Computers in Biology and Medicine 146 (2022) 105621

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A semi-supervised decision support system to facilitate antibiotic stewardship for urinary tract infections

Sjoerd de Vries [a,b,*], Thijs ten Doesschate [c], Joan E.E. Totté [d], Judith W. Heutz [d,e], Yvette G.T. Loeffen [f], Jan Jelrik Oosterheert [c], Dirk Thierens [a], Edwin Boel [d]

[a] Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC, Utrecht, the Netherlands
[b] Department of Digital Health, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, the Netherlands
[c] Department of Internal Medicine, Infectious Diseases, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, the Netherlands
[d] Department of Medical Microbiology, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, the Netherlands
[e] Department of Rheumatology, Erasmus Medical Center, Dr. Molewaterplein 40, 3015 GD, Rotterdam, the Netherlands
[f] Division of Pediatric Immunology and Infectious Diseases, Wilhelmina Children's Hospital Utrecht, Lundlaan 6, 3584 EA, Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Urinary tract infection
Clinical decision support
Semi-supervised learning
Ensemble learning
RESSEL
Antibiotic stewardship

ABSTRACT

Urinary Tract Infections (UTIs) are among the most frequently occurring infections in the hospital. Urinalysis and urine culture are the main tools used for diagnosis. Whereas urinalysis is sufficiently sensitive for detecting UTI, it has a relatively low specificity, leading to unnecessary treatment with antibiotics and the risk of increasing antibiotic resistance. We performed an evaluation of the current diagnostic process with an expert-based label for UTI as outcome, retrospectively established using data from the Electronic Health Records. We found that the combination of urinalysis results with the Gram stain and other readily available parameters can be used effectively for predicting UTI. Based on the obtained information, we engineered a clinical decision support system (CDSS) using the reliable semi-supervised ensemble learning (RESSEL) method, and found it to be more accurate than urinalysis or the urine culture for prediction of UTI. The CDSS provides clinicians with this prediction within hours of ordering a culture and thereby enables them to hold off on prematurely prescribing antibiotics for UTI while awaiting the culture results.

1. Introduction

Urinary Tract Infections (UTIs) comprise a large part of all bacterial infections [1], causing serious health problems for patients as well as imposing a significant workload on diagnostic laboratories [2–4]. In women, cystitis (lower-UTI) is the most common cause for a visit to the general practice in the Netherlands [5] and UTIs are the fifth most common healthcare-associated infection in the United States [6].

Although definitions of UTI vary internationally [7,8], predominantly one or more associated signs or symptoms of UTI accompanied by the presence of bacteria in the urine (bacteriuria) are required. Antibiotics are effective and indicated in case of a diagnosis of UTI [2].

In the hospital, the diagnosis of UTI is often difficult. Clinicians frequently have to decide on day one whether to start antibiotics, at which time only the signs and symptoms and the urinalysis results may be available. Symptoms can be nonspecific for UTI [9,10] and although urinalysis is highly sensitive for UTI, it has low specificity [11]. The detection and identification of uropathogens by urine culture provide vital information for the diagnosis of UTI, but it can take up to three days before the results are available. This uncertain process predisposes to the unnecessary use of antibiotics.

Unnecessary treatment in the context of UTI diagnosis is an important part of inappropriate antibiotic administration and its prevention is a crucial element of modern antibiotic stewardship programs [12]. Inappropriate and unnecessary administration of antibiotics, estimated to occur in 20–50% of all hospital prescriptions [13], are the main drivers of increased antibiotic resistance. The increase in antibiotic resistance among pathogenic microorganisms in recent years is of growing concern [14,15]. The use of a predictive system which is able to accurately assess whether a patient is at high or low risk of having a UTI at an early stage in the diagnostic process could support the clinician in deciding not to start antibiotics. Thereby unnecessary antibiotic administration can be reduced.

In this paper, we report on the design and evaluation of a clinical

* Corresponding author. Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC, Utrecht, the Netherlands.
E-mail address: s.devries1@uu.nl (S. de Vries).

https://doi.org/10.1016/j.compbiomed.2022.105621
Received 31 December 2021; Received in revised form 18 February 2022; Accepted 19 March 2022
Available online 20 May 2022

decision support system (CDSS) to predict UTI before the urine culture results are available. An expert panel enriched our data set with a label for UTI, which we use as the target variable for a prediction model, rather than the urine culture result. This label is a more suitable endpoint for our goal of reducing unnecessary administration of antibiotics, as the presence of uropathogens in the urine in absence of symptoms, i.e. asymptomatic bacteriuria, generally does not require treatment with antibiotics. We first analyze the current diagnostic process using this expert-based label as reference, examining the predictive value of urinalysis screening and the Gram stain results. Based on the insights from this analysis the CDSS is constructed, using a two-step approach. In the first step the CDSS identifies cases with a negative urinalysis screening (based on leukocyte esterase and nitrite) as this reliably excludes UTI. In the second step the system estimates the risk of UTI for the remaining cases by using a predictive model based on urinalysis, Gram stain and other readily available parameters. This provides the clinician with an accurate prediction around the same time the urinalysis is available, while the culture results take one or several more days. In step two, an automated pipeline is used to first optimize feature subsets and hyperparameters and then to train and validate different models. The reliable semi-supervised ensemble learning (RESSEL) [16] method is then used to enrich these models using unlabeled data to further increase their predictive performance, since labeled data is scarce while a substantial amount of unlabeled data is available. Next, the performance of the CDSS as a whole is calculated for all included cases as well as different patient subgroups. Finally, we provide a potential impact analysis, estimating the reduction in the number of unnecessary antibiotic prescriptions that could be achieved by using the predictive framework.

2. Related work

Although UTIs have been studied extensively, the number of works describing the design or evaluation of predictive CDSSs is limited. Possibly the first work describing the use of a model for predicting urine culture results in patients with suspected UTI stems from 1985 by Wigton [17]. They use Discriminant Analysis to obtain the coefficients used in a linear equation to obtain a final score which translates to a range of possible decision rules, depending on which cutoff value is used.

Most retrospective cohort studies over large patient populations into UTI prediction use the urine culture as outcome instead of UTI, as the absence of structured data on symptoms or suspicion of UTI makes retrospective classification of UTI as outcome difficult on a large scale. Kim et al. [18] conducted such a study in a large hospital in Korea. Similar to the research by Wigton, they developed a prediction algorithm for culture outcome, the UTOPIA score. The coefficients included in the score were obtained by using multivariate binary logistic regression. Burton et al. [19] designed and successfully implemented a machine learning system to predict urine culture results in a routine clinical microbiology lab, significantly reducing the workload while retaining high sensitivity.

A number of studies have been performed in which patients were included based on being symptomatic and the culture result was used as outcome. These studies were generally performed on smaller populations. Heckerling et al. [20] employed genetic algorithms in order to search for valuable combinations of input features to a neural network. Another study [21] focused on feature selection, having available both clinical markers as well as immunological biomarkers. In contrast, Taylor et al. [22] were able to include a large number of patients presenting to the emergency departments at several hospital sites, by using regular expressions to find UTI related symptoms in clinical notes. Additionally, they compared the model outcomes to an alternative label based on antibiotic administration and documented diagnosis.

Additionally, many studies were conducted where the outcome was defined as UTI by the combination of inclusion via symptoms and a positive urine culture, but decision rules were constructed by specifying thresholds based on univariate relationships between predictors and the target variable. Little et al. [11] conducted such a study, reporting on all stages from development of the scores, to their validation, both retrospectively and prospectively, to the measurement of their impact on antibiotic administration along with the economic impact of the scores. In a meta-analysis [23] pooling the results of four previous studies, the authors compute the diagnostic accuracy of urinalysis results as well as the added value of history and physical examination for uncomplicated UTI.

Rather than using traditional predictors, from urinalysis or patient symptoms, some studies have focused on alternative approaches for detecting UTI. Li et al. [24] developed a model to predict positive urine culture in stroke patients who presented with symptoms, based on a stroke scale score and serum biomarkers. In another study [25], targeted at patients who suffer from dementia, the authors combine sensory device readings with bidaily physiological recordings to monitor changes in activity patterns from which they generate alarms signaling high probability of UTI. Other studies have researched alternative manners for analyzing urine. Turra et al. [26] employ hyperspectral image analysis using machine learning models in order to distinguish different potential uropathogens, reducing the time to outcome compared to the urine culture. Kodogiannis et al. [27] have researched the use of gas-sensing technology, or electronic nose, to extract sensor parameters from the urine samples on the basis of which a UTI prediction was made.

3. Methods

3.1. Study design and data collection

This retrospective single center study was conducted in cooperation with the antibiotic stewardship program of the University Medical Center Utrecht (UMCU) as part of the former Applied Data Science in Medicine (ADAM) project, currently the Digital Health Department. Routine care data from inpatients of the UMC Utrecht were included from January 2017 to December 2018. The use of these data for research purposes was approved by the Medical Research Ethics Commission (MREC) to be exempt from the Medical Research Involving Human Subjects Act (dutch: WMO), and the study adheres to the UMCU data management policy for Non-WMO Research. The data were obtained from the Research Data Platform that extracts data from the Electronic Health Records and the laboratory information management system GLIMS.

3.2. Data inclusion and patient subgroups

3.2.1. Data inclusion

A description of the data inclusion process is shown in Fig. 1. In total, data from 16,987 urine cultures from 7737 patients were collected. Cultures with non-standardized, missing or multiple conflicting results and those that could not be interpreted by the lab due to contamination (more than two micro-organisms in the urine other than *Enterobacteriaceae* en *Pseudomonas aeruginosa*) were discarded, as well as those which were not ordered for the purpose of finding pathogenic micro-organisms (PMO). Cultures for which either the admission data or basic patient characteristics (sex, age) were not available, were excluded. Another filtering was applied to respect the UMCU opt-out policy, which enables patients to prevent their data from being used for research purposes. This resulted in a "Useable Data Set" containing 13,286 cultures of 7295 patients.

After these filter steps a further exclusion occurred based on the treating specialism and admission ward, followed by the exclusion of neutropenic patients. Cultures from patients that were under care of the hematologist (664) or patients that resided on the adult (486) or pediatric (494) intensive care units were excluded because of the complex population, the high level of antibiotic prophylaxis use and the low-

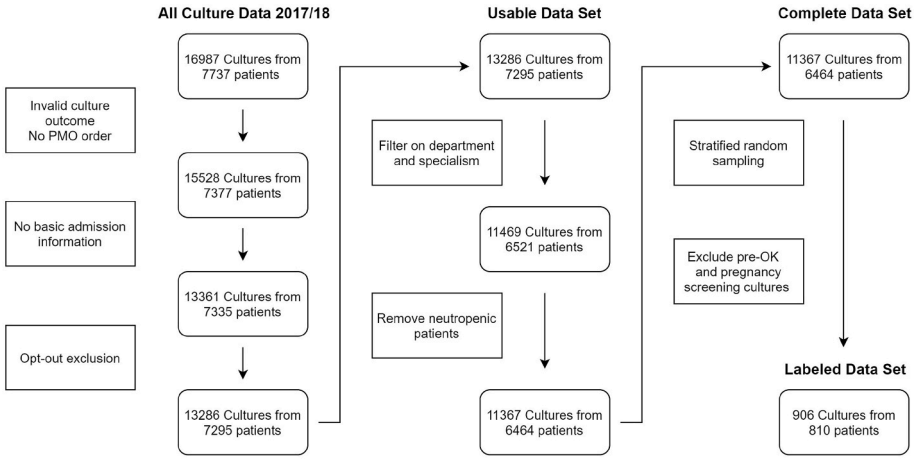


Fig. 1. Patient inclusion flowchart. PMO stands for pathogenic micro-organisms.

threshold antibiotic prescribing. Cultures from neutropenic patients (296) were left out since these patients have a shortage of leukocytes by definition, while the presence of leukocytes in urine plays an important role in UTI detection [28,29]. Patients were defined as neutropenic if their neutrophil count was $< 0.5{\cdot}10^9/L$ on the day of the culture, or both the day before and after [30]. Furthermore, cultures from a small number of other departments where cultures would normally be ordered for a different purpose than UTI detection were excluded (173).

The "Complete Data Set" remaining after these steps contained 11,367 cultures from 6464 patients.

Finally, during the labeling procedure of the urine cultures, described in detail in Section 3.3, cultures that were used for screening purposes before an urological procedure were removed, as these serve a different purpose from regular UTI diagnosis and no reduction in administration of antibiotics is to be expected. Similarly, cultures obtained from pregnant women were excluded, as the guidelines prescribe treating asymptomatic bacteriuria in this group, making a prediction model for UTI obsolete in the presence of a culture result.

3.2.2. Patient subgroups

We identified four patient groups that we suspected could differ from the overall population to the extent that a model might underperform for them: immunosuppressed patients (as defined in Section 3.4), elderly ($\geq$75), children (< 18) and urological patients. They were not excluded up-front, but instead model performance was monitored separately for these groups. The number of patients corresponding to these groups in the data set is shown in Table 1.

3.3. Expert panel labeling

Even though the urine culture is the main diagnostic tool in detecting UTI, the culture result alone is insufficient for the diagnosis of UTI, mostly due to the existence of asymptomatic bacteriuria that does not require antibiotic treatment, with some exceptions e.g. pregnant women and before certain urological procedures [31]. Therefore, an expert panel, consisting of an infectious disease specialist, a fellow medical microbiologist, a pediatrician-infectious disease specialist and a fellow infectious disease specialist, provided 906 cultures from 810 patients with a UTI label. The process by which the experts decided upon this label is described in the following.

A detailed sheet of information was provided to the experts, containing for each patient all information about the variables listed in Tables B.14, B.15, B.16 and B.17 in Appendix B. These data were collected from 7 days prior to the date of the culture order to 7 days after and included results of the corresponding urinalysis, urine culture, Gram stain, laboratory test and any previous cultures as well as registered antibiotics started, stopped or active during this period. Importantly, the clinical notes were available, to provide the necessary context such as signs, symptoms and alternative diagnoses. Since these data were only available in unstructured format, they were not suitable for use in the prediction models without the application of sufficiently reliable text mining methods.

Along with the UTI label, each expert provided a confidence score ranging from 1 to 10. If the experts were not sufficiently confident, i.e. a score of 6 or below, the case was discussed further with another expert until consensus was reached. During this process, any pre-operative screening cultures for urological procedures and cultures of pregnant patients were removed from the set as described in Section 3.2.1.

Cultures were selected for labeling using stratified random sampling to preserve the distribution of the corresponding patients over the different departments of the Complete Data Set. The distributions of the data before and after labeling are shown in Table 2.

3.4. Predictor variables

A selection of variables to include for predictive modeling was made based on prior literature and clinical experience. An overview of all of the variables that we considered can be found in Tables B14, B15, B16 and B.17 in Appendix B. Only data that were available on the day of the culture order were included as the model will be used to make a

Table 1
Number of included patients per (sub)group.

Patient group	Cultures	%	Labeled	%
Immunosuppressed	4452	39.2	345	38.1
Elderly	2285	20.1	181	20.0
Children	1440	12.7	168	18.6
Urology	975	8.6	93	10.3
Other	3599	31.7	256	28.3
Total	11367	100.0	906	100.0

Table 2
Number of included cultures per department. Other medical wards include e.g. cardiology, respiratory and neurology wards.

Department	Labeled	%	Total	%
Surgical	198	21.9	2730	24.0
Internal medicine	177	19.5	2231	19.6
Emergency	172	19.0	1888	16.6
Pediatric	121	13.4	1240	10.9
Other medical wards	111	12.3	1707	15.0
Urological	93	10.3	975	8.6
Gynecological	34	3.8	596	5.2
Total	906	100.0	11367	100.0

prediction at that point in time.

The data included the patient characteristics sex and age, whether the urine was extracted midstream or via catheter, urinalysis (nitrite, leukocytes, erythrocytes, protein, glucose, pH), urine sediment microscopy (hyaline/granular casts), clinical chemistry urine measurements (sodium, osmolality, creatinine, urea, potassium), Gram stain measurements (Gram-negative/-positive rods/cocci, leukocytes, epithelial cells), inflammation parameters in the blood (CRP, leukocytes, neutrophils), temperature measurements and whether a chest X-ray was performed as an indication of a broader search for infection.

Patient characteristics, the method by which the urine was collected and the Gram stain results were available for every culture included in the Complete Data Set. The availability of the other variables varied and has been reported in Appendix B. The data from the laboratory as well as the temperature were considered up to 4 days prior to the culture order. Previous urine cultures were included up to 7 days prior to culture ordering. Leukocytes and neutrophils from blood were included up to 14 days before the culture order. For all of these the last measured value was used if multiple measurements were present. The occurrence of a chest X-ray was considered up to 14 days prior to the culture order.

Furthermore, we include diabetes and immunosuppression as predictors by processing prescribed medication data: A patient was classified as diabetic if any medication with an Anatomical Therapeutic Chemical (ATC) code [32] starting with A10 was administered in the year previous to the culture order. A patient was classified as immunosuppressed if any medication with an ATC code starting with A07E, A14A, H02, L01 or L04 was administered in the previous 90 days.

Antibiotic prescriptions were taken into account as well. Antibiotics were categorized into three classes: UTI specific (U) [5], Broad-spectrum (B) and Not for UTI (N). The exact classifications can be found in Table C.18 in C. Whether a patient was on any of the specific types of antibiotics on the day of the culture order was available as a feature in the model, in addition to whether the patient was on any antibiotics regardless of type.

The culture result itself was not included as a predictor, as it becomes available at a much later stage in the diagnostic process than the other variables.

3.5. Model development

3.5.1. Preprocessing

In order to transform the collected data into a suitable format for machine learning models to handle, the data had to be preprocessed. As we are interested in building a model which is applicable in clinical

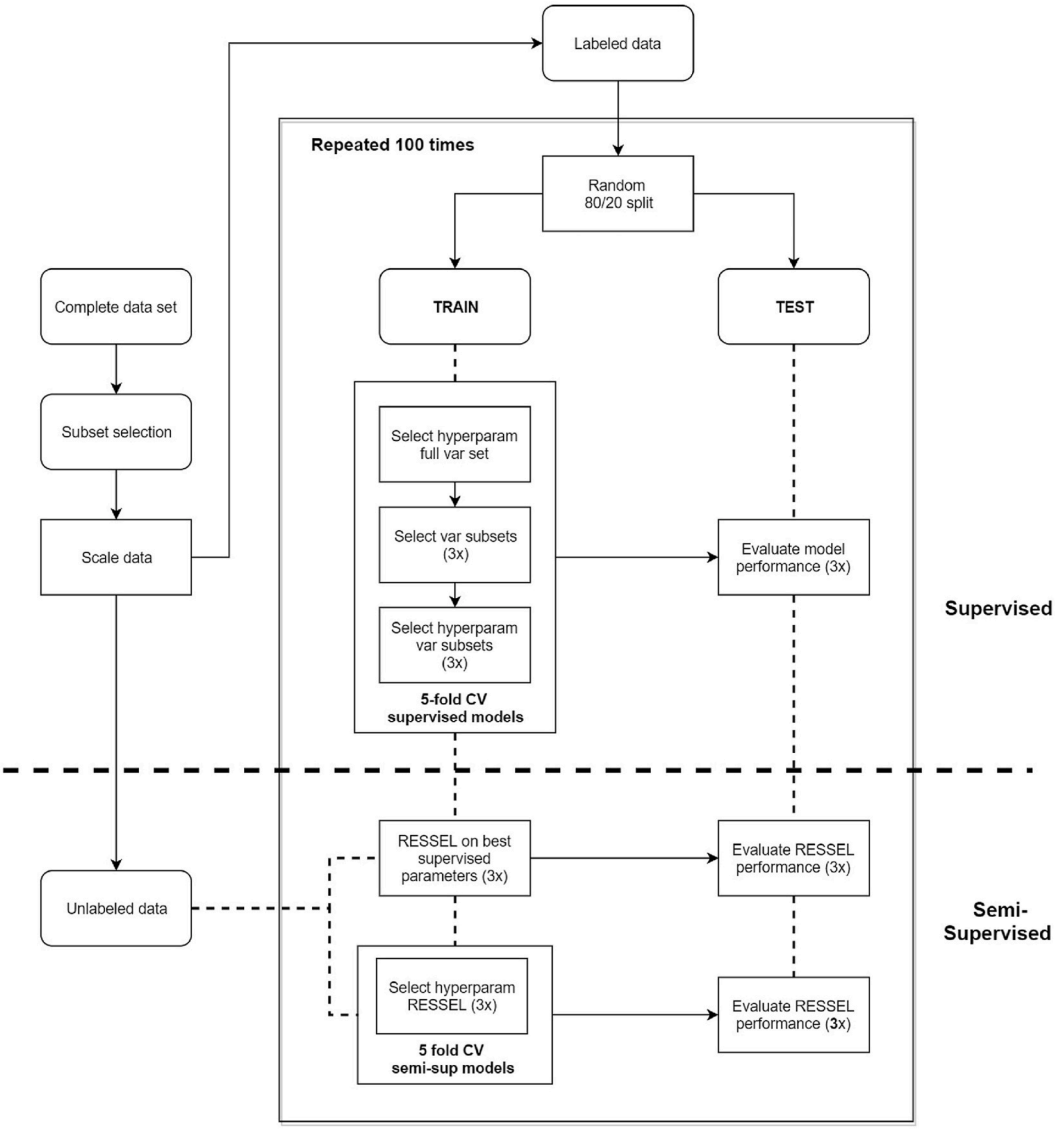


Fig. 2. Experimental setup.

practice every time a UTI is suspected, given the patient meets the inclusion criteria, a record in our data set represents an episode corresponding to a single urine culture. This episode contained all information that was available within a couple of hours after the culture order, thus including the urinalysis and Gram stain results, but not the culture result itself.

Four of the features in the data set were filled using imputation when they were missing: temperature, CRP and blood leukocytes and neutrophils. These were imputed with their respective median values over the population as a whole. Other missing values were set to 0 and an additional dummy variable was introduced for the model to distinguish their missingness. This distinction in method of imputation was made based on the difference in the availability of the features, as seen in Table B.14 in Appendix B.

3.5.2. Experimental setup

We constructed a prediction pipeline to optimize, train and validate prediction models, depicted in Fig. 2. The steps in the pipeline are explained in the following.

The first step was to take a subset of the complete data if required for the current experiment, e.g. cultures associated with positive urinalysis.

Next, the predictors were standardized (Z-score normalization) to make them suitable for each of the different classifiers tested.

Then, a distinction is made depending on the type of machine learning used to train the models: supervised learning (i.e. training using only the labeled data) was applied in every experiment, while semi-supervised learning (i.e. training using the unlabeled data as well) was only applied to the subpopulation that was ultimately selected as the target population for which a model would be most effective if implemented in clinical practice.

We opted to employ a repeated re-sampling strategy as part of our experimental setup: the Labeled Data Set was repeatedly split into a separate training set, consisting of 80% of the data, and a test set, consisting of 20% of the data. This split was constructed by random sampling without replacement and the process was repeated 100 times to obtain 100 different train/test splits. The 20% data in the test set was set apart and only used to evaluate model performance during the final step of the experimental pipeline. All tuning of the hyperparameters and variable selection was limited to the training set. The experiments that included both supervised and semi-supervised models employed the same random split, to allow for a fair comparison of the calculated metrics on an identical test set.

The supervised part of the pipeline consists of the following steps:

1. Preliminary hyperparameter selection using 5-fold cross validation within the train set on the full feature set. The Area Under the Receiver Operating Characteristic Curve (AUC) was used as the metric to optimize at this stage, as we believe it would be most beneficial to first optimize the predictive power of the model over the entire range of cutoff points in the early steps of the pipeline. The accuracy is maximized for the threshold which will be used by the model in practice in the final optimization step of the pipeline.
2. From the candidate features described in Section 3.4, the most informative were selected using sequential backward feature selection on the train set, as implemented by Mlxtend [33], with floating enabled. The hyperparameters selected in the previous step were used. This entailed using 5-fold cross validation while removing features in turn, training the model on the reduced feature sets and measuring the performance. Three feature subsets resulted from this selection procedure: a Fixed feature set was predefined by the expert panel to include the features that were deemed very likely to be of predictive value and whose exclusion might lead to decreased trust in the system: any active antibiotics at time of culture, urine nitrite, urine leukocytes, Gram positive cocci, Gram negative rods, urine collection method, age and sex. Furthermore, the feature set which was found to have optimal cross validated AUC, was selected. We observed this often resulted in features being included that resulted in minimal increases in performance, which we suspected could be attributed to the randomness in the train-test split. Therefore a third, Sparse set was retained as well, which was restricted to be the smallest number of features for which the cross validated performance was within 0.25 times the standard error of the optimal result. This threshold was established by visually inspecting the feature subset performance figures (see Fig. 3 for an example) and found to be a good trade-off between performance and sparsity.
3. For each of the three reduced feature sets another round of hyperparameter optimization followed, as the optimal setting may vary between the different feature sets. The target metric was set to accuracy at this stage, as we will be using the model with a single cutoff point and want its accuracy to be maximized.
4. The predictive performance of these supervised classifiers, with optimized feature sets and hyperparameters, was measured on the data in the test set of the corresponding train/test split. The final supervised models were trained on the entire training set consisting of 80% of the data using the optimized reduced feature sets and corresponding hyperparameter values.

If the experiment included semi-supervised learning to further enrich the models by including the unlabeled data, the following additional steps were appended to the pipeline:

1. The models were refined using the unlabeled records through semi-supervised learning, using the RESSEL method. The feature subsets and hyperparameter values found in the supervised steps were maintained. These models were then evaluated on the test set.
2. The hyperparameters were optimized in a separate 5-fold cross validation for each of the feature subsets on the training data. Again, as this is the last step in the optimization, accuracy was chosen as the metric to optimize.
3. The models were retrained on the train set using the optimized hyperparameter values for each feature subset. The performance of the semi-supervised models was measured on the test set of the corresponding train/test split.

The results over the 100 train/test splits were averaged to obtain robust estimates of the predictive performance of the different models.

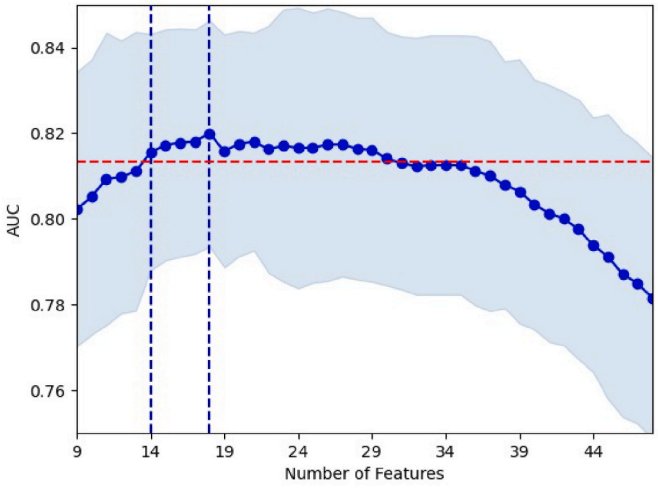


Fig. 3. Sequential Feature Selection. The red line indicates the cutoff for the Sparse feature set. The leftmost blue lines indicates the number of features that was included in the Sparse feature set. The rightmost blue line indicates the number of features for which the best cross-validated accuracy was found, i.e. the size of the Selected feature set.

3.5.3. Supervised models

We compared a total of five different supervised classifiers implemented in the Scikit-learn package [34]: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), eXtreme Gradient Boosting (XGB) and k-Nearest Neighbors (NN).

The hyperparameter settings for these models, as well as the ranges of the pre-specified grids used in hyperparameter selection by means of a cross-validated grid search, are shown in Table A.12 in Appendix A.

3.5.4. RESSEL

In addition to the ($n = 906$) Labeled Data Set, a large number of unlabeled cultures ($n = 10,461$) was available. In order to achieve the best predictive accuracy, we applied the reliable semi-supervised ensemble learning (RESSEL) [16] method to improve upon the supervised models.

RESSEL is an ensemble based wrapper method and as such it takes a collection of trained classifiers and uses the available unlabeled data to attempt to improve their predictive performance. These classifiers are initially trained on bootstrap samples of the labeled training data, setting apart the out-of-bag samples. Each classifier is then individually enhanced using a portion of the unlabeled data by applying the semi-supervised self-training technique. During this process the out-of-bag error is monitored, thus assuring the accuracy can at most deteriorate to the extent that the out-of-bag samples are not representative of the overall data. The resulting classifiers are used to make an ensemble prediction.

The ensemble benefits from the diversity introduced through the bootstrapping as well as self-training procedures, while the early-stopping mechanism based on the error measurements as well as the merging of the individual classifiers into an ensemble provide the necessary robustness to the collective.

In previous work we found RESSEL to improve predictive performance for a wide range of data sets by using unlabeled data [16], while no explicit checking of difficult assumptions on underlying problem structure is required. The method was therefore deemed a good fit for the current problem of predicting UTI from limited labeled and a large amount of unlabeled examples. For an in-depth description of the method as well as the algorithmic details, we refer the reader to the original paper [16].

The RESSEL method required a number of hyperparameter settings to be specified. The settings used in our experiments are shown in Table A.13 in Appendix A.

3.5.5. Modeling process

The following sections are structured such as to reflect the process we followed in developing the CDSS, as depicted in Fig. 4: First, we performed a traditional analysis of the study population, compared the expert panel label to the culture result and determined the effectiveness of different thresholds for a urinalysis screening rule in Section 4. Based on the insights generated by these analyses, we determined the population for which a model would be the most effective in reaching our goal of reducing the administration of unnecessary antibiotics, as described in Section 5. The decision for the use of a two-step approach in the CDSS was made, the first part of which is a urinalysis screening rule and the second part consisting of a predictive model for patients who had a positive urinalysis result. The predictive values of both the semi-supervised RESSEL model and the CDSS as a whole are then reported in Section 6.

4. Clinical findings

4.1. Description of the study population

A total of 906 cultures from 810 patients were included in the Labeled Data Set, which forms the base set for the supervised training of the prediction models, as shown in Fig. 2. The Complete Data Set consisting of 11,367 cultures from 6464 patients is used for semi-supervised learning to further enhance the supervised models. Descriptive statistics for the predictor variables that were included into the models by default, i.e. the Fixed set, are shown in Table 3 for the Labeled Data Set.

The median age at the time of a urine culture was slightly lower (57.0 years) in the non UTI group, than it was for the UTI group (63.0 years). The percentage of cultures that corresponds to a UTI is higher for women than for men (33.4%–25.3%). Most urine was collected as midstream clean-catch urine (58.3%), with the remainder collected via a catheter (41.7%). Those patients from whom the urine was collected via a catheter had a higher chance of having a UTI compared to the midstream group (32.3%–27.3%). The percentage of cultures with the UTI label was lower (24.6%) for patients who were already on antibiotics of any kind than for those who were not (32.3%). We observe 191 out of the 906 patients (21.1%) did not have a urinalysis performed in the days up to the culture order. Finally, we observe that the percentage of cultures with a UTI label increases as more Gram-negative rods/-positive cocci are detected in the corresponding urine.

A more detailed report on the variables not included in the Fixed set can be found in Appendix B, with a description of the binary variables, numerical variables, results from the urinalysis and the variables pertaining to the culture shown in Tables B.15, B.14, B.16 and B.17 respectively. The results of the urine culture of interest itself are further discussed in the following section.

4.2. Accuracy of the urine culture for detecting UTI

We evaluated the effectiveness of the culture result for detecting UTI as labeled by the expert panel. A positive culture was defined as the growth of $\geq 10^4$ colony forming units (cfu)/mL of a common

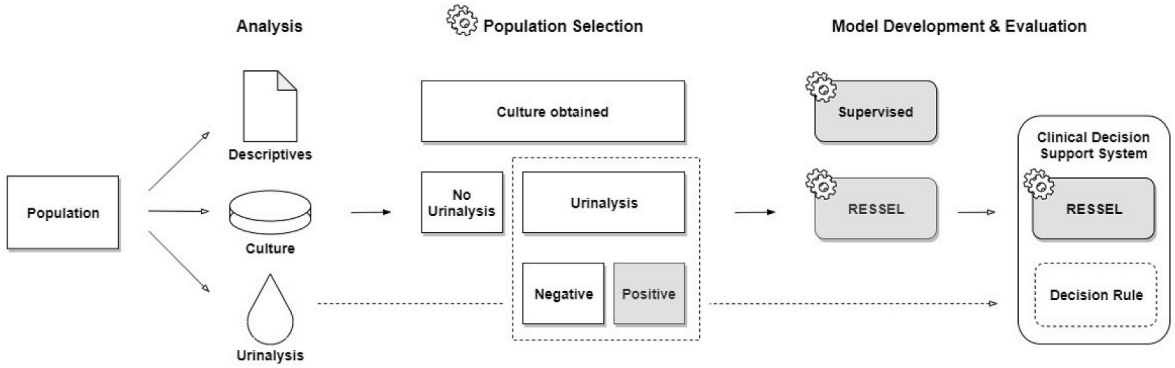


Fig. 4. Modeling Process. The order of the different actions depicted from left to right corresponds to the order in which development took place. A cogwheel indicates the training and evaluation of a model.

Table 3
Description of the predictor variables included in the Fixed set.

Variable	Value	Number (%)	No UTI(% or IQR)	UTI(% or IQR)
Age	Years	906 (100.0)	57.0 (27.8–71.0)	63.0 (38.0–74.0)
Sex	Female	452 (49.9)	301 (66.59)	151 (33.4)
	Male	454 (50.1)	339 (74.67)	115 (25.3)
Collection method	Catheter	378 (41.7)	256 (67.72)	122 (32.3)
	Midstream	528 (58.3)	384 (72.73)	144 (27.3)
Any antibiotics	No	564 (62.3)	382 (67.73)	182 (32.3)
	Yes	342 (37.7)	258 (75.44)	84 (24.6)
Urine leukocytes ($/\mu$L)	Not Measured	191 (21.1)	156 (81.68)	35 (18.32)
	Negative	304 (33.6)	287 (94.41)	17 (5.6)
	Ca. 25	70 (7.7)	57 (81.43)	13 (18.6)
	Ca. 75	71 (7.8)	49 (69.01)	22 (31.0)
	Ca. 250	36 (4.0)	20 (55.56)	16 (44.4)
	Ca. 500	234 (25.8)	71 (30.34)	163 (69.7)
Urine nitrite	Not Measured	191 (21.1)	156 (81.68)	35 (18.32)
	Negative	628 (69.3)	466 (74.2)	162 (25.8)
	Positive	87 (9.6)	18 (20.69)	69 (79.3)
Gram negative	Negative	528 (58.3)	461 (87.31)	67 (12.7)
rods (/1000x)	<1	69 (7.6)	59 (85.51)	10 (14.5)
	1–5	91 (10.0)	63 (69.23)	28 (30.8)
	6–30	87 (9.6)	28 (32.18)	59 (67.8)
	>30	131 (14.5)	29 (22.14)	102 (77.9)
Gram positive	Negative	648 (71.5)	480 (74.07)	168 (25.9)
cocci (/1000x)	<1	74 (8.2)	57 (77.03)	17 (23.0)
	1–5	90 (9.9)	67 (74.44)	23 (25.6)
	6–30	55 (6.1)	21 (38.18)	34 (61.8)
	>30	39 (4.3)	15 (38.46)	24 (61.5)

uropathogen in case of a midstream urine sample, or the growth of $\geq 10^3$ cfu/mL in catheter samples or midstream samples specifically if the patient was female and the uropathogen found was *Escherichia coli*. The isolation of more than two organisms may suggest contamination and such cultures were classified as negative. The comparison is shown in Table 4.

The culture was found to be 77.3% accurate in detecting UTI using the expert panel labeling as gold standard. Cohen's kappa was found to be 0.51, which can be seen as moderate agreement. The largest discrepancy was found for the positive cultures, which if the culture was used directly for UTI detection would lead to false positives 157 out of 374 times (42.0%) corresponding to a PPV of 58.0%. These false positives are cases of asymptomatic bacteriuria, i.e. the presence of uropathogens in the urine in absence of symptoms, and are therefore not classified as UTI. The number of false negatives found was 49 out of 532 (9.2%) corresponding to a NPV of 90.8%. The culture was found to have a sensitivity of 81.6% and a specificity of 75.5%. These false negatives could for example be due to a uropathogen being detected for which the number of cfu/mL was below the threshold or the culture being deemed negative as per the contamination criterion, while the patient was determined to have an UTI.

Table 4
The culture result compared to the expert panel labels.

		UTI		Metrics	
		No	Yes		
Culture	Negative	483	49	NPV: 90.8%	Specificity: 75.5%
	Positive	157	217	PPV: 58.0%	Sensitivity: 81.6%

4.3. Predictive value of the urinalysis

An important tool in the process of UTI diagnosis is the urinalysis. The measurements include: the number of leukocytes and erythrocytes, the presence of nitrite and glucose, the amount of protein and the pH of the urine. Especially the presence of nitrite and/or leukocytes are strong predictors of UTI and their combination is regularly used as a screening test in the current diagnostic process in the hospital.

The predictive values associated with the decision rules based on the different possible cut-off value combinations of the nitrite and leukocyte values are shown in Table 5.

In total, data from 715 cultures, belonging to 659 patients are shown in this table. Combining the different thresholds to calculate the AUC of a rule based classifier based on the aforementioned definitions results in an AUC of 85.84.

We observe that a high sensitivity is achieved by some of the threshold combinations, with the most conservative decision rule achieving a sensitivity of 93.94%. Since the urinalysis results are used as a screening test, a high sensitivity is necessary. The increased sensitivity logically comes at the cost of a reduced specificity. For the rules to be effective in ruling out UTI, it is important to have reasonable specificity as well. We observe a shift to the second tier of leukocytes (from ca. 25 to ca. 75 leukocytes) results in a loss in sensitivity of 4.33% (10 additional false negatives), but in a gain of the specificity of 11.37% (55 fewer false positives), and the rule overall is found to be 6.29% more accurate. As sensitivity remains high, this is the level we believe to be most effective in achieving our goal of reducing unnecessary antibiotic prescriptions while preserving patient safety. In the remainder of this work, we use the latter threshold to define positivity, i.e. the screening is positive if either nitrite is positive or ca. 75 leukocytes or more were found, and refer to the resulting decision rule as the urinalysis screening rule, which forms the first step of the CDSS.

5. Target population specification

5.1. Target population selection

A number of different candidate populations for which a prediction model could be deployed can be distinguished, as seen in Fig. 5.

Our aim was to decide which model would be most likely to have an impact on clinical practice. In order to make this decision, we weighted both model performance and the ease of integration into the current work process.

In the following we describe the considerations for each of the candidate groups for which a model could be developed and ultimately for the selection of the positive urinalysis group. The accuracies referred to are shown in Table 6.

1. Culture Obtained. This group consists of all cultures included in the Labeled Data Set. The increase in accuracy found in this group compared to the distribution of UTI (29.4%–70.6%) can be completely explained by the weighted combination of the accuracies of the models of its subgroups, the No Urinalysis and Urinalysis groups, i.e. $(715 \cdot 84.38 + 191 \cdot 82.47)/906 = 83.98\%$, compared to 84.01% found for the Culture Obtained group.
2. No Urinalysis. This group of cultures was not accompanied by urinalysis results. The predictive models achieved an accuracy not much better than the percentage of negative cultures in this group. A model predicting no UTI for every culture would thus have nearly equal performance and the added value of a model on this group is negligible. This confirms the urinalysis is an essential part of the

Table 5
Predictive values for different cut-off value combinations for leukocytes and nitrite measured during urinalysis. The decision rule corresponding to a combination of a selected nitrite and leukocyte value is to predict negative if the nitrite value is less than or equal to its selected value (with negative defined as less than positive) AND the number of leukocytes is less than or equal to its selected value as well.

Urinalysis	Leukocytes (/μL)	UTI	Yes (%)	Total	Metrics		Sensitivity	Specificity	Accuracy
Nitrite		No (%)			PPV	NPV			
Negative	0	285 (95)	14 (5)	299	52.16	95.32	93.94	58.88	70.21
	ca. 25	55 (85)	10 (15)	65	58.97	93.41	89.61	70.25	76.50
	ca. 75	48 (76)	15 (24)	63	66.67	90.87	83.12	80.17	81.12
	ca. 250	19 (56)	15 (44)	34	69.69	88.29	76.62	84.09	81.68
	ca. 500	59 (35)	108 (65)	167	79.31	74.20	29.87	96.28	74.83
Positive	0	2 (40)	3 (60)	5	52.07	94.41	92.64	59.30	70.07
	ca. 25	2 (40)	3 (60)	5	58.94	91.98	87.01	71.07	76.22
	ca. 75	1 (12)	7 (88)	8	66.30	88.31	77.49	81.20	80.00
	ca. 250	1 (50)	1 (50)	2	69.66	85.86	70.56	85.33	80.56
	ca. 500	12 (18)	55 (82)	67	-	-	-	-	-
Total		484 (68)	231 (32)	715	-	-	-	-	-

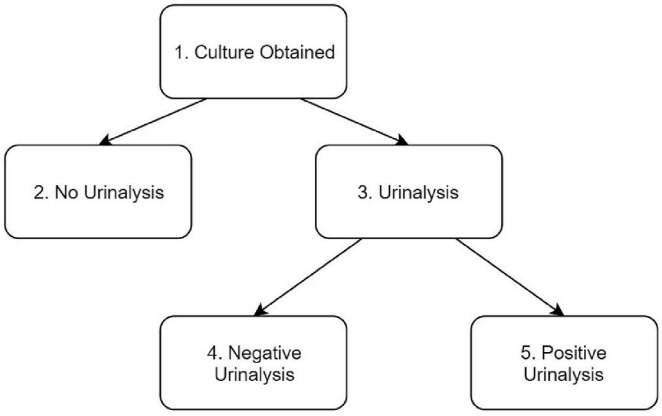


Fig. 5. Candidate populations for model development.

diagnostic process surrounding UTIs and is required for the model to have sufficient predictive performance.

3. Urinalysis. For this group, a lot of additional diagnostic information is available from the urinalysis results. In clinical practice, this information is already used effectively to rule out UTI based on the presence of nitrite and leukocyte counts. The use of such a cutoff rule, as seen in Table 5 achieves high predictive accuracy by itself (77% for our previously selected threshold, at most 82%). While the use of a prediction model for the entire group can further improve this accuracy to 84.4%, a divergence from such an established rule might lead to resistance. Furthermore, in the negative urinalysis group, very few positives remain (6.6%), indicating the good predictive performance is due to distinctions being made in the positive group.
4. Urinalysis Negative. The patients corresponding to the cultures in this group were found overwhelmingly not to have a UTI (93.4%). Model accuracy was found to be similar, effectively on par with a model classifying every culture in this group as negative. This could in part be because the data set is very imbalanced, a known difficulty when training classifiers, and the use of over- or undersampling techniques might further improve accuracy. On the other hand, the low number of positive samples in this group (24) limits the potential impact of a model, even if it was further improved. Simply using the urinalysis rule as a screening rule, as is current practice, is already very effective.
5. Positive Urinalysis. This group consists of patients for whom the urinalysis decision rule was positive. When the urinalysis results are in the clinician is often faced with the decision whether to prescribe antibiotics or wait for the culture result, which will take at least another day even if it was ordered at the same time as the urinalysis. From our analysis, we observe that in 41% of cases patients in this group have no UTI, however. A model which uses the Gram stain results for this group increases the accuracy from 67.56% to 75.63% compared to a model without the Gram stain results. With the incidence of UTI at 59% for this group, we conclude that the model that includes the Gram stain is able to provide important additional information to the clinician at this stage in the diagnostic process which can be used to indicate the absence of UTI, allowing for postponement of treatment.

The model for the positive urinalysis group was found to have large improvement of accuracy over the baseline distribution of UTI and thereby the potential for delay and prevention of antibiotics prescriptions. Furthermore, it was found to be a good fit into the current diagnostic process, as clinicians already use the urinalysis to base treatment decisions on. Therefore, we suspected a model for this group was most likely to be of added value in clinical practice.

5.2. Predictive value of the Gram stain in different candidate groups

The Gram stain is not commonly used in the diagnostic process for UTI, as it is not routinely measured in every hospital. In the analysis of the predictive results for the different candidate populations, shown in

Table 6
Different candidate groups for UTI prediction. In the Accuracy column the highest accuracy found among the different supervised classifiers for the corresponding candidate group is shown. The Accuracy without Gram stain column describes the best accuracy found by the supervised models without using the features from the Gram stain.

Candidate group	Total	Labeled	Positive (%)	Negative (%)	Accuracy	Accuracy without Gram stain
1. Culture obtained	11,367	906	266 (29.36)	640 (70.64)	84.01	81.23
2. No Urinalysis	2823	191	35 (18.32)	156 (81.68)	82.47	81.74
3. Urinalysis	8544	715	231 (32.31)	484 (67.69)	84.38	80.62
4. Negative Urinalysis	4647	364	24 (6.59)	340 (93.41)	93.40	93.82
5. Positive Urinalysis	3897	351	207 (58.97)	144 (41.03)	75.63	67.56

Table 6, we found that for some of these groups the Gram stain features played a significant role in the classification of UTIs. This can be seen from the differences in accuracy between the Accuracy and the Accuracy without Gram Stain column. The latter shows for each group the predictive accuracy found by the best model from which all the variables pertaining to the Gram measurement (negative/positive rods/cocci, leukocytes, epithelial cells) were removed.

We observe no noteworthy difference in predictive power for two of the five groups when comparing the model with the Gram stain predictors to the model without these predictors: the group without urinalysis (2) and the group with a negative urinalysis result (4). For these groups, no meaningful distinction could be made on the basis of the other variables available, likely due to the large class imbalance and the most informative predictor (the urinalysis) already being factored out.

An increase in predictive performance can be observed for the three other groups, however. For the group for which the culture was obtained (1), the group for which urinalysis was available (3) and the group with positive urinalysis (5). It is important to note, that group 5 is in its entirety a subgroup of group 3, which in turn is a subgroup of group 1, as shown in 6. When we take the accuracy increase from adding the Gram stain variables to the model in group 5, measured to be 8.07%, and scale it by 351/715 and 351/906, we obtain an increase of 3.96% and 3.13% respectively. The advantages found in groups 3 and 1 were measured to be 3.76% and 2.78%. We hypothesize these differences can thus be entirely explained by the increased predictive performance in the positive urinalysis group.

The Gram stain is found to be of value in the group where the screening step based on urinalysis is already determined to be positive and the risk of UTI relatively high.

6. Predictive results

6.1. Supervised model performance for cultures with positive urinalysis screening

In Table 7 the predictive power of the different supervised models applied to cultures that were found positive by the urinalysis screening is shown. These models used the predictor variables of the Fixed feature set, as this set performed better than the Sparse and Selected sets (see Section 6.5). The Fixed feature set includes: any active antibiotics at time of culture, urine nitrite, urine leukocytes, Gram positive cocci, Gram negative rods, urine collection method, age and sex.

From these results, we observe that the best accuracy is obtained by the RF classifier, while the best AUC is achieved by the SVM classifier. The sensitivity and specificity of RF for predicting UTI in cultures with positive urinalysis were 80.18% and 69.73% respectively.

6.2. Semi-supervised model performance for cultures with positive urinalysis screening: RESSEL

To further improve upon the results obtained by the supervised classifiers, the RESSEL method was applied. In addition to the 351 labeled data points, the 3546 unlabeled data points from the positive urinalysis group were used to increase predictive performance. The results for each of the different supervised classifiers in combination with RESSEL are shown in Table 8.

We observe the maximum accuracy is once more obtained by the RF model and has further improved by 1.14 to be 76.77%. The AUC of the SVM and RF models decreased, while that of the XGB, NN and LR classifiers increased. We are most interested in the accuracy, however, as in practice a single cutoff for prediction will be used. For this reason, the models were optimized for maximum accuracy in the final hyperparameter optimization step as explained in Section 3.5.2. The sensitivity and specificity of the RF-RESSEL model are increased to 81.28% and 70.75% respectively.

In Table D.19 in Appendix D the culture results for the positive urinalysis group are compared to the UTI label assigned by the expert panel. We observe the accuracy for the culture for this group is found to be 76.35. This accuracy is well within the 95% Confidence Interval of the RF-RESSEL classifier and a one sample t-test finds a t-statistic of 0.851 and corresponding p-value $= 0.396$. The sensitivity and specificity of the culture for detecting UTI are found to be 82.61% and 67.36% respectively.

6.3. Predictive performance by patient subgroup

In Section 3.2.2 we distinguished four subgroups for which we suspected model performance might be worse than for the overall population: immunosuppressed patients, elderly ($\geq$75), children ($<$18) and urological patients. For each of these patient groups, we recorded the predictions made during the test set evaluation. The resulting predictive values for patients in these groups who also had a positive urinalysis screening result, are described in Table 9.

The predictive accuracy of the model for each of these subgroups was lower than that of the model on the entire subpopulation with a positive urinalysis screening test. We observe the confidence intervals associated with each of these patient groups were much larger, however, as each of the subsets was by definition smaller than the full set. The differences found were sufficiently small as not to necessitate the exclusion of one or more of these groups from application of the model.

6.4. Performance of the clinical decision support system

The clinical decision support system as a whole, consisting of the screening step based on the urinalysis results, followed by the prediction by the RF - RESSEL model, was found to predict UTI with an accuracy of 85.24% and a specificity and NPV of 91.30% and 87.46% respectively,

Table 7
Prediction results of supervised models on the cultures that are accompanied by a positive urinalysis result. The number within parenthesis represents the 95% Confidence Interval.

	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Classifier						
SVM	74.86 (±0.87)	79.57 (±1.24)	68.66 (±1.50)	78.10 (±1.22)	70.55 (±1.87)	80.43 (±0.95)
XGB	75.41 (±0.87)	79.44 (±1.13)	69.96 (±1.55)	79.02 (±1.15)	70.42 (±1.57)	79.87 (±1.06)
RF	75.63 (±0.86)	80.18 (±1.27)	69.73 (±1.70)	78.99 (±1.28)	71.26 (±1.71)	80.22 (±0.97)
NN	73.57 (±0.84)	74.38 (±1.16)	72.72 (±1.31)	79.49 (±1.14)	66.46 (±1.50)	78.11 (±0.97)
LR	65.36 (±1.55)	71.77 (±2.37)	55.80 (±3.43)	71.36 (±1.64)	57.58 (±2.39)	69.86 (±1.81)

Table 8
Prediction results of supervised models enhanced through RESSEL on the cultures that are accompanied by a positive urinalysis result.The number within parenthesis represents the 95% Confidence Interval.

	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Classifier						
SVM	74.51 (±0.99)	78.83 (±1.31)	69.07 (±1.57)	78.01 (±1.38)	69.98 (±1.88)	78.20 (±0.96)
XGB	76.23 (±0.88)	80.01 (±1.14)	70.94 (±1.47)	79.73 (±1.05)	71.27 (±1.60)	80.05 (±1.08)
RF	76.77 (±0.97)	81.28 (±1.16)	70.75 (±1.85)	79.76 (±1.35)	72.51 (±1.64)	80.02 (±1.00)
NN	73.77 (±0.85)	75.33 (±1.15)	71.84 (±1.33)	79.15 (±1.14)	66.94 (±1.52)	78.59 (±1.03)
LR	73.87 (±0.94)	75.44 (±1.31)	71.72 (±1.67)	79.88 (±1.28)	66.43 (±1.42)	77.47 (±1.06)

Table 9
Predictive performance of the best performing semi-supervised model (RF enhanced with RESSEL) for different patient groups on the positive urinalysis group.

	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Patient group						
Immunosuppressed	75.24 (±1.51)	80.75 (±1.81)	68.06 (±2.83)	77.48 (±1.84)	71.43 (±2.68)	79.88 (±1.55)
Elderly	73.96 (±1.99)	84.89 (±2.32)	59.92 (±3.53)	73.21 (±2.74)	76.32 (±3.81)	75.77 (±2.43)
Children	76.15 (±2.36)	80.75 (±3.06)	70.56 (±4.49)	79.54 (±2.86)	70.84 (±4.64)	78.79 (±2.88)
Urological	75.11 (±2.74)	73.50 (±4.16)	73.47 (±6.02)	83.79 (±3.85)	64.01 (±4.75)	81.91 (±3.65)

as shown in Table 10. The system provides similar predictive accuracy for the subgroups of children and patients on immunosuppressants, while it is less accurate for the elderly and urological patients. These results demonstrate the effectiveness of the system for predicting UTI for cultures that had urinalysis results available, i.e. 715 out of 906 labeled cultures (78.92%).

To determine the effectiveness of the clinical decision support system, we established the negative predictive values of the screening step and combined them with the prediction results from the models seen in Tables 8 and 9 in the following manner:

$$Accuracy = NPV^{-} \cdot \frac{N^{-}}{N} + Acc^{+} \cdot \frac{N^{+}}{N}, \tag{1}$$

$$Sensitivity = Sensitivity^{+} \cdot \frac{N^{+}_{Pos}}{N_{Pos}}, \tag{2}$$

$$Specificity = \frac{Specificity^{+} \cdot N^{+}_{neg} + N^{-}_{neg}}{N_{neg}}, \tag{3}$$

$$PPV = PPV^{+}, \tag{4}$$

$$NPV = NPV^{-} \cdot \frac{N^{-}_{negpred}}{N_{negpred}} + NPV^{+} \cdot \frac{N^{+}_{negpred}}{N_{negpred}}, \tag{5}$$

where the superscript $(-)$ indicates the value is associated with the cultures classified negative by the screening step, the superscript $(+)$ indicates the value is associated with the cultures that were classified as positive in the screening step and thus presented to the model, N is the number of labeled cultures in the group, N_{Pos} the number of UTI positive labeled cultures, N_{neg} the number of UTI negative labeled cultures and $N_{negpred}$ the number of cultures which were predicted to be negative.

While the number of cultures that was found to be UTI positive or negative in each of the two groups was constant, the number of positive and negative predictions differed per iteration. $N^{-}_{negpred}$ is constant however, as it is equal to N^{-}, as all cultures with negative screening result are predicted to be negative. $N_{negpred}$ can be calculated from $N_{negpred} = N^{-}_{negpred} + N^{+}_{negpred}$, so it suffices to calculate an average $N^{+}_{negpred}$ to calculate the average NPV of the system as a whole. $N^{+}_{negpred}$ can be calculated as:

$$\frac{Acc - PPV}{-PPV + NPV}, \tag{6}$$

given $-PPV \neq NPV$. All of these quantities are known for the group with positive screening result, and therefore the total NPV can be calculated. The results of these calculations are shown in Table 10.

The predictive values of the culture for the UTI label are shown in Table D.20 in Appendix D. Compared to the CDSS, the culture has significantly lower accuracy: 77.90% compared to 85.24%. This is also true for its PPV (62.05%–79.76%) and specificity (76.24%–91.30%). In terms of sensitivity, however, the culture outperforms the pipeline for the thresholds selected (81.39%–72.84%) and for NPV as well (89.56%–87.46%).

6.5. Model optimization: variable selection

As part of model optimization, variable selection was applied as described in Section 3.5.2. As each experiment consisted of 100 repetitions, each variable presented to the model was selected between 0 and 100 times during these experiments.

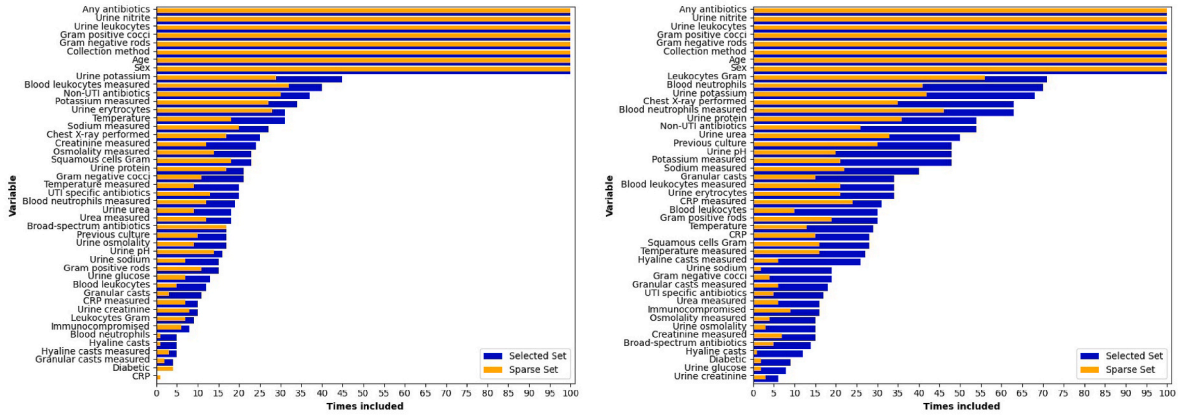
We visualize the selection frequencies of the variables for the RF and SVM classifiers in Fig. 6. The variables contained in the Fixed set were: any active antibiotics at time of culture, urine nitrite, urine leukocytes, Gram positive cocci, Gram negative rods, urine collection method, age and sex. These were included by default. The number of times that a feature was included in the Selected set is shown in blue, the number of times it was included in the Sparse set is shown in orange. Although the Sparse set is by definition smaller or equal to the Selected set in size for each iteration, the number of times an individual feature was included overall in the Sparse set can be larger than the number of times it was included in the Selected set due to the floating option of Mlxtend.

We observe the total number of features included was much larger for the SVM classifier (2007 Selected, 1443 Sparse) than for RF (1466 Selected, 1251 Sparse), for both the Sparse and Selected sets. Interestingly, none of the optional features was consistently selected over the 100 iterations. The most frequently selected feature for RF was potassium in the urine (45 times), followed by whether blood leukocytes were measured (40 times) and active non-UTI antibiotics (both 46 times). In the Sparse set these same three variables were the most included, with 29, 32 and 30 inclusions respectively. In case of the SVM classifier, the most frequently selected features was leukocytes - Gram (71 times), followed by blood neutrophils (70 times) and potassium in the urine (68 times). The most frequently selected features for the Sparse set were leukocytes - Gram (56 times), whether blood neutrophils were measured (46 times) and potassium in the urine (42 times).

In Fig. 7 the relative performance of the different feature subsets is shown for each of the classifiers included in the experiments. In Fig. 7a the results for the supervised models are shown, in Fig. 7b the results for RESSEL are depicted. Feature set selection was limited to the supervised part of the pipeline as discussed in Section 3.5.2.

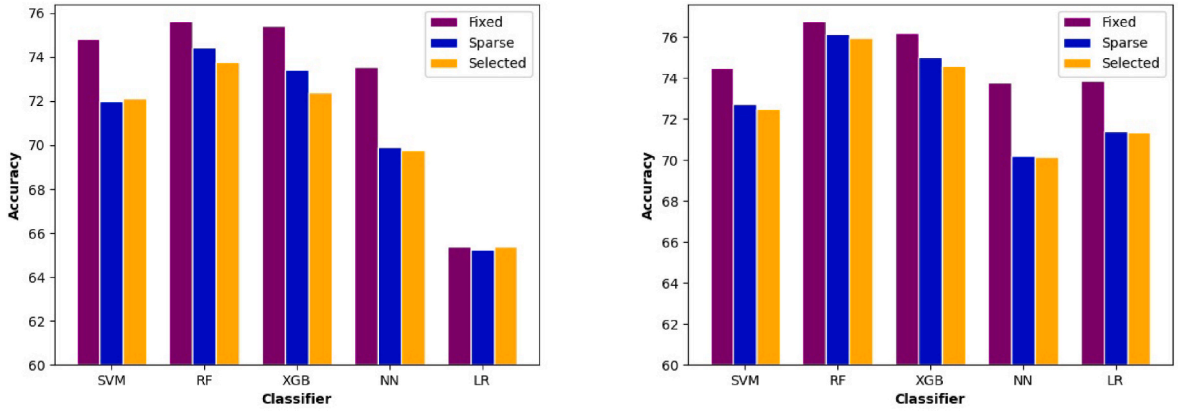
Table 10
Predictive performance of the Clinical Decision Support System for patients for whom urinalysis results were available. The results shown represent the combined performance of the urinalysis screening step with the RF - RESSEL model.

	NPV^{-}	N^{-}_{neg}	N^{-}_{pos}	N^{+}_{neg}	N^{+}_{pos}	Accuracy	Sensitivity	Specificity	PPV	NPV
Patient group										
All	93.41	340	24	144	207	85.24	72.84	91.30	79.76	87.46
Immunosuppressed	94.12	144	9	54	77	84.38	72.30	89.73	77.48	87.70
Elderly	87.04	47	7	38	49	78.97	74.28	82.08	73.21	84.04
Children	91.67	66	6	27	36	84.43	69.21	91.45	79.54	86.37
Urological	88.24	15	2	15	25	79.03	68.06	86.74	83.79	75.93



(a) RF (b) SVM

Fig. 6. Number of times a particular feature was selected through Sequential Feature Selection for the Sparse and Selected feature sets for the (a) RF and (b) SVM classifiers. All features that were selected 100 times belong to the Fixed set and were included by default.



(a) Supervised (b) Semi-supervised: RESSEL

Fig. 7. Accuracies for all of the classifiers trained using the different feature sets using (a) supervised and (b) semi-supervised learning.

We observe that for each of the classifiers, the best results are obtained from the Fixed feature set. The performance of the Sparse and Selected sets is comparable, although the Sparse set performs slightly better. We hypothesize the most predictive features are already included in the Fixed set. While there may be some predictive value in the inclusion of additional individual features, the inclusion of an additional set based on cross validation AUC is not beneficial.

This may be due to the relatively small size of the labeled positive urinalysis data set (n = 351), causing a feature to perform better on the train set, be selected and not perform as well on the test set. Furthermore, if any individual feature had been a strong predictor, we would expect it to be included more frequently in the Selected and Sparse sets.

6.6. Model optimization: hyperparameter selection

The predictive pipeline depicted in Fig. 2 includes a separate hyperparameter selection for the supervised and semi-supervised parts of the pipeline. RESSEL is a computationally heavy method [16], since it both combines multiple classifiers into an ensemble and repeatedly re-trains the individual classifiers during its self-training process. The hyperparameter optimization step further increases the training time, scaling linearly with the number of feature combinations tried and the number of cross validation folds, e.g. a $12 \cdot 12 \cdot 5 = 720$ times increase in time spent for the SVM classifier with the settings used in the experiments compared to a single fit with preset features on a data set of equal size.

We investigate the performance difference of using the hyperparameter settings found earlier, in the supervised hyperparameter selection step, to the use of a separate step for RESSEL. The results for using the supervised hyperparameters are shown in Table 11.

We compare to the results of using a separate hyperparameter tuning step in combination with RESSEL, shown in Table 8. Overall, results are very similar. In terms of both accuracy and AUC, differences are well within each other's confidence intervals. Therefore, at least on this data set, the additional step is not demonstrably of added value and can be omitted to reduce computational costs.

Table 11
Predictive performance of the classifiers enhanced with RESSEL using the best hyperparameters found during supervised optimization.

	Accuracy	AUC
Classifier		
SVM	74.83 (±0.95)	78.74 (±0.98)
XGB	75.93 (±0.86)	79.10 (±0.98)
RF	76.90 (±0.80)	79.82 (±0.98)
NN	73.60 (±0.86)	78.66 (±1.04)
LR	73.83 (±0.95)	77.28 (±1.06)

7. Discussion

7.1. Safety

In current practice the clinician has to decide whether to start antibiotics based on the clinical picture and urinalysis. The CDSS we developed provides the clinician with a prediction for UTI at a stage of the diagnostic process at which currently no additional quantitative information becomes available until the culture results are in, at least a day later. We show the addition of this system, consisting of a screening step based on the urinalysis followed by a machine learning model, leads to a desirable situation in terms of both safety and potential reduction of inappropriate antibiotic prescriptions: at the time the Gram stain results are in, the system is run and provides a prediction as to whether a patient has a UTI with high accuracy and NPV of 85.24% and 87.46% respectively (Table 10). This enables the health care provider to confidently hold off on prescribing antibiotics in case of a no UTI prediction, unless there is another indication that warrants the use of antibiotics. For the population with positive urinalysis result, who are most at risk of UTI and for whom the model is used, we found there to be no statistically significant different in accuracy between our model and the urine culture for predicting UTI, with a p-value of 0.40. In the worst case scenario, the system falsely predicts no UTI, resulting in delayed treatment. In clinical practice, however, if patient health deteriorates antibiotics will likely be administered, regardless of the previously predicted outcome by the system.

7.2. Potential impact estimation

The primary aim of the implementation of a system such as we propose in this paper is to lead to a reduction in the inappropriate prescription of antibiotics. In the following, we analyze the current state of antibiotic use surrounding urine cultures and estimate the size of the impact our system might have.

In order to be able to carry out such an analysis we have had to make some assumptions that we were not able to verify as the data required was not available. The largest assumption is that broad-spectrum antibiotics will be halted in case no UTI is predicted. In practice, these antibiotics will frequently be continued in case of systemic illness, albeit for a different focus. Although this analysis thus provides us with an upper limit to the possible effect size, an early prediction of no UTI does provide the added benefit of aiding the search for an alternative diagnosis.

In the following analysis we make use of three categories of antibiotics, further specified in Table C.18 in Appendix C:

- UTI specific (U): antibiotics within this category are solely used to treat a UTI and are never prescribed for a different focus.
- Broad-spectrum (B): these antibiotics may be prescribed when UTI is suspected, but could also be prescribed for a different focus.
- Not for UTI (N): these should, with few exceptions, never be prescribed if the focus is UTI.

In measuring the impact the application of our system could have on the reduction of unnecessary antibiotics, we are mostly interested in the UTI specific and Broad-spectrum categories. We assume the Other category was not prescribed for suspected UTI and therefore a reduction in its use is not within the scope of this work. It is included in the remainder of this section, however, to provide context.

An overview of the current situation with regard to antibiotic prescriptions surrounding the urine culture, for the group of patients who have had urinalysis performed, is shown in Fig. 8 with respect to UTI and Figure E.9 in Appendix E with respect to the culture result as the UTI label is not available for the majority of the cultures.

Of the 715 labeled cultures included in this group, only 148 did not have any active antibiotic prescriptions registered in the 7 days before or after the date of the culture. Of the 567 that did have a prescription in this time span, 185 (32.63%) were already on antibiotics at the date of the urine culture.

The use of the predictive system proposed in this paper would be able to reduce the antibiotic prescriptions in the group where no antibiotics were active before the culture order, so the 382 cultures. Of these, 261 (68.32%) were prescribed antibiotics on the day of the culture, while only about half end up having a UTI according to the expert labeling. More precisely, of those patients for whom we are certain they were prescribed antibiotics specifically for a UTI (U: 34), 32.35% ended up not having a UTI after all. For the patients who were prescribed antibiotics for which it is uncertain if they were prescribed for a UTI (E: 212) 50.47% ended up not having a UTI.

To provide a rough estimate of the potential in antibiotic reduction, we take the specificity of 91.3% found for the CDSS and multiply it by the percentage of cultures that were ultimately not corresponding to a UTI. Given this specificity, the 32.35% of cultures (11 total) which were falsely prescribed antibiotics of type U could potentially be reduced to 2.81% (~1). The 50.47% of cultures (107 total) which were falsely prescribed antibiotics of type B could potentially be reduced to 4.39% (~9).

Next we apply the same calculations to all culture data within our target population collected in two years, as shown in Figure E.9 in Appendix E, but assuming an equal percentage of UTI as in the labeled data. We assume 32.35% of cultures would be falsely started on antibiotics of type U (~134 out of 413 in total), which could potentially be reduced to ~12. Assuming again 50.47% of cultures (~1293 out of 2561 in total) were falsely prescribed antibiotics of type E, the total number of cultures accompanied by inappropriately started antibiotics could potentially be reduced to ~112.

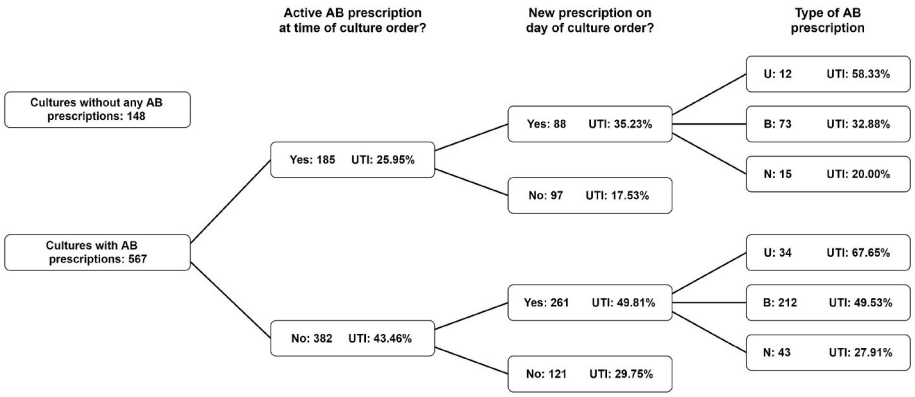


Fig. 8. Analysis of antibiotic (AB) prescriptions surrounding the culture order with UTI as outcome. The percentage indicates the percentage of the corresponding group who was labeled as having an UTI. Antibiotic prescriptions are defined into the classes UTI specific (U), Broad-spectrum (B) and Not for UTI (N).

In total, this would mean inappropriate antibiotic prescriptions could be avoided with high certainty for $134 - 12 = 122$ cultures, in addition to possibly avoiding prescriptions for $1293 - 112 = 1181$ cultures. Combined, these add up to 1303 cultures over the two years of data, out of a total of 8544 cultures that qualify to participate in the CDSS, i.e. a potential reduction of up to 15.2%. In clinical practice this number will be lower, mainly due to antibiotics in the Broad-spectrum category, which might have been prescribed for a different focus, but a significant reduction could be achieved nonetheless.

These rough estimates assumes equal predictive performance for each subgroup antibiotic prescription type, which is an over-simplification. Nevertheless, it is clear that the use of such a system has great potential for the reduction of unnecessarily prescribed antibiotics. Additionally, some cases which are found to be a UTI would be missed by the system. As stated before, however, the culture result will still follow between one and three days after the prediction and the condition of the patient will be monitored during this time as well. Therefore, we believe antibiotic treatment might be slightly delayed, but will seldom be completely discarded in case there is a UTI present.

7.3. Limitations

This research has a number of limitations. First of all, the models are developed and tested on a large part of the hospital population, but we opted to exclude some specific types of patients beforehand. The inability to use the model for every type of patient might hinder its adoption. Additionally, as the system is developed for the specific population of the University Medical Center Utrecht, which has a more complex patient population than the average hospital, it is uncertain how well the model performance will generalize to other medical centers.

Another limiting factor is that the Gram stain is required. While a Gram stain is routinely performed for each urine culture in our laboratory, currently the workflow is not designed to have it performed immediately after the culture order, causing the time before the results are in to vary. For the predictive system described in this work to be effective, the Gram stain needs to be processed more consistently, so health care professionals can rely upon the system to make a prediction within a set time frame after the culture order. Moreover, the Gram stain not routinely performed in every hospital, preventing the application of system as reported in this work.

Furthermore, the number of labeled samples is limited, especially for the group of patients with positive urine screening result. A lot of the labeled data used in this research was used in data analysis from which we came to the conclusion a predictive model for the screening positive group was most advantageous. If the model is to be further improved, a new labeling can focus entirely on this sub group such that every additional label is used for model development. Finally, there was no structured information available about the signs and symptoms. We believe their inclusion would lead to greater predictive accuracy in detecting UTI.

8. Conclusion

In this work we present a study of urinary tract infections (UTI) in an academic center. We use the presence of UTI, established through expert panel review, as gold standard. We compare the culture result with this label and find that the two are only moderately in agreement with each other within the hospital population as a whole, thereby confirming the added value of this more informative study outcome.

We demonstrate the benefit of using the Gram stain results, showing that its predictive value for UTI is increased in patients with positive urinalysis who have previously been deemed to have relatively high risk of a UTI. We then propose a two-step clinical decision support system (CDSS), using a screening step based on the urinalysis followed by a prediction model for high risk patients. We show that a number of supervised models have good accuracy for detecting UTI in this subgroup and that this accuracy can be further improved upon by using the reliable semi-supervised ensemble learning (RESSEL) method to learn from the unlabeled data.

The model used in the second part of the CDSS is evaluated and found to be as accurate as the urine culture is for predicting the expert label for UTI in patients with a positive urinalysis result. The system provides the clinician with this prediction at an earlier stage in the diagnostic process. This enables the clinician to withhold the administration antibiotics, to look for other causes of infection or to recognize asymptomatic bacteriuria in case no UTI is predicted. In an exploratory analysis, we combine the predictive performance of the CDSS as a whole with an analysis of antibiotic prescriptions surrounding urine cultures to provide an estimate of the potential impact the system might have on prevention of inappropriate prescription of antibiotics. We calculated that the use of the system could result in a reduction of antibiotic prescriptions of up to 15.2% for the target population in our setting, although this number is expected be lower in clinical practice.

Future research directions include the inclusion of signs and symptoms from clinical notes, which would require the application of Text Mining techniques. Our next step is to implement and to validate this model and to measure the clinical benefits as well as the risks prospectively in an impact study.

Acknowledgements

This work was supported by the University Medical Center Utrecht. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Hyperparameter Settings

Table A.12
Base classifier hyperparameter settings.

Classifier (parameter)	Value
XGB	
η	[0.1, 0.25, 0.4, 0.55, 0.7, 0.85, 1]
col sample by tree	[0.2,0.4,0.6,0.8,1]
Features	[1, all]
SVM	
Kernel	rbf
C	1E[-2, 0.5, 12 steps]
γ	1E[-2.5, 0, 12 steps]
NN	
Neighbors	[1,25]
RF	

(*continued on next page*)

Table A.12 (*continued*)

Classifier (parameter)	Value
Depth	[1,10]
Features	[1, all]
LR	
Implementation	SGD
Loss	log
Penalty	[L1, L2, elasticnet]

Table A.13
RESSEL hyperparameter settings.

Setting	Value
Unlabeled fraction	0.35
Ensemble size	25
Bootstrap	True
Stratify	False

Appendix B. Clinical Findings

Table B.14
Numerical predictors used in the modeling process.

	N (%)	No UTI	UTI
Variable			
Age (years)	906 (100.0)	57.0 (27.8–71.0)	63.0 (38.0–74.0)
CRP (mg/L)	729 (80.5)	52.0 (15.0–126.0)	55.0 (16.0–116.0)
Blood leukocytes ($x10^9$/L)	780 (86.1)	10.7 (7.5–14.2)	11.3 (8.4–15.9)
Blood neutrophils ($x10^9$/L)	359 (39.6)	8.4 (5.0–11.8)	8.5 (4.8–12.4)
Temperature (°C)	811 (89.5)	37.7 (37.1–38.6)	37.9 (37.2–38.8)
Sodium urine (mmol/L)	97 (10.7)	44.0 (23.0–91.0)	40.5 (29.8–62.0)
Osmolality urine (mOsmol/kg)	54 (6.0)	505.5 (377.2–631.2)	345.0 (268.5–393.5)
Creatinine urine (mmol/L)	106 (11.7)	5.8 (3.6–9.9)	7.0 (4.9–9.4)
Urea urine (mmol/L)	52 (5.7)	197.0 (117.0–245.5)	164.0 (113.8–210.8)
Potassium urine (mmol/L)	40 (4.4)	30.0 (20.0–52.5)	33.0 (23.0–57.0)

Table B.15
Binary predictors used in the modeling process.

Variable	Value	N (%)	No UTI	UTI
Sex	Female	452 (49.9)	301 (66.59)	151 (33.4)
	Male	454 (50.1)	339 (74.67)	115 (25.3)
Collection method	Catheter	378 (41.7)	256 (67.72)	122 (32.3)
	Midstream	528 (58.3)	384 (72.73)	144 (27.3)
Chest X-ray performed	No	577 (63.7)	385 (66.72)	192 (33.3)
	Yes	329 (36.3)	255 (77.51)	74 (22.5)
Immunosuppressed	No	561 (61.9)	392 (69.88)	169 (30.1)
	Yes	345 (38.1)	248 (71.88)	97 (28.1)
Diabetic	No	707 (78.0)	497 (70.3)	210 (29.7)
	Yes	199 (22.0)	143 (71.86)	56 (28.1)
UTI specific antibiotics	No	828 (91.4)	588 (71.01)	240 (29.0)
	Yes	78 (8.6)	52 (66.67)	26 (33.3)
Broad-spectrum antibiotics	No	685 (75.6)	469 (68.47)	216 (31.5)
	Yes	221 (24.4)	171 (77.38)	50 (22.6)
Non-UTI antibiotics	No	788 (87.0)	546 (69.29)	242 (30.7)
	Yes	118 (13.0)	94 (79.66)	24 (20.3)
Any antibiotics	No	564 (62.3)	382 (67.73)	182 (32.3)
	Yes	342 (37.7)	258 (75.44)	84 (24.6)

Table B.16
Predictors from urinalysis used in the modeling process.

Variable	Value	N (%)	No UTI	UTI
Leukocytes	Not Measured	191 (21.1)	156 (81.68)	35 (18.32)
(/microL)	Negative	304 (33.6)	287 (94.41)	17 (5.6)
	ca. 25	70 (7.7)	57 (81.43)	13 (18.6)
	ca. 75	71 (7.8)	49 (69.01)	22 (31.0)
	ca. 250	36 (4.0)	20 (55.56)	16 (44.4)
	ca. 500	234 (25.8)	71 (30.34)	163 (69.7)
Nitrite	Not Measured	191 (21.1)	156 (81.68)	35 (18.32)
	Negative	628 (69.3)	466 (74.2)	162 (25.8)
	Positive	87 (9.6)	18 (20.69)	69 (79.3)
Protein	Not Measured	191 (21.1)	156 (81.68)	35 (18.32)
(g/L)	Negative	215 (23.7)	176 (81.86)	39 (18.1)
	Trace	96 (10.6)	71 (73.96)	25 (26.0)
	ca. 0.15	3 (0.3)	1 (33.33)	2 (66.7)
	ca. 0.20	97 (10.7)	64 (65.98)	33 (34.0)
	ca. 0.30	62 (6.8)	39 (62.9)	23 (37.1)
	ca. 0.50	53 (5.8)	29 (54.72)	24 (45.3)
	ca. 0.70	61 (6.7)	36 (59.02)	25 (41.0)
	ca. 1.00	43 (4.7)	19 (44.19)	24 (55.8)
	ca. 1.50	22 (2.4)	12 (54.55)	10 (45.5)
	ca. 2.00	20 (2.2)	11 (55.0)	9 (45.0)
	>2.00	43 (4.7)	26 (60.47)	17 (39.5)
Erythrocytes	Not Measured	191 (21.1)	156 (81.68)	35 (18.32)
free Hb	Negative	316 (34.9)	258 (81.65)	58 (18.4)
(/mircoL)	ca. 10	73 (8.1)	48 (65.75)	25 (34.2)
	ca. 20	54 (6.0)	29 (53.7)	25 (46.3)
	ca. 30	64 (7.1)	38 (59.38)	26 (40.6)
	ca. 60	72 (7.9)	34 (47.22)	38 (52.8)
	ca. 150	55 (6.1)	33 (60.0)	22 (40.0)
	ca. 300	42 (4.6)	23 (54.76)	19 (45.2)
	>300	39 (4.3)	21 (53.85)	18 (46.2)
Glucose	Not Measured	191 (21.1)	156 (81.68)	35 (18.32)
	Negative	628 (69.3)	421 (67.04)	207 (33.0)
	Trace	16 (1.8)	12 (75.0)	4 (25.0)
	Positive	18 (2.0)	15 (83.33)	3 (16.7)
	Strong Positive	53 (5.8)	36 (67.92)	17 (32.1)
pH	Not Measured	191 (21.1)	156 (81.68)	35 (18.32)
	5.0	12 (1.3)	11 (91.67)	1 (8.3)
	5.5	300 (33.1)	208 (69.33)	92 (30.7)
	6.0	202 (22.3)	132 (65.35)	70 (34.7)
	6.5	102 (11.3)	64 (62.75)	38 (37.3)
	7.0	66 (7.3)	46 (69.7)	20 (30.3)
	7.5	24 (2.6)	20 (83.33)	4 (16.7)
	8.0	7 (0.8)	3 (42.86)	4 (57.1)
	9.0	2 (0.2)	0 (0.0)	2 (100.0)
Hyaline casts	Not Measured	742 (81.9)	531 (71.56)	211 (28.44)
	1–2	37 (4.1)	20 (54.05)	17 (45.9)
	>2	127 (14.0)	89 (70.08)	38 (29.9)
Granular casts	Not Measured	799 (88.2)	560 (70.09)	239 (29.91)
	None	5 (0.6)	3 (60.0)	2 (40.0)
	1–2	49 (5.4)	36 (73.47)	13 (26.5)
	>2	53 (5.8)	41 (77.36)	12 (22.6)

Table B.17
Predictors from urine culture and Gram stain used in the modeling process.

Variable	Value	N (%)	No UTI	UTI
Culture result	Negative	532 (58.7)	483 (90.79)	49 (9.2)
	Positive	374 (41.3)	157 (41.98)	217 (58.0)
Number of bacteria	1	755 (83.3)	555 (73.51)	200 (26.5)
	2	126 (13.9)	68 (53.97)	58 (46.0)
	3	19 (2.1)	13 (68.42)	6 (31.6)
	4	5 (0.6)	3 (60.0)	2 (40.0)
	5	1 (0.1)	1 (100.0)	0 (0.0)
Leukocytes Gram	Not Measured	2 (0.2)	2 (100.0)	0 (0.0)
staining (/100x)	<1	323 (35.7)	288 (89.16)	35 (10.8)
	1–9	211 (23.3)	166 (78.67)	45 (21.3)
	10–25	147 (16.2)	92 (62.59)	55 (37.4)
	>25	223 (24.6)	92 (41.26)	131 (58.7)
Squamous cells	Not Measured	25 (2.8)	18 (72.0)	7 (28.0)
Gram staining (/100x)	<1	664 (73.3)	485 (73.04)	179 (27.0)
	1–9	171 (18.9)	109 (63.74)	62 (36.3)

(*continued on next page*)

Table B.17 (*continued*)

Variable	Value	N (%)	No UTI	UTI
	10–25	25 (2.8)	16 (64.0)	9 (36.0)
	>25	21 (2.3)	12 (57.14)	9 (42.9)
Gram negative rods (/1000x)	Not Measured	528 (58.3)	461 (87.31)	67 (12.7)
	<1	69 (7.6)	59 (85.51)	10 (14.5)
	1–5	91 (10.0)	63 (69.23)	28 (30.8)
	6–30	87 (9.6)	28 (32.18)	59 (67.8)
	>30	131 (14.5)	29 (22.14)	102 (77.9)
Gram negative cocci (/1000x)	Not Measured	901 (99.4)	638 (70.81)	263 (29.2)
	<1	1 (0.1)	1 (100.0)	0 (0.0)
	1–5	1 (0.1)	0 (0.0)	1 (100.0)
	6–30	1 (0.1)	0 (0.0)	1 (100.0)
	>30	2 (0.2)	1 (50.0)	1 (50.0)
Gram positive rods (/1000x)	Not Measured	818 (90.3)	580 (70.9)	238 (29.1)
	<1	27 (3.0)	21 (77.78)	6 (22.2)
	1–5	38 (4.2)	24 (63.16)	14 (36.8)
	6–30	16 (1.8)	10 (62.5)	6 (37.5)
	>30	7 (0.8)	5 (71.43)	2 (28.6)
Gram positive cocci (/1000x)	Not Measured	648 (71.5)	480 (74.07)	168 (25.9)
	<1	74 (8.2)	57 (77.03)	17 (23.0)
	1–5	90 (9.9)	67 (74.44)	23 (25.6)
	6–30	55 (6.1)	21 (38.18)	34 (61.8)
	>30	39 (4.3)	15 (38.46)	24 (61.5)
Previous culture (common uropathogens, cfu)	Not Measured	721 (79.6)	517 (71.71)	204 (28.29)
	No Relevant Bacteria	101 (11.1)	83 (82.18)	18 (17.8)
	$< 10^3$	2 (0.2)	2 (100.0)	0 (0.0)
	10^3	2 (0.2)	1 (50.0)	1 (50.0)
	10^4	15 (1.7)	9 (60.0)	6 (40.0)
	10^5	65 (7.2)	28 (43.08)	37 (56.9)

Appendix C. Antibiotic Classes

Table C.18
Antibiotic Classes. The different antibiotic classes are defined as: UTI specific (U), Broad-spectrum (B) and Not for UTI (N).

ATC code	name	class
J01EA01	trimethoprim	U
J01XE01	nitrofurantoin	U
J01XX01	fosfomycin	U
J01CR02	amoxicillin and beta-lactamase inhibitor	B
J01EE01	sulfamethoxazole and trimethoprim	B
J01MA02	ciprofloxacin	B
J01DD04	ceftriaxone	B
J01GB03	gentamicin	B
J01DH02	meropenem	B
J01DH51	imipenem and cilastatin	B
J01CR05	piperacillin and beta-lactamase inhibitor	B
J01DC02	cefuroxime	B
J01MA06	norfloxacin	B
J01MA01	ofloxacin	B
J01DH03	ertapenem	B
J01DI54	ceftolozane and beta-lactamase inhibitor	B
J01AA12	tigecycline	B
J01DD08	cefixime	B
J01FA10	azithromycin	N
J01CA04	amoxicillin	N
J01DB04	cefazolin	N
J01XA01	vancomycin	N
J01FA01	erythromycin	N
J01XD01	metronidazole	N
J01CF05	flucloxacillin	N
J01DD02	ceftazidime	N
J01FF01	clindamycin	N
J01XB01	colistin	N
J01AA02	doxycycline	N
J01GB01	tobramycin	N
J01CE05	pheneticillin	N
J01DD01	cefotaxime	N
J01CE01	benzylpenicillin	N
J01MA14	moxifloxacin	N
J01MA12	levofloxacin	N
J01XA02	teicoplanin	N
J01DB01	cefalexin	N

(*continued on next page*)

Table C.18 (*continued*)

ATC code	name	class
J01FA09	clarithromycin	N
J01DD14	ceftibuten	N
J01DF01	aztreonam	N
J01DC04	cefaclor	N
J01GB06	amikacin	N
J01CE02	phenoxymethylpenicillin	N
J01CE08	benzathine benzylpenicillin	N
J01AA08	minocycline	N
J01XX09	daptomycin	N
J01XX08	linezolid	N
J01XX05	methenamine	N
J01XC01	fusidic acid	N
J01EC02	sulfadiazine	N
J01DI02	ceftaroline fosamil	N

Appendix D. Culture Results

Table D.19
Culture result compared to UTI label for all cultures with positive associated urinalysis.

UTI	Neg	Pos	All	P (%)	N (%)	PPV	NPV	Sensitivity	Specificity	Accuracy
Culture										
Neg	97	36	133	27.07	72.93	78.44	72.93	82.61	67.36	76.35
Pos	47	171	218	78.44	21.56	-	-	-	-	-
All	144	207	351	58.97	41.03	-	-	-	-	-

Table D.20
Culture result compared to UTI label for all cultures with associated urinalysis.

UTI	Neg	Pos	All	P (%)	N (%)	PPV	NPV	Sensitivity	Specificity	Accuracy
Culture										
Neg	369	43	412	10.44	89.56	62.05	89.56	81.39	76.24	77.9
Pos	115	188	303	62.05	37.95	-	-	-	-	-
All	484	231	715	32.31	67.69	-	-	-	-	-

Appendix E. Antibiotic Impact Analysis - Culture as Outcome

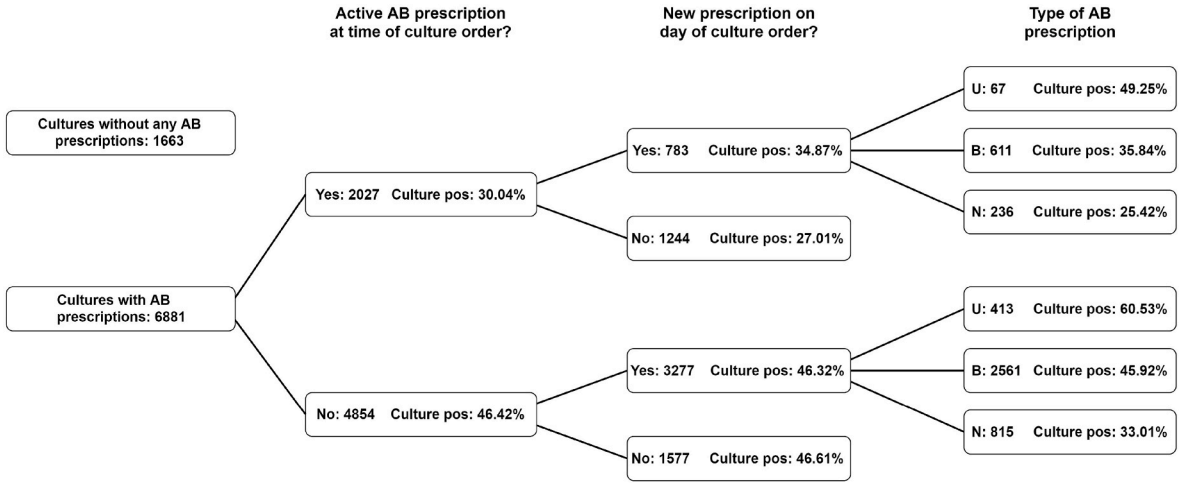


Figure E.9. Analysis of antibiotic (AB) prescriptions surrounding the culture order with the urine culture as outcome. The percentage indicates the percentage of the corresponding group who was labeled as having an UTI. Antibiotic prescriptions are defined into the classes UTI specific (U), Broad-spectrum (B) and Not for UTI (N).

References

[1] W.E. Stamm, S.R. Norrby, Urinary tract infections: disease panorama and challenges, J. Infect. Dis. 183 (Supplement 1) (2001) S1–S4.

[2] A.L. Flores-Mireles, J.N. Walker, M. Caparon, S.J. Hultgren, Urinary tract infections: epidemiology, mechanisms of infection and treatment options, Nat. Rev. Microbiol. 13 (5) (2015) 269–284.

[3] S.L. Bermingham, J.F. Ashe, Systematic review of the impact of urinary tract infections on health-related quality of life, BJU Int. 110 (2012) E830–E836.

[4] Department of Health, Lord Carter of Coles, Report of the review of nhs pathology services in england, retrieved on: 31-12-2021, https://www.networks.nhs.uk/nhs-networks/peninsula-pathology-network/documents/CarterReviewPathologyReport.pdf, 2006. URL.

[5] M. Bouma, S.E. Geerlings, S. Klinkhamer, B.J. Knottnerus, T.N. Platteel, E. A. Reuland, H.S. Visser, R.J. Wolters, Nhg standaard: urineweginfecties, retrieved on: 31-12-2021 (2020). URL, https://richtlijnen.nhg.org/standaarden/urineweginfecties.

[6] CDC, NHSN, Urinary Tract Infection (Catheter-associated Urinary Tract Infection [cauti] and Non-catheter-associated Urinary Tract Infection [uti]) and Other Urinary System Infection [usi]) Events, Surveillance Definitions for Specific Types of Infections.

[7] M.L. Terpstra, S.E. Geerlings, C. van Nieuwkoop, E.P. van Haarst, H. Boom, B. J. Knottnerus, W. Rozemeijer, C.J. de Groot, C.M.P.M. Hertogh, P.D. van der Linden, et al., Swab Guidelines for Antimicrobial Therapy of Urinary Tract Infections in Adults, 2020.

[8] T.E.B. Johansen, H. Botto, M. Cek, M. Grabe, P. Tenke, F.M.E. Wagenlehner, K. G. Naber, Critical review of current definitions of urinary tract infections and proposal of an eau/esiu classification system, Int. J. Antimicrob. Agents 38 (2011) 64–70.

[9] N.C. Tan, A.Y.L. Koong, L.P. Ng, P.L. Hu, E.Y.L. Koh, K.T. Tan, P.K.S. Moey, M. X. Tan, C.S. Wong, T.Y. Tan, et al., Accuracy of urinary symptoms and urine microscopy in diagnosing urinary tract infection in women, Fam. Pract. 36 (4) (2019) 417–424.

[10] D. Medina-Bombardó, M. Seguí-Díaz, C. Roca-Fusalba, J. Llobera, What is the predictive value of urinary symptoms for diagnosing urinary tract infection in women? Fam. Pract. 20 (2) (2003) 103–107.

[11] P. Little, S. Turner, K. Rumsby, G. Warner, M. Moore, J.A. Lowes, H. Smith, C. Hawke, D. Turner, G.M. Leydon, A. Arscott, M. Mullee, Dipsticks and diagnostic algorithms in urinary tract infection: development and validation, randomised trial, economic analysis, observational cohort, and qualitative study, Health Technol. Assess. 13 (19) (2009) 1–96.

[12] G. Bonkat, R. Bartoletti, F. Bruyère, et al., Eau guidelines: urological infections, the European association of UrologyRetrieved on: 31-12-2021, URL, https://uroweb.org/guideline/urological-infections/#3.

[13] M.E.J.L. Hulscher, R.P.T.M. Grol, J.W.M. Van Der Meer, Antibiotic prescribing in hospitals: a social and behavioural scientific approach, Lancet Infect. Dis. 10 (3) (2010) 167–175.

[14] B.G. Bell, F. Schellevis, E. Stobberingh, H. Goossens, M. Pringle, A systematic review and meta-analysis of the effects of antibiotic consumption on antibiotic resistance, BMC Infect. Dis. 14 (1) (2014) 1–25.

[15] W.H. Organization, et al., Antimicrobial Resistance: Global Report on Surveillance, World Health Organization, 2014.

[16] S. de Vries, D. Thierens, A reliable ensemble based approach to semi-supervised learning, Knowl. Base Syst. 215 (2021), 106738.

[17] R.S. Wigton, V.L. Hoellerich, J.P. Ornato, V. Leu, L.A. Mazzotta, I.-H.C. Cheng, Use of clinical findings in the diagnosis of urinary tract infection in women, Arch. Intern. Med. 145 (12) (1985) 2222–2227.

[18] D. Kim, S.C. Oh, C. Liu, Y. Kim, Y. Park, S.H. Jeong, Prediction of urine culture results by automated urinalysis with digital flow morphology analysis, Sci. Rep. 11 (1) (2021) 1–8.

[19] R.J. Burton, M. Albur, M. Eberl, S.M. Cuff, Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections, BMC Med. Inf. Decis. Making 19 (1) (2019) 171.

[20] P.S. Heckerling, G.J. Canaris, S.D. Flach, T.G. Tape, R.S. Wigton, B.S. Gerber, Predictors of urinary tract infection based on artificial neural networks and genetic algorithms, Int. J. Med. Inf. 76 (4) (2007) 289–296.

[21] A.A.H. Gadalla, I.M. Friberg, A. Kift-Morgan, J. Zhang, M. Eberl, N. Topley, I. Weeks, S. Cuff, M. Wootton, M. Gal, et al., Identification of clinical and urine biomarkers for uncomplicated urinary tract infection using machine learning algorithms, Sci. Rep. 9 (1) (2019) 1–11.

[22] R.A. Taylor, C.L. Moore, K.-H. Cheung, C. Brandt, Predicting urinary tract infections in the emergency department with machine learning, PLoS One 13 (3) (2018), e0194085.

[23] L. Meister, E.J. Morley, D. Scheer, R. Sinert, History and physical examination plus laboratory testing for the diagnosis of adult female urinary tract infection, Acad. Emerg. Med. 20 (7) (2013) 631–645.

[24] Y.-m. Li, J.-h. Xu, Y.-x. Zhao, Predictors of urinary tract infection in acute stroke patients: a cohort study, Medicine 99 (27).

[25] S. Enshaeifar, A. Zoha, S. Skillman, A. Markides, S.T. Acton, T. Elsaleh, M. Kenny, H. Rostill, R. Nilforooshan, P. Barnaghi, Machine learning methods for detecting urinary tract infection and analysing daily living activities in people with dementia, PLoS One 14 (1) (2019), e0209909.

[26] G. Turra, N. Conti, A. Signoroni, Hyperspectral image acquisition and analysis of cultured bacteria for the discrimination of urinary tract infections, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2015, pp. 759–762.

[27] V.S. Kodogiannis, J.N. Lygouras, A. Tarczynski, H.S. Chowdrey, Artificial odor discrimination system using electronic nose and neural networks for the identification of urinary tract infection, IEEE Trans. Inf. Technol. Biomed. 12 (6) (2008) 707–713.

[28] C. Sandoval, B. Sinaki, R. Weiss, J. Munoz, M.F. Ozkaynak, O. Tugal, S. Jayabose, Urinary tract infections in pediatric oncology patients with fever and neutropenia, Pediatr. Hematol. Oncol. 29 (1) (2012) 68–72.

[29] M.L. Wilson, L. Gaido, Laboratory diagnosis of urinary tract infections in adult patients, Clin. Infect. Dis. 38 (8) (2004) 1150–1158.

[30] J. Klastersky, J. De Naurois, K. Rolston, B. Rapoport, G. Maschmeyer, M. Aapro, J. Herrstedt, Management of febrile neutropaenia: Esmo clinical practice guidelines, Ann. Oncol. 27 (2016) v111–v118.

[31] L.E. Nicolle, K. Gupta, S.F. Bradley, R. Colgan, G.P. DeMuri, D. Drekonja, L. O. Eckert, S.E. Geerlings, B. Köves, T.M. Hooton, et al., Clinical practice guideline for the management of asymptomatic bacteriuria: 2019 update by the infectious diseases society of America, Clin. Infect. Dis. 68 (10) (2019) e83–e110.

[32] WHO Collaborating Centre for Drug Statistics Methodology Norwegian Institute of Public Health, Atc/ddd index 2021, retrieved on: 31-12-2021, https://www.whocc.no/atc_ddd_index/, 2021. URL.

[33] S. Raschka, Mlxtend: providing machine learning and data science utilities and extensions to python's scientific computing stack, J. Open Source Softw 3 (24), retrieved on: 31-12-2021. doi:10.21105/joss.00638.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.