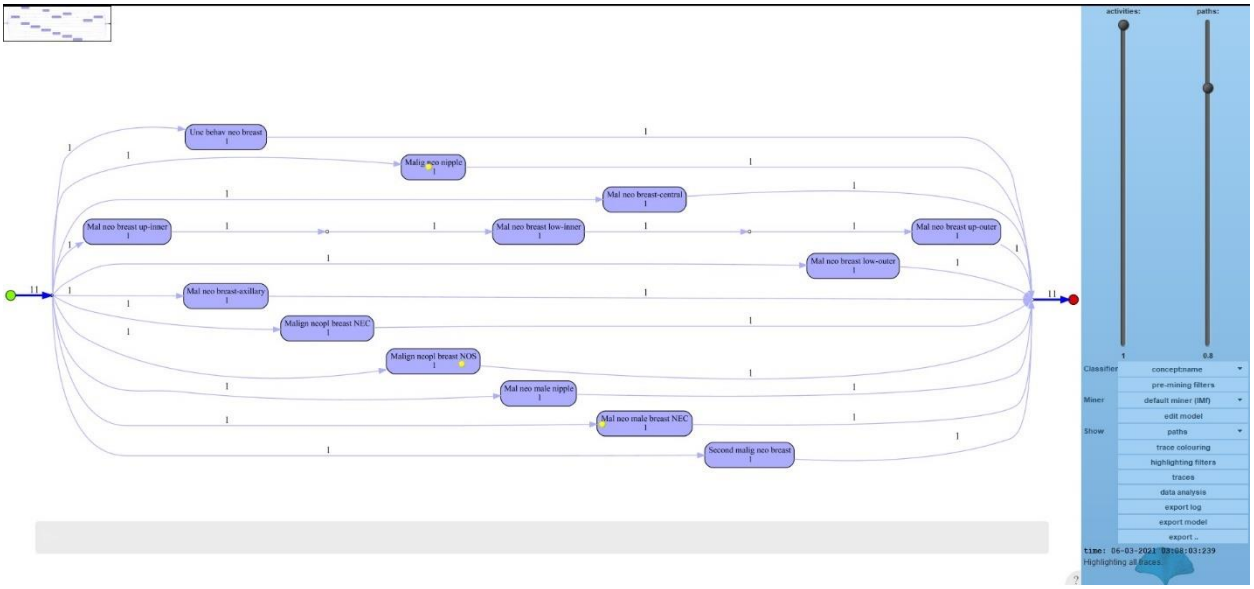


Figure 14. Inductive Visual Minor with cleaned data - Breast Cancer Data.

In *Figure 15* the paths and service times of the inductive visual minor are loaded. It can be clearly seen that there are two diseases which consist of faulty data. Therefore, the dataset has been cleaned so that the data is consistent and can provide qualitative models. Afterwards, a model of the prepared data has been loaded in *Figure 16*. It can be observed that the patients are 11, with two less than the model with invalid data. However, the complex patient remains the same.

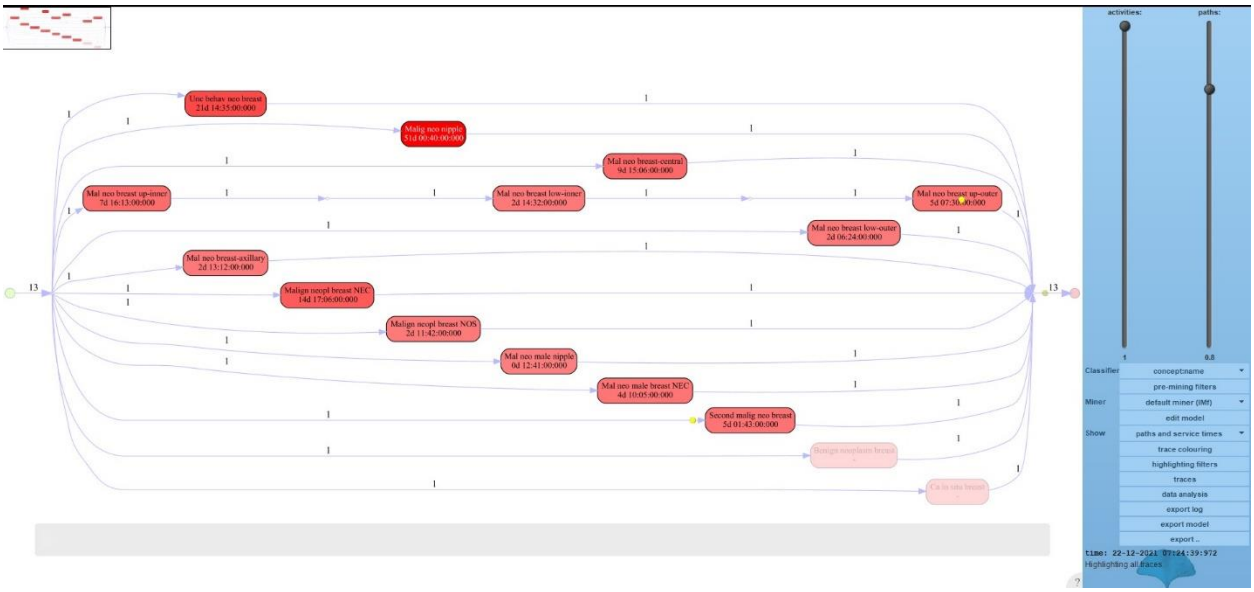


Figure 15. Inductive Visual Minor – Breast Cancer dataset with invalid data - visualizing paths and service times.

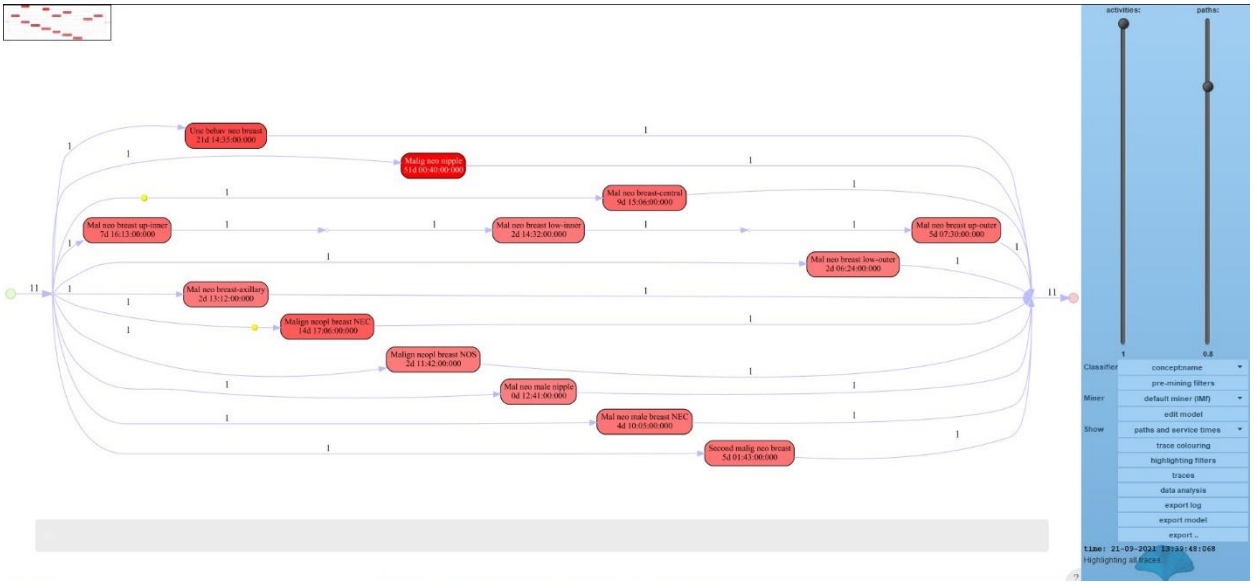


Figure 16. Inductive Visual Minor – Breast Cancer dataset with cleaned data - visualizing paths and service times.

6.4. Petri Net for Breast Cancer Data

In *Figure 17* there are 15 instances of breast cancer types and 13 patients. It can be observed that there is one patient who has been admitted to the hospital for three different breast cancer types: malignant neoplasm of upper-inner quadrant of female breast, malignant neoplasm of lower-inner quadrant of female breast, malignant neoplasm of upper-outer quadrant of female breast. Therefore, this can be considered as a high-risk patient in this dataset. The reason could be disease progression - when the cancer cells are growing out rapidly and affect different parts of the breast which causes more cancer variations. Moreover, there are 12 subjects being admitted at the hospital for one breast cancer decease. All patients have been discharged.

Figure 18 represents a petri net of cleaned breast cancer data. In comparison to *Figure 17* it can be noticed that the disease types decreased to eleven because the invalid data has been removed. However, the complex patient remains the same with the same number of admissions because of different breast cancer types as in was found in *Figure 17*. Overall, the petri nets of inconsistent, and cleaned and proceeded data are similar as there are no major discrepancies of the observed outcomes.

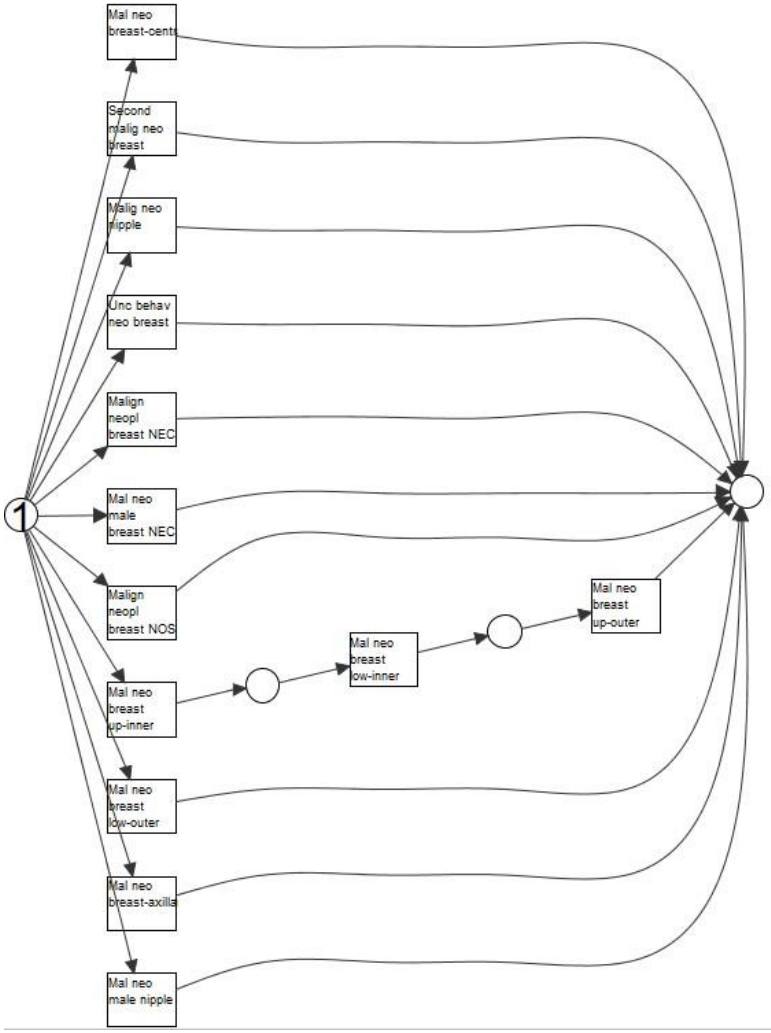


Figure 17. Petri Net of the Breast cancer dataset with invalid data.

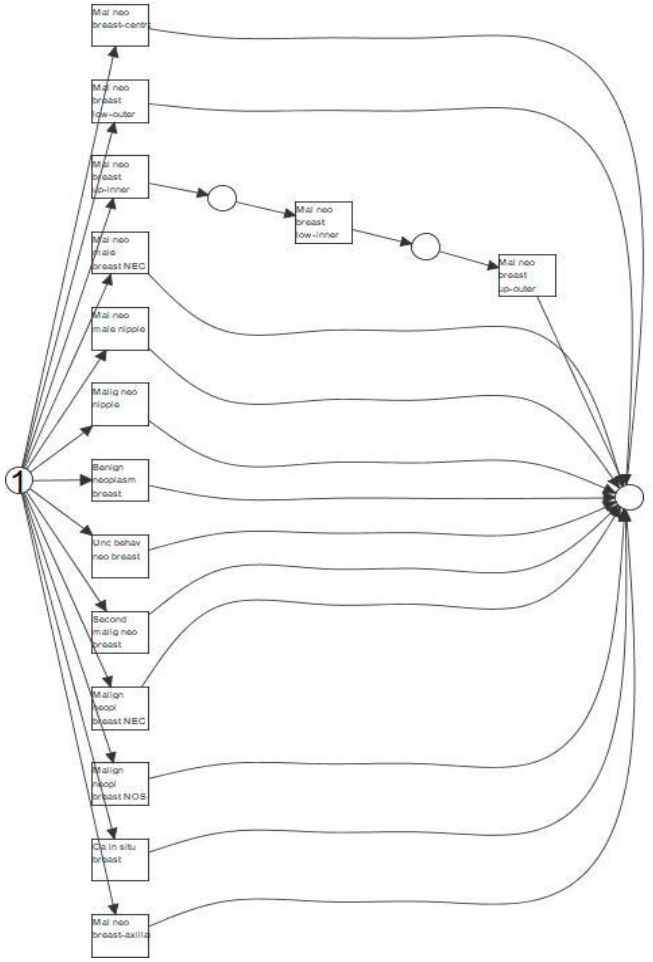


Figure 18. Petri Net of the Breast cancer dataset with cleaned data.

6.5. Machine learning techniques in ProM

For more complex processes in ProM, a petri net of the breast cancer data is used together with the event log of cancer data (a XES file). With this combination there can be applied machine learning plugin from ProM based on which it can be evaluated how machine learning and process mining techniques work together.

The discovery of process dataflow using the decision tree miner with ”Discovery of the Process Data-Flow (Decision-Tree Miner)” plugin is shown in *Figure 19*. This plugin is an enhanced version of “Decision Miner” which shows more precise data flow results and traces which are not conforming [31].

It can be observed that the root of the decision tree is the admission time. However, all child nodes end together in one point which represents the discharge time. Overall, only the breast cancer disease types are visualized. The reason for this could be because they are the only ones that overlap with the cancer disease. Moreover, the average fitness of the model is less than 0.5 which means that the observations of the event log are poorly represented. Consequently, the defined guards are not fully reliable and control-flow alignment needs to be generated. To fix the model and produce satisfying results, an “Align Log to The Model” plugin can be used. Nonetheless the input data does not contain any information related to decisions which can be made; hence, the control-flow alignment cannot be applied.

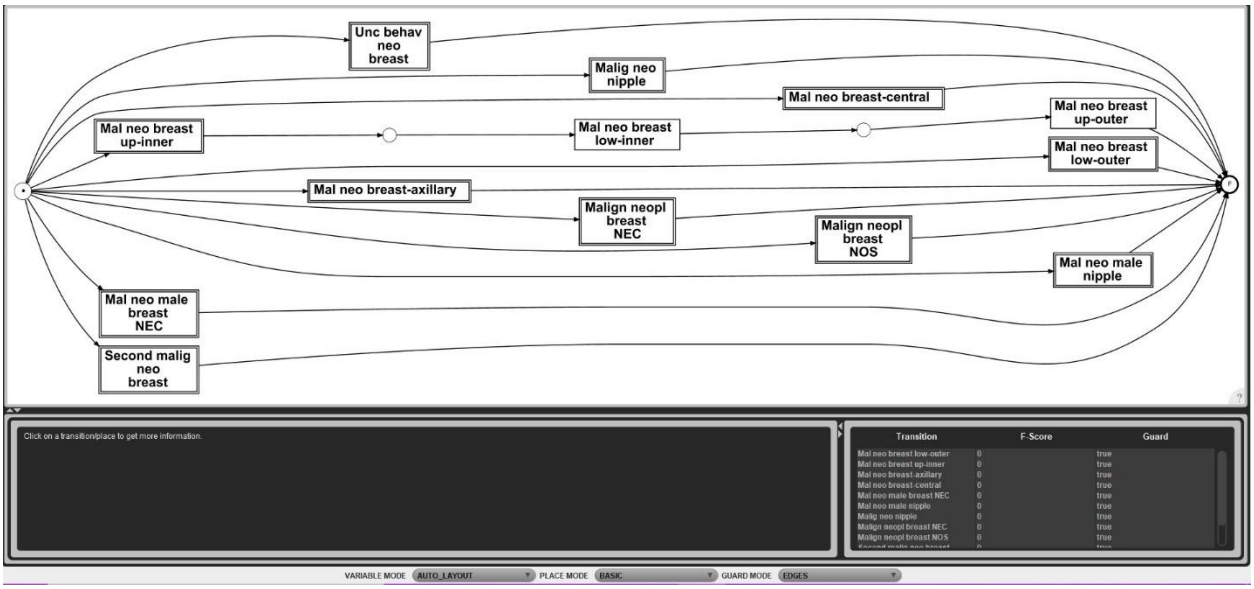
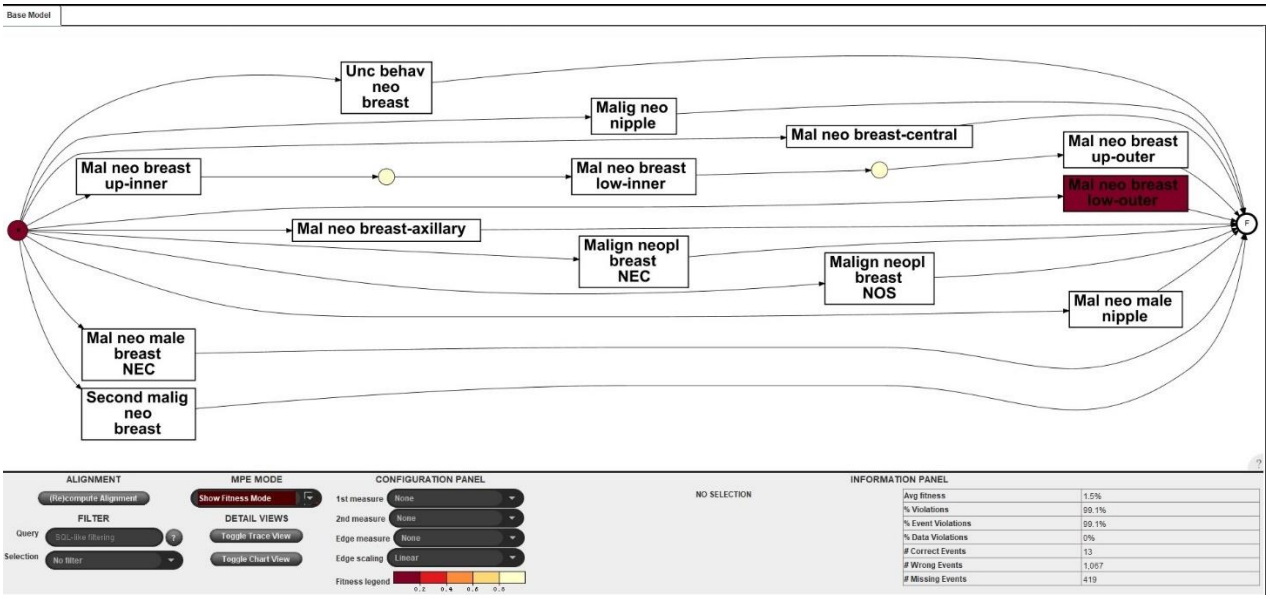


Figure 19. Discovery of Process Data-Flow (Decision-Tree Miner).

In *Figure 20* ”Multi-perspective Process Explorer” plugin of process mining is applied. It consists of different modes for integration of discovery and conformance checking techniques [31]. To be generated, it uses the same input as the plugin of decision tree miner. It processes the petri net of breast cancer dataset and the event log of cancer dataset. This could be one of the reasons why both models look so similar. The multi-perspective process explorer is set to show the fitness mode of the model. It can be observed that all events which are not visualized are going through one breast cancer disease and violates

its fitness value by 99.1%. The transitional fitness of the correct events is 0.2.There are 419 missing events and 13 correct ones which represent the patients who have been admitted to the hospital.



20. *Multi-perspective Process Explorer showing Fitness mode.*

Besides, in the model in *Figure 28* in the Appendix C, the same plug in is processed. The only difference is that Model mode is selected for the visualization. Thus, it can be observed that the first event happened on 6th April 2020 and the last registered one – on 19th November 2022.

Referring to RQ3, it can be concluded that machine learning techniques and ProM Software are working separately. However, their combination does not provide useful results for the selected cancer and breast cancer datasets for this research.

7. Conclusion

The main goal of this research project was to make a comparison of the care paths of subpopulations suffering from cancer and breast cancer by using data from MIMIC-III Critical Care database and expand the research in regard of data quality assurance and evaluation of machine learning techniques in combination with process mining. To reach this goal, a related work has been selected and used as a baseline following Barbara Kitchenham's et al. [6] methodology. Then, the CRISP-DM model was used for improving data quality of cancer and breast cancer datasets extracted from MIMIC-III database. Therefore, the data has been prepared and used for process models in ProM. Various process mining techniques are applied, including machine learning. Hence, the models were evaluated, and the results were used for providing answers to the three sub-questions defined for this project.

In conclusion, an answer to the main Research Question can be produced based on the findings from the research process. The main RQ is:

How can we assure quality evaluation of data extracted from public, provided online healthcare dataset - MIMIC-III Critical Care database, related to cancer and breast cancer?

Firstly, a systematic literature review needs to be made so that relevant papers to the topic are selected. Moreover, this approach can help the researchers to learn what has already been invented and build upon these guidelines. Secondly, following CRISP-DM model for data analysis methodology helped for the improvement of data quality techniques. Furthermore, after applying process mining techniques on the selected cancer and breast cancer datasets, it could be concluded that comparison between subpopulations is possible. Lastly, the ProM Software supports one machine learning algorithm under the form of plugin - ”Discovery of the Process Data-Flow (Decision-Tree Miner)”. Therefore, machine learning and process mining techniques can work together and process models for evaluation. However, the Decision-Tree Miner does not provide applicable results of the selected datasets for this research. More explanation why. For that reason, to make use of both process mining and machine learning processes, careful selection of data is necessary.

8. Discussion

Process mining is a relatively new technique used for analyzing processes from which researchers aim to identify subsets of bigger data. This research project gave insight into the qualitative evaluation of models generated by using process mining techniques. Moreover, it focuses on discovering how machine learning methods work with process mining.

By following CRISP-DM methodology, the data quality has been improved. The first reason for this is data understanding. Once the researcher is well acknowledged with the database, they can go to preparation phase and select the most relevant data for the research. In this case, it was subject identifier, type of cancer disease, admission, and discharge times. In addition, if there are any invalid data or inconsistencies noticed during the selection, the researcher can fix them, or if this is not possible - clean them. Therefore, the data is modelled, and for this research, by using process mining techniques. This way if there is any faulty data it can be easily noticed, which makes the analyst go to the previous step and fix or clean up the inconsistencies. This loop part of CRISP-DM model helps the most with making qualitative data which therefore contributes to more accurate results and conclusions.

In this research work, there were found inconsistencies in the admission and discharge date. There were 15 instances where the admission date was after the discharge from the hospital data. Therefore, they were removed from the dataset. Analogously, there could be data of patients who have stayed in the hospital for too long such as for years. In such situations when the data analyst is questioning the data, the data provider should be contacted and asked about the specific case which is unclear to the data analyst.

The application of process mining techniques on the selected for this project data was useful for subpopulation comparison. However, there was only one machine learning technique imported in the process mining software. It showed that process mining and machine learning can work together and produce interesting models. Nevertheless, the processed model was not applicable for the selected datasets.

Additional process mining techniques can be seen in Appendix D. In *Figure 29* Pom-Pom View process mining technique is loaded. It is generated by using the petri net of breast cancer dataset and event log of cancer dataset. In addition, in *Figure 30* a conformance checking of DPN technique can be seen. It is processed with the same input as Pom-Pom View, and without using approximate matches of the data. The yellow color means that there are parts which do not cover each other fully, and that another data input is needed or the current input data needs to be fixed or modified.

8.1. Limitations and Future work

One of the biggest limitations of this research is the lack of time. If there was more time, real-world data could be processed. Future research could expand on this one and focus on extracting data from a local or central hospital, for example, and compare the results to the ones extracted from MIMIC-III database. The outcome may show differences regarding data quality and data preparation.

Another limitation is the use of only one machine learning technique in combination with process mining. The only ML algorithm provided in ProM Software as plugin is Decision Tree. For future work, detailed research on machine learning methods in process mining can be made. The results could show that other ML techniques can be imported in ProM Software by writing a premade code or downloading them from a specific source. Therefore, the combination of these techniques with process mining may be more applicable to the selected data for this research. Hence, more data inconsistencies may be found or interesting cancer and breast cancer patients' subpopulations.

For future research on this topic, it can also be made evaluation of another selected dataset from MIMIC-III database. The data can be conscientiously selected so that Decision Tree algorithm in ProM can be applied, and the models can be evaluated providing compelling results.

Overall, the combination of machine learning and process mining techniques is a new field and there are a lot of places for further development.

References

[1] W. van der Aalst, *Process Mining*, Second Edition. Berlin Heidelberg 2011, 2016: Springer Heidelberg New York Dordrecht London, 2016. [Online]. Available: https://link.springer.com/content/pdf/10.1007/978-3-662-49851-4.pdf

[2] A. Fitri, Yasmin, A. Faiza, Bukhsh, P. De, and A. Silva, "Process Enhancement in Process Mining: A Literature Review." Accessed: Sep. 15, 2022. [Online]. Available: https://ris.utwente.nl/ws/portalfiles/portal/84624357

[3] A. Jain et al., "Overview and Importance of Data Quality for Machine Learning Tasks," *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Aug. 2020, doi: 10.1145/3394486.3406477.

[4] M. D. Cero *et al.*, "Evaluation of data quality in the Spanish EURECCA Esophagogastric Cancer Registry," *European Journal of Surgical Oncology*, vol. 47, no. 12, pp. 3081–3087, Dec. 2021, doi: 10.1016/j.ejso.2021.04.025.

[5] K. A. Kerr, T. Norris, and R. Stockdale, "The strategic management of data quality in healthcare," *Health Informatics Journal*, vol. 14, no. 4, pp. 259–266, Dec. 2008, doi: 10.1177/1460458208096555.

[6] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," Information and Software Technology, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009.

[7] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci Data*, vol. 3, no. 1, Art. no. 1, May 2016, doi: 10.1038/sdata.2016.35.

[8] S. V. Poucke et al., "Scalable Predictive Analysis in Critically Ill Patients Using a Visual Open Data Analysis Platform," PLOS ONE, vol. 11, no. 1, p. e0145791, Jan. 2016, doi: 10.1371/journal.pone.0145791.

[9] J. Wu, Y. Lin, P. Li, Y. Hu, L. Zhang, and G. Kong, "Predicting Prolonged Length of ICU Stay through Machine Learning," Diagnostics, vol. 11, no. 12, Art. no. 12, Dec. 2021, doi: 10.3390/diagnostics11122242.

[10] D. Chicco and G. Jurman, "Survival prediction of patients with sepsis from age, sex, and septic episode number alone," Scientific Reports, vol. 10, no. 1, Oct. 2020, doi: 10.1038/s41598-020-73558-3.

[11] F. Marazza et al., "Automatic Process Comparison for Subpopulations: Application in Cancer Care," International Journal of Environmental Research and Public Health, vol. 17, no. 16, p. 5707, Jan. 2020, doi: 10.3390/ijerph17165707.

[12] H. R. Darabi, D. Tsinis, K. Zecchini, W. F. Whitcomb, and A. Liss, "Forecasting Mortality Risk for Patients Admitted to Intensive Care Units Using Machine Learning," Procedia Computer Science, vol. 140, pp. 306–313, Jan. 2018, doi: 10.1016/j.procs.2018.10.313.

[13] P. L. Peissig et al., "Relational machine learning for electronic health record-driven phenotyping," Journal of Biomedical Informatics, vol. 52, pp. 260–270, Dec. 2014, doi: 10.1016/j.jbi.2014.07.007.

[14] N. Hong et al., “Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries,” Journal of Biomedical Informatics, vol. 99, p. 103310, Nov. 2019, doi: 10.1016/j.jbi.2019.103310.

[15] M. P. Pacheco et al., “Identifying and targeting cancer-specific metabolism with network-based drug target prediction,” EBioMedicine, vol. 43, pp. 98–106, May 2019, doi: 10.1016/j.ebiom.2019.04.046.

[16] L. T. Slater et al., “Towards similarity-based differential diagnostics for common diseases,” Computers in Biology and Medicine, vol. 133, p. 104360, Jun. 2021, doi: 10.1016/j.compbiomed.2021.104360.

[17] D. Zhao and C. Weng, “Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction,” Journal of Biomedical Informatics, vol. 44, no. 5, pp. 859–868, Oct. 2011, doi: 10.1016/j.jbi.2011.05.004.

[18] C.-M. Chao, Y.-W. Yu, B.-W. Cheng, and Y.-L. Kuo, “Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine, Logistic Regression and Decision Tree,” Journal of Medical Systems, vol. 38, no. 10, Aug. 2014, doi: 10.1007/s10916-014-0106-1.

[19] A. Alexandridis and E. Chondrodima, “A medical diagnostic tool based on radial basis function classifiers and evolutionary simulated annealing,” Journal of Biomedical Informatics, vol. 49, pp. 61–72, Jun. 2014, doi: 10.1016/j.jbi.2014.03.008.

[20] S. Famouri, L. Morra, L. Mangia, and F. Lamberti, “Breast Mass Detection With Faster R-CNN: On the Feasibility of Learning From Noisy Annotations,” IEEE Access, vol. 9, pp. 66163–66175, 2021, doi: 10.1109/ACCESS.2021.3072997.

[21] A. Rahimi and M. Gönen, “Efficient Multitask Multiple Kernel Learning With Application to Cancer Research,” IEEE Transactions on Cybernetics, vol. 52, no. 9, pp. 8716–8728, Sep. 2022, doi: 10.1109/TCYB.2021.3052357.

[22] K. Alghatani, N. Ammar, A. Rezgui, and A. Shaban-Nejad, “Precision Clinical Medicine Through Machine Learning: Using High and Low Quantile Ranges of Vital Signs for Risk Stratification of ICU Patients,” IEEE Access, vol. 10, pp. 52418–52430, 2022, doi: 10.1109/ACCESS.2022.3175304.

[23] N. Emaminejad et al., “Fusion of Quantitative Image and Genomic Biomarkers to Improve Prognosis Assessment of Early Stage Lung Cancer Patients,” IEEE Transactions on Biomedical Engineering, vol. 63, no. 5, pp. 1034–1043, May 2016, doi: 10.1109/TBME.2015.2477688.

[24] B. H. Cho, H. Yu, J. Lee, Y. J. Chee, I. Y. Kim, and S. I. Kim, “Nonlinear Support Vector Machine Visualization for Risk Factor Analysis Using Nomograms and Localized Radial Basis Function Kernels,” IEEE Transactions on Information Technology in Biomedicine, vol. 12, no. 2, pp. 247–256, Mar. 2008, doi: 10.1109/TITB.2007.902300.

[25] S. Liu, B. Fu, W. Wang, M. Liu, and X. Sun, “Dynamic Sepsis Prediction for Intensive Care Unit Patients Using XGBoost-Based Model With Novel Time-Dependent Features,” IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 8, pp. 4258–4269, Aug. 2022, doi: 10.1109/JBHI.2022.3171673.

[26] A. P. Kurniati, G. Hall, D. Hogg, and O. Johnson, “Process mining in oncology using the MIMIC-III dataset,” Journal of Physics: Conference Series, vol. 971, p. 012008, Mar. 2018, doi: 10.1088/1742-6596/971/1/012008.

[27] Wirth R, J. Hipp, “CRISP-DM: Towards a standard process model for data mining,” In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, vol. 1, pp. 29-39, 11 Apr. 2000.

[28] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, & R. Wirth, "The CRISP-DM user guide," 4th CRISP-DM SIG Workshop in Brussels in March, vol. 1999. Sn, 18 Mar. 1999.

[29] Y. Amelia Effendi and R. Sarno, “Conformance Checking Evaluation of Process Discovery Using Modified Alpha++ Miner Algorithm,” IEEE Xplore, Sep. 01, 2018. https://ieeexplore.ieee.org/document/8549770.

[30] S. Weerapong, P. Porouhan, and W. Premchaiswadi, “Process mining using α-algorithm as a tool (A case study of student registration),” IEEE Xplore, Nov. 01, 2012. https://ieeexplore.ieee.org/document/6408558.

[31] F. A. Yasmin, R. Bemthuis, M. Elhagaly, and F. A. Bukhsh, “A Process Mining Starting Guideline for Process Analysts and Process Owners: A Practical Process Analytics Guide using ProM”.

Appendix

A. Data Extraction from Literature for the main Research Question.

Serial No.	Authors	Title	Year	Source title	Cited by	Affiliations	Keywords	Document type
8	Sven Van Poucke, Zhongheng Zhang, Martin Schmitz, Milan Vukicevic, Margot Vander Laenen, Leo Anthony Celi, Cathy De Deyne (Sven et al., 2016)	Scalable Predictive Analysis in Critically Ill; Patients Using a Visual Open Data Analysis Platform	2016	PLOS ONE	38	Department of Anesthesiology, Intensive Care, Emergency Medicine and Pain Therapy, Ziekenhuis OostLimburg, Genk, Belgium; Department of Critical Care Medicine, Jinhua Hospital of Zhejiang University, Zhejiang, P.R. China; RapidMiner GmbH, Dortmund, Germany; Department of Organizational Sciences, University of Belgrade, Belgrade, Serbia; MIT Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States; Limburg Clinical Research Program, Faculty of Medicine, University Hasselt UH, Hasselt, Belgium		Research article
9	J. Wu, Y. Lin, P. Li, Y. Hu, L. Zhang, and G. Kong	Predicting Prolonged Length of ICU Stay through Machine Learning	2021	PubMed Diagnostics, MDPI	1	National Institute of Health Data Science, Peking University, Beijing 100191, China; Advanced Institute of Information Technology, Peking University, Hangzhou 311215, China;	prolonged length of ICU stay; machine learning; clinical decision rules; medical informatics	Research Article

						Department of Medicine and Therapeutics, LKS Institute of Health Science, The Chinese University of Hong Kong, Hong Kong, China; Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China; Medical Informatics Center, Peking University, Beijing 100191, China; Renal Division, Department of Medicine, Peking University First Hospital, Peking University Institute of Nephrology, Beijing 100034, China		
10	Davide Chicco and Giuseppe Jurman	Survival prediction of patients with sepsis from age, sex, and septic episode number alone	2020	Nature, Scientific Reports	11	Krembil Research Institute, Toronto, ON, Canada; Fondazione Bruno Kessler, Trento, Italy.		Research Article
11	Francesca Marazza, Faiza Allah Bukhsh, Jeroen Geerdink, Onno Vijlbrief, Shreyasi Pathak, Maurice van Keulen, and Christin Seifert	Automatic Process Comparison for Subpopulations: Application in Cancer Care	2020	International Journal of Environmental Research and Public Health	3	Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7522 NB Enschede, The Netherlands; Hospital Group Twente (ZGT), 7555 DL Hengelo, The Netherlands	process mining; process comparison; quality control; cancer types; breast cancer care; MIMIC database	Journal Article
12	Hamid R. Darabi, Daniel Tsinis, Kevin Zecchini, Winthrop F.	Forecasting Mortality Risk for Patients Admitted to Intensive	2018	Elsevier B.V.; Procedia Computer	17	Remedy Partners, 5 Penn Plaza, Floor 6, New York City, NY 10001, U.S.A.	Mortality Risk Prediction, Machine Learning in Healthcare,	Journal Article

	Whitcomb, Alexander Liss	Care Units Using Machine Learning		Science 140 (2018) 306–313			Predictive Models in Healthcare, Electronic Health Records	
13	Peggy L. Peissig, Vitor Santos Costa, Michael D. Caldwell, Carla Rottscheit, Richard L. Berg, Eneida A. Mendonca, David Page	Relational machine learning for electronic health record-driven phenotyping	2014	Elsevier	39	Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI, USA; DCC-FCUP and CRACS INESC-TEC, Department de Ciência de Computadores, Universidade do Porto, Portugal; Department of Surgery, Marshfield Clinic, Marshfield, WI, USA; Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, USA; Department of Pediatrics, University of Wisconsin–Madison, USA; Department of Computer Sciences, University of Wisconsin–Madison, USA	Machine learning, Electronic health record, Inductive logic programming, Phenotyping, Relational machine learning	Journal Article
14	Na Hong, Andrew Wen, Daniel J. Stone, Shintaro Tsuji, Paul R. Kingsbury, Luke V. Rasmussen, Jennifer A. Pacheco, Prakash Adekkanattu, Fei Wang, Yuan Luo, Jyotishman Pathak, Hongfang	Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries	2019	Elsevier	13	Mayo Clinic, Rochester, MN, USA; Northwestern University Feinberg School of Medicine, Chicago, IL, USA; Weill Cornell Medicine, New York City, NY, USA	Clinical phenotyping, HL7 Fast Healthcare Interoperability Resources (FHIR), Electronic Health Records (EHRs), Natural language processing, Algorithm portability	Journal Article

	Liu, Guoqian Jiang							
15	Maria Pires Pacheco, Tamara Bintener, Dominik Ternes, Dagmar Kulms, Serge Haan, Elisabeth Letellier, Thomas Sauter	Identifying and targeting cancer-specific metabolism with network-based drug target prediction	2019	Elsevier; EBioMedicine	31	Life Sciences Research Unit, University of Luxembourg, Esch-Alzette, Luxembourg; Experimental Dermatology, Department of Dermatology, Technical University Dresden, Dresden, Germany; Center for Regenerative Therapies, Technical University Dresden, Dresden, Germany	Metabolic modelling, Cancer, Machine learning, Drug repurposing	Research paper
16	Luke T. Slater, Andreas Karwath, John A. Williams, Sophie Russell, Silver Makepeace, Alexander Carberry, Robert Hoehndorf, Georgios V. Gkoutos	Towards similarity-based differential diagnostics for common diseases	2021	Elsevier; Computers in Biology and Medicine 133 (2021)	5	College of Medical and Dental Sciences, Institute of Cancer and Genomic Sciences, University of Birmingham, UK; Institute of Translational Medicine, University Hospitals Birmingham, NHS Foundation Trust, UK; NIHR Experimental Cancer Medicine Centre, UK; NIHR Surgical Reconstruction and Microbiology Research Centre, UK; NIHR Biomedical Research Centre, UK; MRC Health Data Research UK (HDR UK) Midlands, UK; Computer, Electrical and Mathematical Sciences & Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Saudi Arabia; University Hospitals Birmingham	Semantic web, Ontology, Differential diagnosis, Mimic-III, Semantic similarity	Journal Article

						NHS Foundation Trust, Edgbaston, Birmingham, UK		
17	Di Zhao, Chunhua Weng	Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction	2011	Elsevier; Journal of Biomedical Informatics 44 (2011) 859–868	89	Department of Biomedical Informatics, Columbia University, New York, NY 10032, United States	Electronic Health Records, Pancreatic neoplasms, Text mining, Bayesian method	Journal Article
18	Cheng-Min Chao, Ya-Wen Yu, Bor-Wen Cheng, Yao-Lung Kuo	Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine, Logistic Regression and Decision Tree	2014	Springer Science+Business Media New York	48	Department of Business Administration, National Taichung University of Science and Technology, Taichung, Taiwan; National Yunlin University of Science and Technology, Douliu, Yunlin Country, Taiwan	Breast cancer, Support vector machine, Logistic Regression, C5.0 decision tree, 10-fold cross-validation	Journal Article
19	Alex Alexandridis, Eva Chondrodima	A medical diagnostic tool based on radial basis function classifiers and evolutionary simulated annealing	2014	Elsevier; Journal of Biomedical Informatics 49 (2014) 61–72	33	Department of Electronic Engineering, Technological Educational Institute of Athens, Agiou Spiridonos, Aigaleo 12210, Greece	Decision support systems, Evolutionary computation, Medical diagnosis, Neural networks, Radial basis function, Simulated annealing	Journal Article
20	Sina Famouri, Lia Morra, Leonardo Mangia, Fabrizio Lamberti	Breast Mass Detection With Faster R-CNN: On the Feasibility of Learning From Noisy Annotations	2021	IEEE	2	Department of Computer and Control Engineering, Politecnico di Torino, 10129 Turin, Italy	Computer aided diagnosis, faster R-CNN, machine learning noise, mammography, object detection	Journal Article

21	Arezou Rahimi, Mehmet Gönen	Efficient Multitask Multiple Kernel Learning With Application to Cancer Research	2021	IEEE	5	Graduate School of Sciences and Engineering, Koç University, Istanbul 34450, Turkey; Department of Industrial Engineering, College of Engineering, Koç University, Istanbul 34450, Turkey	Benders decomposition (BD), biochemical pathways, cancer stage, multiple kernel learning (MKL), multitask learning	Research Article
22	Khalid Alghatani, Nariman Ammar, Abdelmounaam Rezgui, Arash Shaban-Nejad,	Precision Clinical Medicine Through Machine Learning: Using High and Low Quantile Ranges of Vital Signs for Risk Stratification of ICU Patients	2022	IEEE	0	King Fahad Medical City, Riyadh 11525, Saudi Arabia; UTHSC-ORNL Center for Biomedical Informatics, Department of Pediatrics, College of Medicine, The University of Tennessee Health Science Center, Memphis, TN 38103, USA; School of Information Technology, Illinois State University, Normal, IL 61790, USA	Precision medicine, artificial intelligence, ICU patient monitoring, prediction models, biomedical informatics	Journal Article
23	Nastaran Emaminejad, Wei Qian, Yubao Guan, Maxine Tan, Yuchen Qiu, Hong Liu, Bin Zheng	Fusion of Quantitative Image and Genomic Biomarkers to Improve Prognosis Assessment of Early Stage Lung Cancer Patients	2015	IEEE	73	University of Oklahoma; Northeastern University; Guangzhou Medical University; School of Electrical and Computer Engineering, University of Oklahoma, Norman, OK, USA	Computer-aided diagnosis, fusion of image features and genomic biomarkers, prediction of lung cancer recurrence risk, quantitative image feature analysis, radiomics	Journal Article
24	Baek Hwan Cho, Hwanjo Yu,	Nonlinear Support Vector Machine	2008	IEEE	47	Department of Biomedical Engineering, Hanyang University,	Decision support systems, feature	Journal Article

	Jongshill Lee, Young Joon Chee, In Young Kim, Sun I. Kim	Visualization for Risk Factor Analysis Using Nomograms and Localized Radial Basis Function Kernels				Seoul 133-605, Korea; Department of Computer Science, University of Iowa, Iowa City, IA 52242 USA	selection, localized radial basis function (LRBF) kernel, nomograms, support vector machines (SVMs), visualization	
25	Shuhui Liu, Bo Fu, Wen Wang, Mei Liu, Xin Sun	Dynamic Sepsis Prediction for Intensive Care Unit Patients Using XGBoost-Based Model With Novel Time-Dependent Features	2022	IEEE Xplore	1	Chinese Evidencebased Medicine Center, West China Hospital, Sichuan University, Chengdu, Sichuan 610017, China; NMPA Key Laboratory for Real World Data Research and Evaluation in Hainan, Sichuan Center of Technology Innovation for Real World Data, Chengdu 610041, China	Boosting, data preprocessing, machine learning, prediction methods	Journal Article

B. Data Extraction from Literature for Research Questions 1, 2, 3.

Serial No.	Name	Description	Machine learning type (supervised or unsupervised)	Dataset type	Machine learning algorithm	Discussed data quality	Breast cancer (yes/no)	Comparison between datasets	Research gaps (future research direction)
8	Scalable Predictive Analysis in Critically Ill Patients Using a Visual Open Data Analysis Platform (Sven et al., 2016)	This paper shows processes for automatic building, parameter optimization, and evaluation of several predictive ML models integrated by using MIMIC-II database.	Supervised	MIMIC-II database	Decision Stump, Decision Tree, Naive Bayes, Logistic Regression, Random Forest, AdaBoost, Bagging, Stacking, Support Vector Machine with feature weighting and qualitative selection - Correlation, Gini Selection, Information Gain and ReliefF	No	No	No	A similar modeling and feature selection procedure could be used implementing features with no, unknown, or minimal interdependency to provide medical and financial liability.
9	Predicting Prolonged Length of ICU Stay through Machine Learning (Wu et al., 2021)	The goal of this study is to construct machine learning methods for predicting prolonged length of stay (pLOS) in intensive care units (ICU) among general ICU patients.	Supervised	single-center dataset: eICU Collaborative Research Database (eICU-CRD); MIMIC-III	Random forest (RF), Support vector machine (SVM), Deep learning (DL), and Gradient boosting decision tree (GBDT)	No	No	A brief comparison of eICU-CRD and MIMIC-III	ML explainability can be taken into consideration in future pLOS-ICU prediction model development.
10	Survival prediction of patients with sepsis from age, sex, and septic episode number alone (Chicco & Jurman, 2020)	In this paper a machine learning algorithm is provided to show prediction of survival within patient suffering from sepsis. The ML model uses three medical characteristics to make correct computations:	Supervised	Extracted data from Norwegian Patient Registry and Statistics Norway agency	Linear regression, support vector machine with linear kernel (linear SVM), Support vector machine with radial kernel (radial SVM), Gradient boosting, and Naïve Bayes	No	No	No	Further investigate the theme of the minimal clinical record for computational prediction of survival on other diseases such as cervical cancer, neuroblastoma, breast cancer, and

		sex, age, and septic episode number.							amyotrophic lateral sclerosis.
11	Automatic Process Comparison for Subpopulations: Application in Cancer Care (Marazza et al., 2020)	An automatic comparison of processes for different breast cancer patient populations is shown in this study. The process models are extracted from event of electronic health records (EHR).	Unsupervised	Electronic Health Records (EHR), MIMIC and Ziekenhuis Group Twente (ZGT)	No	Yes (in related work)	Yes	A process model comparison of MIMIC and ZGT datasets	Future experiments could evaluate the validity of the human ground truth for process model comparison. Also, make a qualitative investigation. In addition, future work might explore semantic embeddings as features for graph comparison.
12	Forecasting Mortality Risk for Patients Admitted to Intensive Care Units Using Machine Learning	The goal of this research is to provide an accurate prediction of mortality risk for patients in the health care system which can improve care quality and reduce costs.	Supervised	MIMIC III dataset	Gradient, boosted trees and deep neural networks	No	Yes (in related work)	No	Apply the model to a larger dataset of medical claims information to examine the performance of the two modeling techniques by incorporating semi-structured data, such as nursing notes.
13	Relational machine learning for electronic health record-driven phenotyping	In this research an evaluation of relational machine learning (ML) using inductive logic programming (ILP) is made in terms of contribution to identification of phenotype features from EHR records and	Supervised	International Classification of Diseases, Ninth Revision (ICD-9) coded EHR data	ILP, WEKA: PART, J48, and JRIP	No	Yes (in related work)	No	Future research is needed to examine grouping of rules and selection of subjects based on a combination of rule conditions, thereby combining the advantages of ILP and the general “rule-of-N”

		classification of patients at risk.							approach commonly used in phenotyping.
14	Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries	The goal of this study is to develop and evaluate a FHIR-based EHR phenotyping framework for identification of patients with obesity and multiple comorbidities using semi-structured discharge summaries.	Supervised	HL7 Fast Healthcare Interoperability Resources (FHIR)-based EHR, MIMIC-III dataset, i2b2 dataset	Logistic regression, support vector machine, decision tree, and random forest	Quality improvement (in the introduction)	No	Yes	In the future, FHIR-based terminology services should be established and leveraged to tackle the challenge although this is a non-trivial task. Moreover, enhancing the interpretability via methods such as either model-specific or model-agnostic, can be explored.
15	Identifying and targeting cancer-specific metabolism with network-based drug target prediction	The aim of this project was to develop efficient FASTCORMICS RNA-seq workflow to build 10,005 high-resolution metabolic models from the TCGA dataset to capture metabolic rewiring strategies in cancer cells.	Supervised	TCGA dataset	linear and radial support vector machines, and *rpart*	No	No	No	In the future, the rFASTCORMICS metabolic models can be used for drug repurposing, which consist in **finding new indications for already commercialized drugs**.
16	Towards similarity-based differential diagnostics for common diseases	In this study a similarity between patient-patient and patient-disease is classified aiming to set differential diagnosis of common disease based on a technique using unextracted text phenotypes from patients' profiles.	Unsupervised	MIMIC-III dataset	Semantic similarity scores	No	No	No	For future work, consider different automated and semi-automated methods of optimizing and curating text-derived patient phenotype profiles. Also, consider how the use of different text

									mining systems affects performance at semantic similarity tasks.
17	Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction	This paper proposes a method where the weighted BNI using PubMed knowledge and EHR data show remarkable accuracy improvement over existing representative methods for pancreatic cancer prediction.	Supervised	case-control dataset from PubMed and EHR: 98-sample dataset with pancreatic cancer patients, and 14,971-sample dataset of patients who did not have ICD-9 diagnosis of pancreatic cancer.	k-Nearest Neighbor (KNN) and Support Vector Machine (SVM) methods, excel primarily in a high-dimensional feature space, the Bayesian Network Inference (BNI) model	No	Mention that BMI has been use for breast cancer prediction.	No	To continuously improve the weighted BNI model for disease risk prediction: - highly accurate dataset is crucial, therefore, investigate the efficacy of weighing causal edges for improving the predictive accuracy of BNI - efficient discovery of unknown disease features Development, validation, and reuse of sophisticated phenotyping algorithms in the EHR
18	Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine, Logistic	The aim of the paper is to use data mining technology to establish a classification of breast cancer survival patterns and offer a treatment decision-making reference.	Supervised	Diseases Database	Support Vector Machine, Logistic Regression and C5.0 Decision Tree	No	Yes	No	For future research - compare the results of data mining across various studies, comparing the performance of different models.

	Regression and Decision Tree								
19	A medical diagnostic tool based on radial basis function classifiers and evolutionary simulated annealing	In this paper, a methodology for designing data-driven medical diagnostic tools, based on neural network classifiers is presented.	Supervised	Cardiotocography datasets (FHR, NSP), EEG Eye State, Pima Indian Diabetes, Thyroid ANN, Vertebral Column Datasets (3 Classes–2 Classes), Wisconsin Original Breast Cancer (WOBC)– Wisconsin Diagnostic Breast Cancer (WDBC)	Radial basis function (RBF) network, non-symmetric fuzzy means (NSFM) for RBF, Simulated Annealing (SA), Evolutionary simulated annealing (ESA), symmetric fuzzy means (SFM), standard SVM, Gaussian kernel function	No	Yes	Yes	For future work - exploiting the inherent potential of ESA for parallel implementation to design a distributed optimizer.
20	Breast Mass Detection With Faster R-CNN: On the Feasibility of Learning From Noisy Annotations	In this study the impact of noise on the training of object detection networks for the medical domain is researched, and how it can be mitigated by improving the training procedure.	Supervised	CBIS-DDSM (Curated Breast Imaging Subset of DDSM) dataset	Faster R-CNN, Non-Maximum Suppression (NMS), Region Proposal Network (RPN)	No	Yes	No	For future work: - conclusions should be validated in a real-life dataset - improving the quality of the regression - Investigating to which extent overfitting is reduced with larger dataset sizes

21	Efficient Multitask Multiple Kernel Learning With Application to Cancer Research	In this research multitask MKL method is used to discriminate early-stage and late-stage cancers using genomic data and gene sets and compare this algorithm against two other algorithms.	Supervised	Cancer Genome Atlas (TCGA)	Multitask multiple kernel learning (MKL), Cutting-Plane Algorithm for solving variable linearization master problem (VLMP), Cutting-Plane Algorithm for solving objective function linearization master problem (OLMP), BDForest Algorithm, Cutting-Plane Algorithm for Forest Formulation, BD Algorithm for Solving FMPt, Accelerated BD Algorithm for the Forest Formulation	No	No	Yes	Exact time-efficient algorithm for co-clustering problem can be developed.
22	Precision Clinical Medicine Through Machine Learning: Using High and Low Quantile Ranges of Vital Signs for Risk Stratification of ICU Patients	This study significantly improves the functionality of the initial intelligent remote patient monitoring (IRPM) framework by building three machine learning models for readmission, abnormality, and next-day vital sign measurements.	Supervised	MIMIC-III database	logistic regression (LR); linear discriminant analysis (LDA); random forest (RF); k-nearest neighbors (KNN); and support vector machine (SVM)	No	No	No	Exploring the impact of adding other relevant abnormality indicators such as cardiac problems and organ disorders. The patient abnormality model might be improved by including more models in the Intelligent ICU Patient Monitoring (IICUPM) module.
23	Fusion of Quantitative Image and Genomic Biomarkers to Improve Prognosis	This study aims to develop a new quantitative image feature analysis scheme and investigate its role along with two genomic biomarkers.	Supervised	Acquired from First Affiliated Hospital of Guangzhou Medical University,	Naïve Bayesian network-based classifier, simple Multilayer perceptron-based classifier, WEKA data mining software package, SMOTE synthetic data generation algorithm, leave-	No	No	No	· To develop more accurate and reliable CAD schemes for tumor segmentation. · Robustness of the reported results in this

	Assessment of Early Stage Lung Cancer Patients			Guangzhou, China	one-case-out validation method, ROCKIT program				study needs to be tested and validated.
24	Nonlinear Support Vector Machine Visualization for Risk Factor Analysis Using Nomograms and Localized Radial Basis Function Kernels	This study presents new ways for visualization of nonlinear classifiers and improvement of the interpretability of results while maintaining high prediction accuracy.	Supervised	University of California, Irvine repository, and Statlog dataset	Localized radial basis function (LRBF) kernel, RBF, ReliefF Method, recursive feature elimination (RFE) with an SVM (SVM-RFE), Sensitivity Analysis	No	No (usage of breast cancer dataset)	No	To apply the LRBF kernel method and the VRIFA to other problems in the medical domain.
25	Dynamic Sepsis Prediction for Intensive Care Unit Patients Using XGBoost-Based Model With Novel Time-Dependent Features	In this paper a machine learning algorithm is provided to show prediction of survival within patient suffering from sepsis. The ML model uses three medical characteristics to make correct computations: sex, age, and septic episode number.	Supervised	MIMIC-III database, PhysioNet Challenge 2019 datasets	linear regression, support vector machine with linear kernel (linear SVM), support vector machine with radial kernel (radial SVM), gradient boosting, and naïve Bayes	No	No	Yes	To find a way to impute the missing values instead of the linear interpolation and address the predicting issues brought by linear interpolation.

C. Process Mining Models

Figure 21. Inductive Visual Minor – Cancer dataset with invalid data and activities slider set to 0.127.

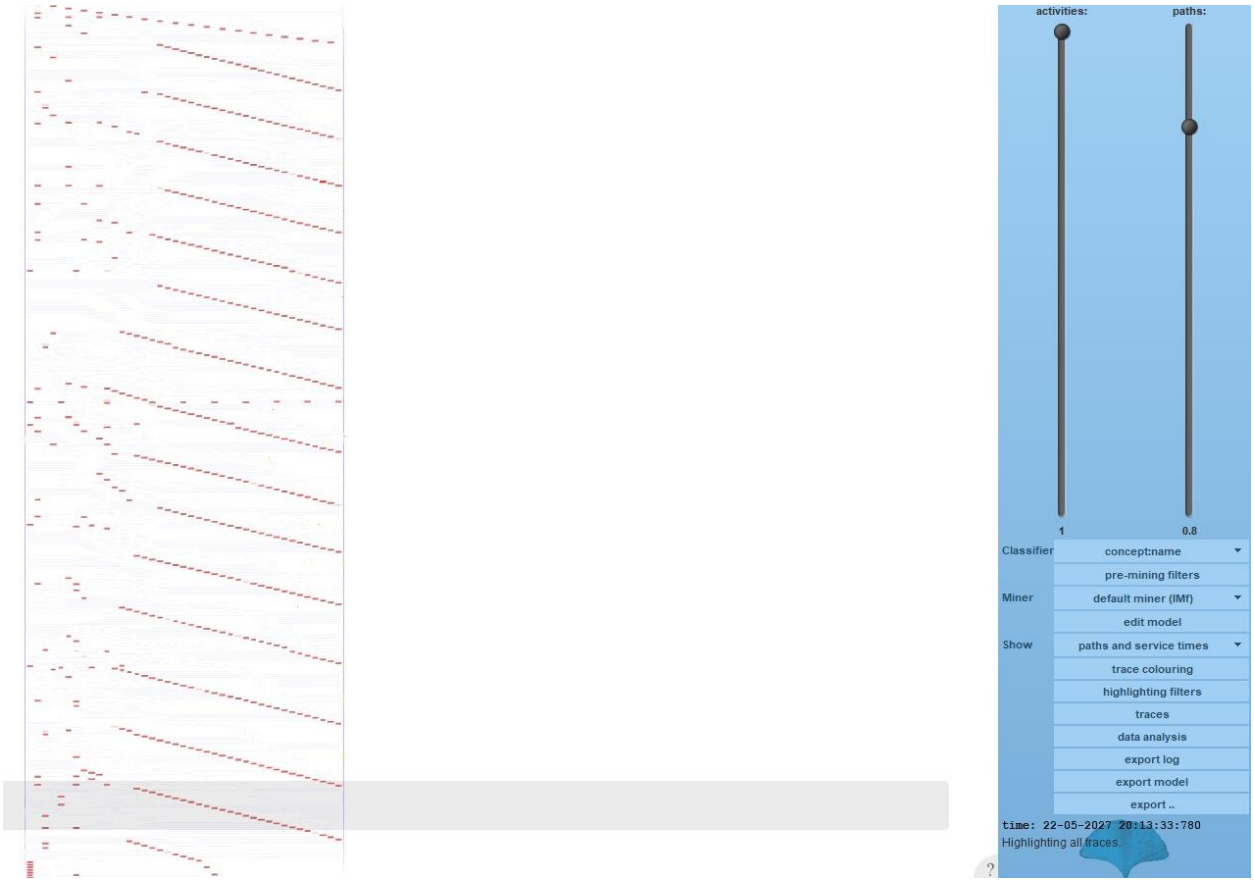


Figure 22. Inductive Visual Minor – Cancer dataset with invalid data and activities slider set to 1.

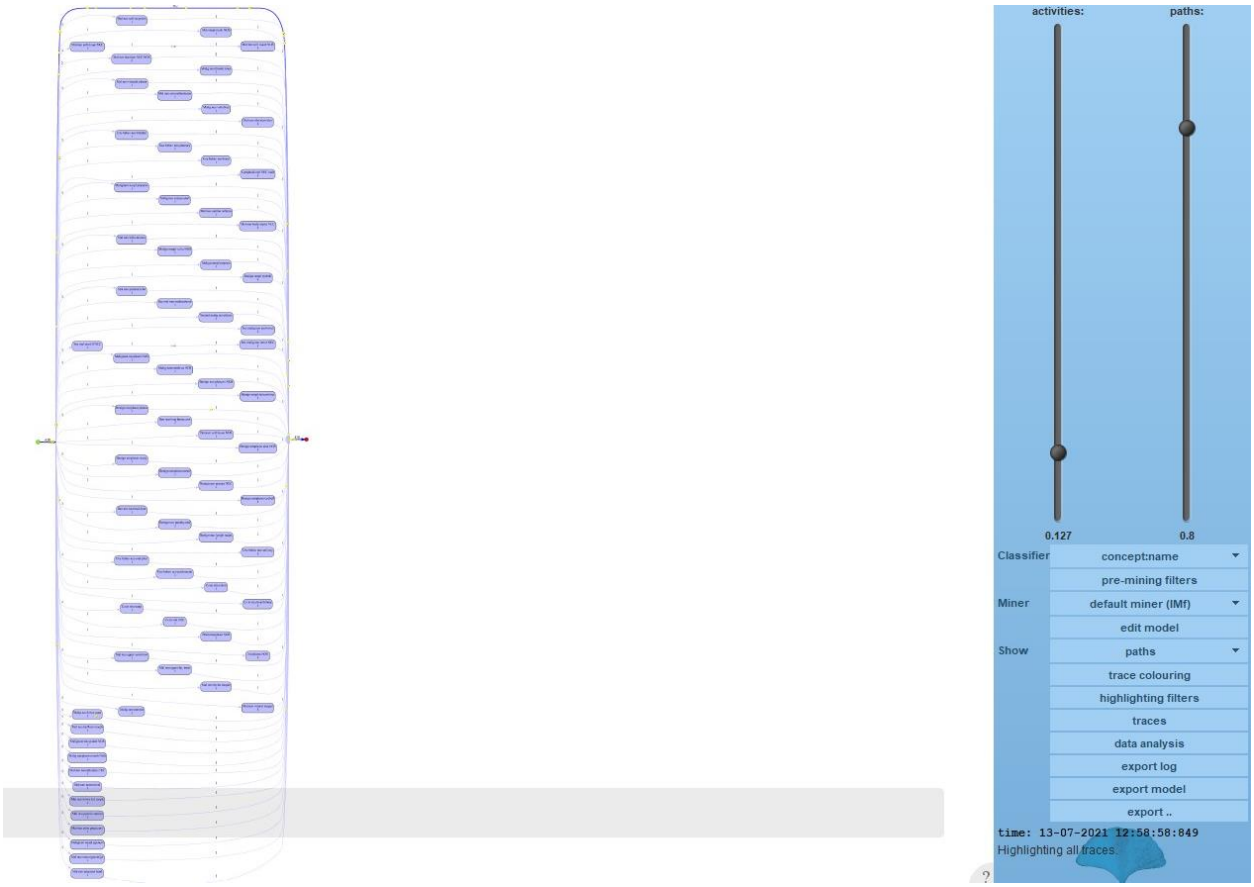


Figure 23. Inductive Visual Minor – Cancer dataset with cleaned data and activities slider set to 0.127.

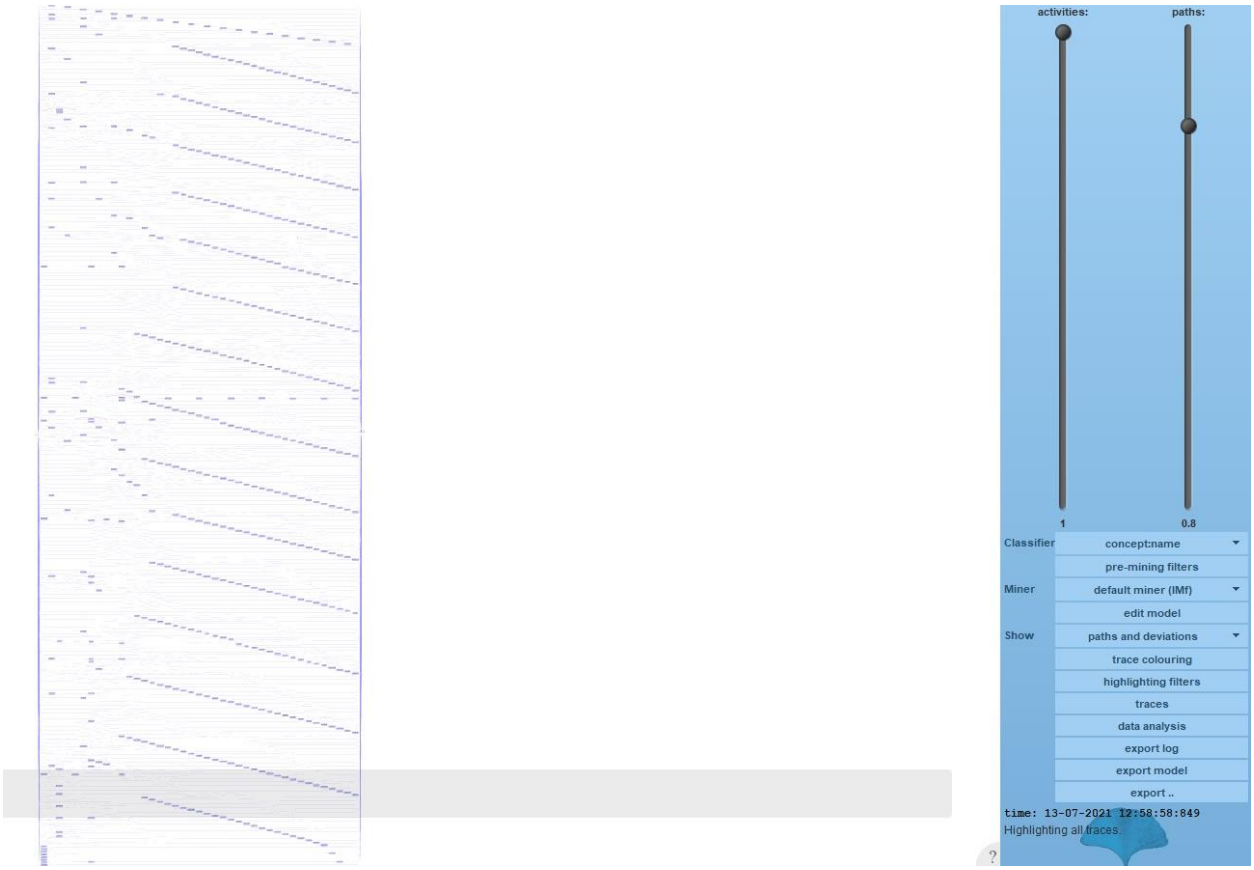


Figure 24. Inductive Visual Minor – Cancer dataset with cleaned data and activities slider set to 1.

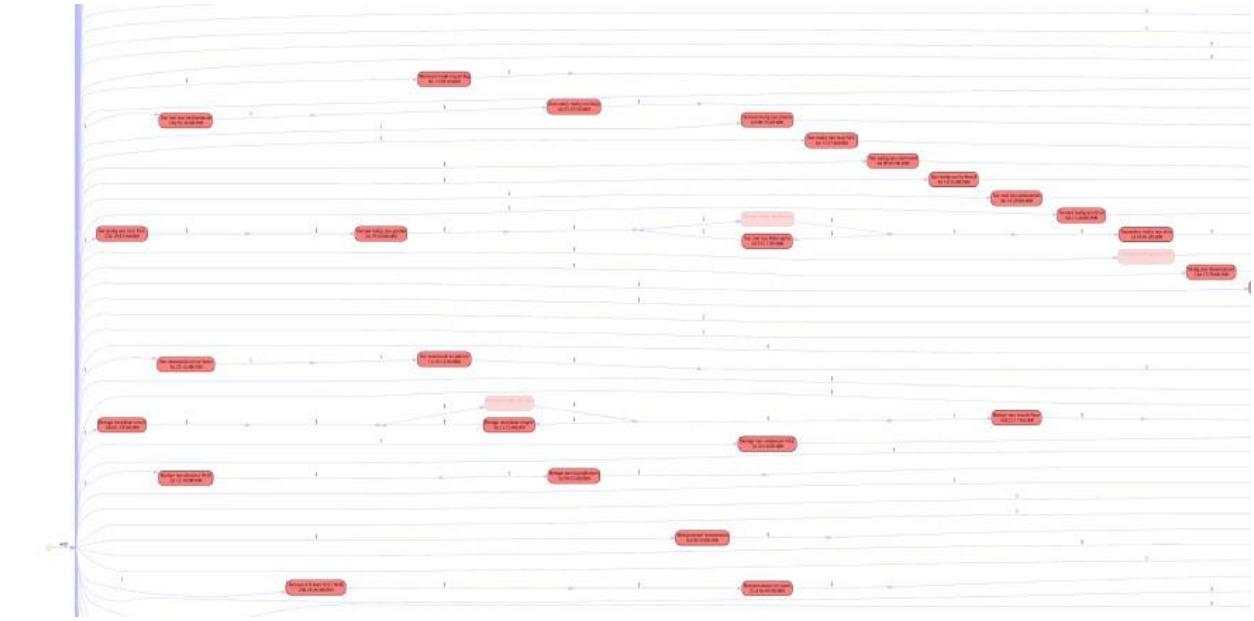


Figure 25. A closer view of Inductive Visual Minor – Cancer dataset with invalid data and activities slider set to 1, showing service times.

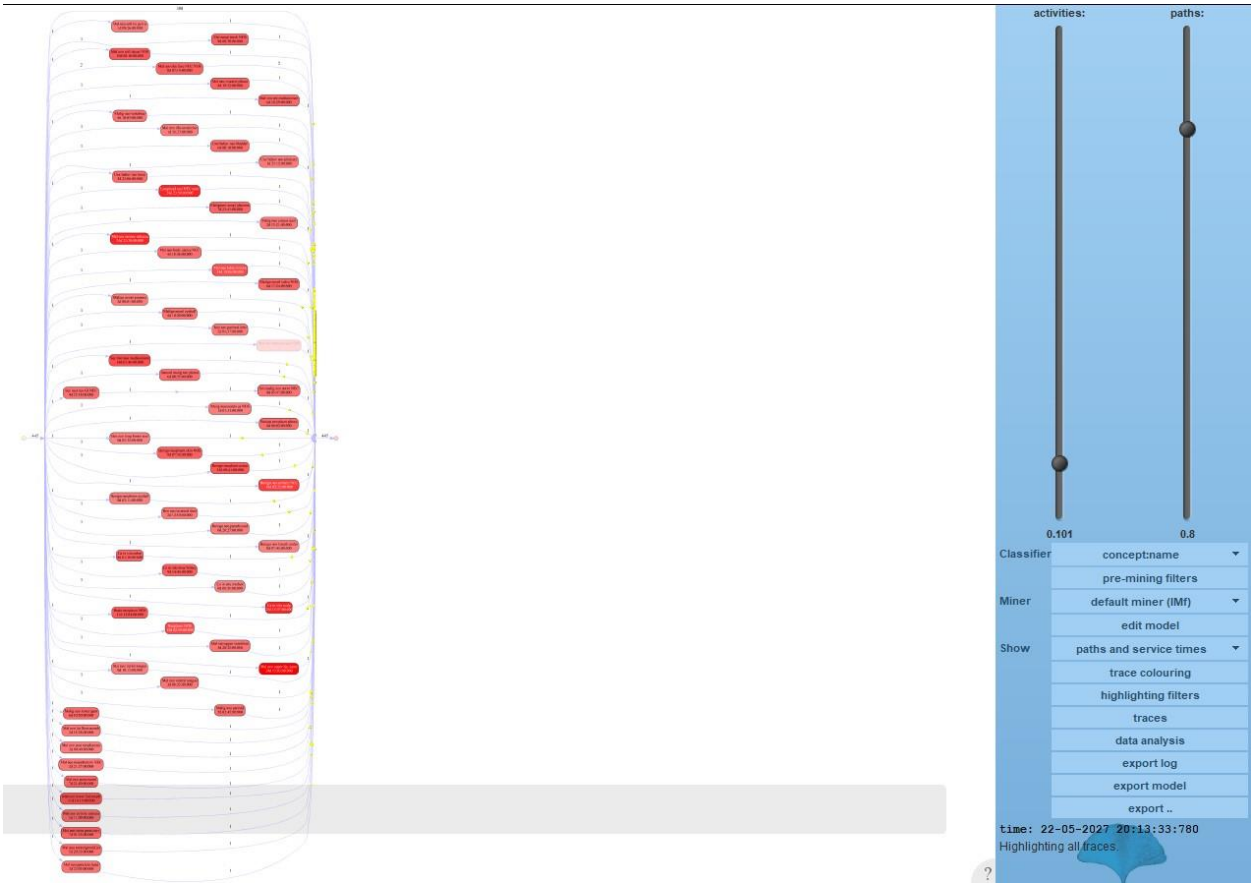


Figure 26. Inductive Visual Minor – Cancer dataset with invalid data and activities slider set to 0.101, showing service times.

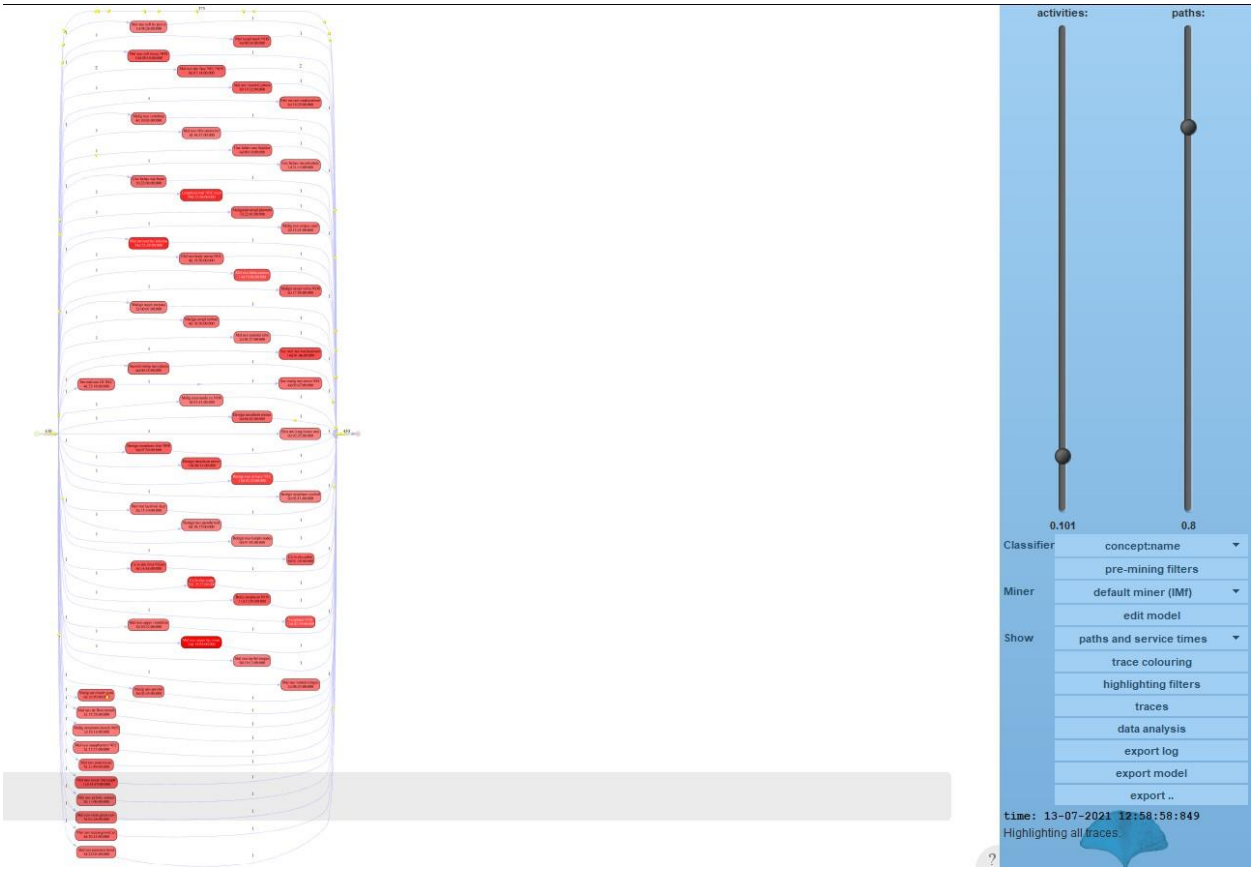
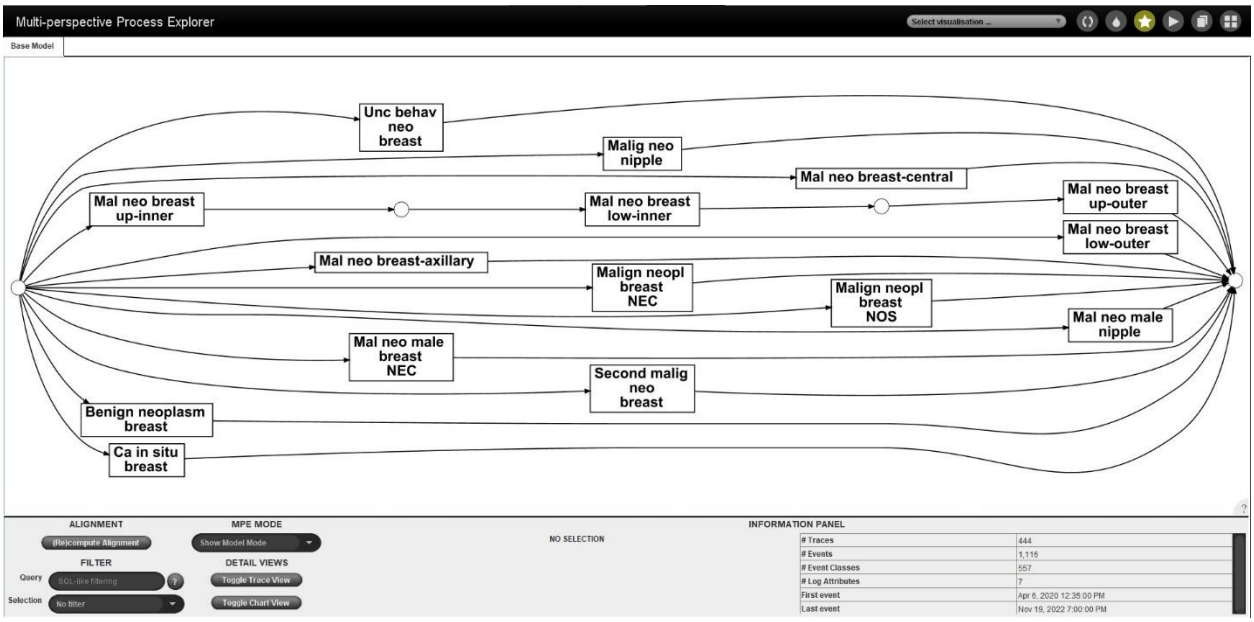
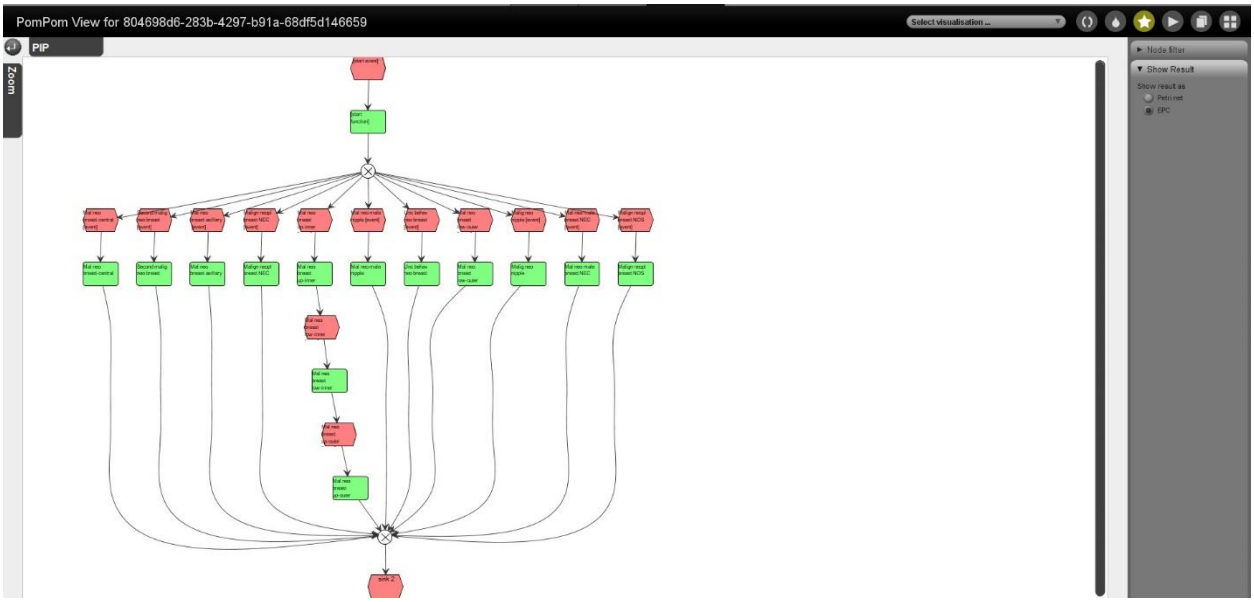


Figure 27. Inductive Visual Minor – Cancer dataset with cleaned data and activities slider set to 0.101, showing service times.



28. Multi-perspective Process Explorer.

D. More Process Mining Models



29. Pom-Pom View - uses petri net of breast cancer dataset and event log of cancer dataset.

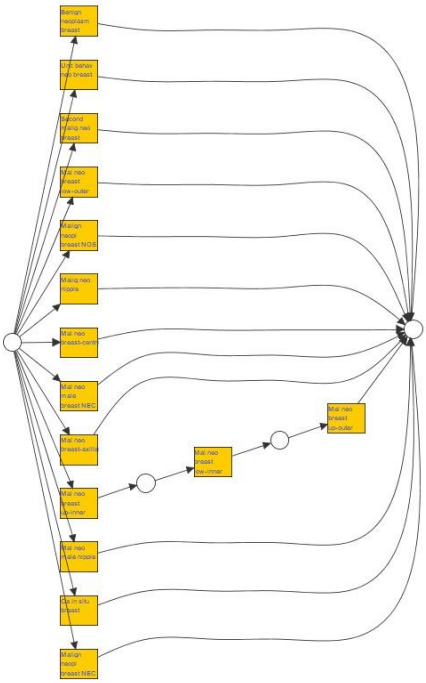


Figure 30. Conformance checking of DPN of a Petri net of Breast cancer data with Cancer data - used without using approximate matches of the data. The yellow color means that there are parts which do not cover each other fully, and that another data input is needed or fixing the current input data.