Universiteit Leiden

Master Computer Science

Exploring the Predictability of Bacterial Vaginosis in Reproductive Age Women using Longitudinal Vaginal Microbe Abundance and pH Data

Name: Stijn Oudshoorn
Student ID: s1925458

Date: August 12, 2022

Specialisation: Bioinformatics

1st supervisor: Dr. Lu Cao
2nd supervisor: Dr. CJ Jenkins

Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Bacterial vaginosis (BV) is the most abundant form of vaginitis, and most common in women of reproductive age. The inflammation it brings gives discomfort to the patient through symptoms of itch, odor, thin white-greyish vaginal discharge and a burning sensation. It also leads to a 2.7-fold increased risk of various STI's, including HIV, chlamydia, herpes, pelvic inflammatory disease and gonorrhea. In pregnant women, BV adds to the probability of the occurrence of a miscarriage and a preterm birth.

A dataset created by Ravel, Brotman, et al. (2013) is used, containing longitudinal data with bacteria abundances, pH, menstruation phase and BV status. It contains 25 women measuring every day for 70 days. First a BV diagnosis test was done with machine learning classifier models as well as a probabilistic dummy classifier. An attempt was done to create more meaningful labels than the moment of diagnosis that is provided in the dataset originally. These labels are based on: a set of rules; adjacent days; days of treatment. The best performance on a test set in terms of f1-score on the diagnosis for these various newly defined was 0.62, with a sensitivity of 0.46.

During the prediction tasks these labels were used to see if the patient gets BV in the next 10 or 30 days, or the next month with 30 days. BV in the next 10 days could be predicted with an f1-score of 0.70 and a sensitivity of 0.57.

For the following 30 days the performance was an f1-score of 0.93 and a sensitivity of 0.94.

For the next 30 days, skipping ahead 30 days, the performance was an f1-score of 0.89 and a sensitivity of 0.93.

A regression task was performed to predict the number of days until BV onsets, which was done with a highest R^2 of 0.75.

To find what bacteria and other vaginal conditions cause this condition, the feature importances were extracted from the best performing models. Most findings were in line with recent literature, where BV associated bacteria like Gardnerella vaginalis and Bifidobacteriaceae were found to be important for short term BV prediction. For longer term BV prediction, Lactobacillus crispatus was found to be more important than the iners variant. This confirms that there is a difference in conditions leading to long term BV risk versus short term BV risk.

Acknowledgements

I want to acknowledge Dr. CJ Jenkins for her mentoring during this project and the amazing time she has given me and my fellow students at Devoteam. I also want to thank Dr. Lu Cao for her realism and regular checks, which helped me to push this thesis over the finish line. It is also thanks to the flexibility in her schedule that I managed to get everything done in time. Dr. Katy Wolstencroft I want to thank for helping me to get things back on track halfway through, and for helping me make a decision on where to do my internship. I acknowledge Dr. Nele Mentens for her support and understanding when my project took an unexpected turn. Thanks to Dr. Roxana Hickey for her advice on the topic for this thesis. Thanks to all other students at Devoteam, and also to Julia Märak Leffler for the amazing coffee breaks with all of you, that were much needed to get this thesis done.

Contents

Chapter 1

Introduction

1.1 The Human Vaginal Microbiome

1.1.1 An Introduction to Microbiomes

Humans have known of the existence of microbes since Antony van Leeuwenhoek observed protozoa in 1676 (Dobell 1920), which he found to be present inside our bodies too shortly after, while inspecting his own stool. We now know that the human body is covered by bacteria amounting to a total of ten times more prokaryotic than human cells (Savage 1977; T. Dobzhansky and T. G. Dobzhansky 1971; Luckey 1972). The various locations of these microbiota include but are not limited to the creases in our skin surface, gut, urinary tract, oral cavity, as well as the vagina. The composition of such microbiota varies widely depending on the environment that arises in these body parts and changes dynamically through time as well (Costello et al. 2009). In this thesis, the scope is set on the reproductive age women's vaginal microbiome, a term that encompasses microbes (prokaryotes, eukaryotes, archaea and viruses), their genes as well as their surrounding environment properties (Marchesi and Ravel 2015).

1.1.2 The Healthy Vaginal Microbiome

The human vagina contains anywhere between 10^{10} and 10^{11} bacteria (C. Chen et al. 2017), forming an ecosystem that is homeostatic and mutually serves the host and the bacteria. The environment is made nourishing, warm and moist for the bacteria to live in, so these microbes can in turn fend off intruding microbes through antimicrobial factors. It has been shown that communities of various bacteria are better at resisting intruders than monocultures (Burmølle et al. 2006). In the case of the vaginal microbiome, core functions were identified which can be conserved despite varying presence of the bacterial taxa since there is no core composition (Ravel, Gajer, et al. 2010). For example, the lactic acid and H_2O_2 produced by vaginal *lactobacilli* lowers the pH to around 4.5, which is believed to suppress pathogenic growth (Linhares et al. 2011; O'Hanlon, Moench, and Cone 2011). Therefore, the presence of these bacteria are associated with a healthy state. It should be noted that the absence of lactobacilli does not necessarily mean that conditions are unhealthy, since other genera of bacteria can also produce acidic compounds to achieve similar protection.

1.1.3 Bacterial Vaginosis

Bacterial Vaginosis (BV) is the most prevalent form of vaginitis, a broad term for inflammation of the vagina. The term vaginitis also includes, for example, candidiasis and trichomoniasis (Elgantri, Mohamed, and Ibrahim 2010). These reproductive tract infections and other infections, like sexually transmitted infections (STI), are more common in and also more damaging to women than men (HL 2006; Philip, Benjamin, and Sengupta 2013). BV alone globally accounts for an annual cost of 13.8 billion US dollars considering all cases (Peebles et al. 2019). A BV infection occurs in an estimate of 10-12% of reproductive age women in the UK (Carol A. Spiegel et al. 1980; Hay et al. 1994), but an even higher prevalence of 15 to 20% is found in pregnant women (Martius et al. 1988). In the USA, studies have shown higher prevalence of BV in pregnant women too ranging between 15% and 30%, where proportionally more cases are found in African American women (Goldenberg et al. 1996).
The cause for a BV infection is mainly the result of surroundings. Research has indicated that the partner penile microbiome plays a role in BV risk (Supriya D. Mehta et al. 2020). For BV to arise, a combined occurrence of certain bacteria is necessary (Z. S. Ma and Ellison 2021). Gardnerella vaginalis presence should be accompanied by a secondary colonizer, like *A. vaginae* or *Sneathia spp.*, for symptomatic BV onset (Muzny et al. 2020).
In cases of BV, there is an increased risk of infertility, endometritis, and pelvic inflammatory disease (Ravel, Moreno, and Simón 2021). BV also causes a 2.7-fold risk increase for STI's like HIV, chlamydia, herpes and gonorrhea (Allsworth and Peipert 2011). Pregnant women with BV have an increased probability of preterm birth (Martius et al. 1988) and having a miscarriage (Al-Memar et al. 2020). Besides this, symptoms include itching, unusual white-grey thin vaginal discharge, strong odor and a burning sensation.
There are various ways to diagnose someone with BV, based on either a nugent score (positive when it is 7 or higher) (Nugent, Krohn, and Hillier 1991; Carol A Spiegel, Amsel, and Holmes 1983), found using a Gram Stain of vaginal fluid, or by Amsel criteria (Richard Amsel et al. 1983; Eschenbach et al. 1988), which involves multiple subjective clinical tests of which at least 3 should be positive to diagnose someone with BV.
Once BV is diagnosed, antibiotics that target the anaerobic bacteria responsible for the BV infection can be used to treat the patient in order to restore the microbiota's balance towards the Lactobacillic bacteria. An example being *Metronidazole*, although this treatment is known to be prone to recurrence (Deng et al. 2018). A more effective way to fight BV is by preventing the infection through use of probiotics (Donders, Zodzika, and Rezeberga 2014).
While this thesis will be limited to using data of genomics, the importance of transcriptomics should be noted in this research field, since some gene activation by Gardnerella vaginalis is correlated to recurrence after metronidazole treatment. This same correlation can not be found by examining genetic abundance (Twin et al. 2013).

1.2 Related Work

Ravel, Brotman, et al. (2013) introduced a new dataset of 25 women whose vaginal microbiome are quantified for a period of 10 weeks. This dataset is further described and explored in Chapter 2. This thesis will use this data to perform experiments on. In their paper, Ravel et al. describe findings on the observations of correlations between certain bacterial taxa abundance and BV initiation or recurrence after treatment. Since the quantitative results are a result of the DNA sequencing of 16S rRNA genes, therefore based on DNA abundance and not RNA, the data is a representation of the functional potential and not the activity that is found for each species of bacteria (France et al. 2022).

A similar set of data has been used to classify woman as BV infected or healthy in previous (Beck and Foster 2014). Since this work showed that solving this classification task accurately is feasible, this thesis will expand on this work by exploring prediction in the future.

Classification or even prediction of BV using microbiota data, could result in earlier detection or even prediction of BV, which could modernise the treatment process (Sharma et al. 2021). On a population basis, microbiota data can be used to group women in four cervicotype (CT) groups, based on which bacterium is dominant. Each CT-group is associated with a certain probability (or risk) of BV (Supriya Dinesh Mehta et al. 2020). Using Markov chain simulations, one can predict on a population scale which group transitions take place, and therefore, how abundant certain risk groups are (Munoz et al. 2021). This knowledge of BV probability for a CT group, as well as the transitions in the future, will be used to define a baseline performance on classifying women as BV positive or healthy, and for predicting BV in the future, respectively.

1.3 Goal and Problem Definitions

This thesis aims to reveal to what extend the vaginal microbiome composition provides enough information for Machine Learning (ML) algorithms to predict the diagnosis and the onset of a case of BV. Inspiration for this question also came from the idea of the predictive qualities of the penile microbiome composition, as described by Supriya D. Mehta et al. (2020). The vaginal microbiome composition has already displayed it's predictive qualities when used by ML algorithms to detect BV in a patient in the paper by Beck and Foster (2014).

In order to answer this question, the dataset as introduced by Ravel, Brotman, et al. (2013) will be used. This dataset was chosen based on it's longitudinal nature (women are monitored for approximately 70 days), as well as the completeness in representation of BV cases, like: asymptomatic BV (ABV); symptomatic BV (SBV); and healthy subjects. Firstly, the diagnosis task of detecting BV will be performed on this dataset. Then, a prediction task will be defined for various amounts of time in the future: longer term and shorter term. Finally a regression task will be defined to find the days until BV onsets.

In this thesis, several attempts will be made to create a BV label that carries the meaning: symptomatic BV is present. In the dataset, the diagnosis of SBV is available as a label. Since a disease onsets before a diagnosis occurs and remains until a treatment is performed, strategies are built around this knowledge and further described in Chapter 2.

These labels and the data, as well as the original SBV diagnosis label, will be used to train and test the performance of several ML algorithms, described in Chapter 3.

The approach of CT group transition probability as described by Munoz et al. (2021) will be used as a baseline model. This baseline is described in more detail in Chapter 3 as well.

The following goals and hypotheses will be considered during the conduction of the experiments.

To inspect whether ML techniques can perform better than the baseline in the task of diagnosis on this dataset with various definitions of BV (SBV, rule based BV (RBV), adjacency based BV (AdBV)). SBV is expected to be very hard to predict, namely due to the low number of data points labeled as such available, as well as the factor that human decisions on when during the onset of BV the diagnosis is made can vary widely and may therefore not be related to the microbiota data. The other custom labels are expected to be more easily predicted, since they will be present more abundantly. They are also expected to carry more actual information on the state of the patient than a moment of diagnosis.

To observe how well one can use various ML and probabilistic techniques to predict SBV, like regression to fit on the number of days until BV onsets or binary classification of BV occurrence in the future.

Find relevant or important features that the ML models use and see if they align with current literature or if there is new patterns emerging. Expected important features for diagnosis are different from prediction in future. From the population based research with CT transitions we know that Lactobacillus crispatus plays a more important role in long term prevention of BV risk than Lactobacillus iners (Munoz et al. 2021). It is also expected that we encounter BV associated bacteria as important features such as A. vaginae, Megaspherea, Bifidobacteriaceae, Sneathia spp., Prevotella amnii, Eggerthella and Gardnerella vaginalis (X. Chen et al. 2021).

Chapter 2

Data

2.1 Description

The dataset gathered by Ravel, Brotman, et al. (2013) comprises of measurements performed on 25 women for a period ranging from 70 to 72 days. For some days, one or more measurements are missing, therefore the number of datapoints per woman vary from 65 to 72. The occurrence of each measurement frequency is plotted in Figure 2.1. The total number of samples is 1756, where 1651 are complete with all measurements and labels.

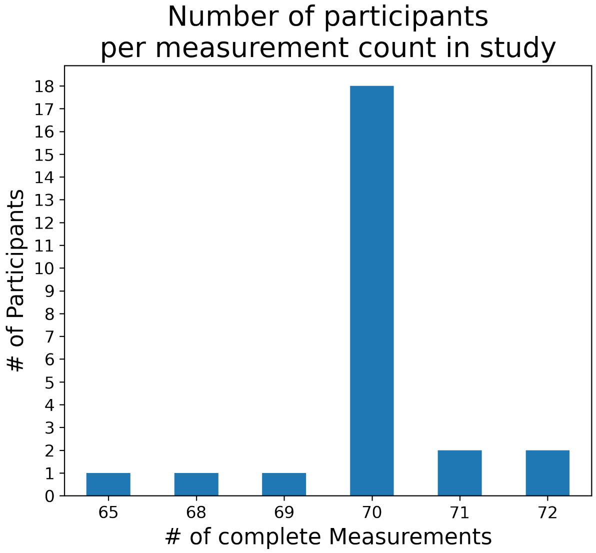


Figure 2.1: The frequency of occurrence for each number of measurement days.

In the next section, the methods of dealing with incomplete measurements are described, as well as the resulting new composition of the data. In Section 4.1, the data is explored on a deeper level. Section 2.2 describes the way that missing labels and measurements are handled, as well as techniques on how new labels were created to add informative value to existing labels.

Table 2.1: The count and frequency of the labels for the dataset as it was initially provided by Ravel, Brotman, et al. (2013), but with manually added labels as described in Table 2.2.

	Total	NBV	SBV	ABV
Count	1651	743	15	21
%	100	45.0	0.9	1.3

2.2 Preprocessing

The two tables, one containing the label data, the other measurement data, are outer merged on the sampleID column, such that no row is left out when either sample is missing. As a result, samples that were present in one of the tables but not in the other will result in missing data in the merged table.

2.2.1 Missing Labels

Compared to the Figure S1 in Additional file 4 of the paper by Ravel, Brotman, et al. (2013), the dataset as provided by the authors is missing several SBV and treatment labels. In Table 2.2 the missing labels, which were added to the dataset manually, are described.

Table 2.2: The SampleID and label type for data alterations to make the set align with the depicted data in Figure S1 of Additional file 4 (Ravel, Brotman, et al. 2013).

SampleID	Added Label
s5.w10d7	SBV
s40.w8d6	SBV
s40.w8d7	BV Med
s40.w9d1	BV Med
s40.w9d2	BV Med
s40.w9d3	BV Med

2.2.2 Missing Measurements

There are also samples present where only certain variables were not reported, like, for example, the pH, some or all of the bacteria abundances. To maximize the amount of available data, the choice was made to interpolate for the missing values. However, some missing values are found on the first or last days in the study, which makes it impossible to interpolate from two surrounding data points. Therefore, these values should be removed (samples are described in Table 2.3), while others are interpolated linearly. To fill both the missing single variable, as well as the entirely missing taxa relative abundances, forward linear interpolation is used. It should be noted that backward would have worked similarly, since the edge cases were removed in the previous preprocessing step.

Table 2.3: Samples that were removed due to the absence of measurement values in the bacteria abundances or pH, excluding the ones that could be interpolated linearly using the neighboring values. This means that these are either the first or last samples taken for the subjects research period.

SampleID	**Missing column(s)**
s3.w1d1	pH
s3.w10d4	Bacteria
s3.w10d5	Bacteria
s3.w10d6	Bacteria
s3.w10d7	Bacteria
s5.w1d1	Bacteria
s130.w10d6	Bacteria
s130.w10d7	Bacteria

The new counts for complete data points and corresponding labels are shown in Table 2.4.

Table 2.4: The count and percentage of the complete data point labels for the same dataset as in Table 2.1, but with manually added labels as described in Table 2.2. The interpolations are also included in these counts as complete data points.

	Total	**NBV**	**SBV**	**ABV**
Count	1748	713	17	21
%	100	40.6	1.0	1.2

2.3 Alterations and Additions

Since the dataset only includes labels for the moment of diagnosis of BV, as well as treatment days, some methods are used to expand the number of labels, while preserving a medical relevance. In the next few sections, each of these label types is described, as well as how they were created.

2.3.1 Rule Based BV

The Rule Based BV (RBV), is an attempt to use logical rules to define which days around a Symptomatic BV diagnosis, the BV is already or still present in a patient. These are the rules, based on which the decision was made to add an RBV label to the dataset:

1. If there is no SBV label yet, but BV treatment is performed, an SBV label will be added on the first day of treatment.
2. If SBV is diagnosed, but treatment starts one or more days later, the days between diagnosis and treatment including the first day of treatment will be given an SBV label.
3. If SBV is diagnosed, but no treatment is started after the diagnosis, the following days to the diagnosis that consecutively have 2 or more symptoms are labelled with SBV.
4. All days preceding an SBV diagnosis of which all consecutively have 2 or more symptoms.

2.3.2 Adjacency Based BV

The Adjacency Based BV (AdBV) label is defined as the day of diagnosis and a number of days before and after this day. Two number of days and a variant with and without the inclusion of treatment days were used to make a total of four variants of this label. This includes the labels with 7 adjacent days - meaning: the day of diagnosis; 3 days before diagnosis; and 3 days after diagnosis - and those with 11 adjacent days, in the same way as the former but with 5 days on either end of the day of diagnosis. Each variant is split in two variants. In the first variant the labeled days that BV treatment is performed are stripped of their label, while in the other variant these labels are kept the same.

Chapter 3

Methods & Materials

3.1 Train & Test Data

The dataset is split into a train and test set, with 80 and 20 percent of the data respectively, using the `train_test_split` function from Sklearn (Pedregosa et al. 2011). The split is stratified, in order to maintain the original ratio of positive and negative labels in both the train and test set.

Besides the stratified split, a class weight can be assigned during training to account for a smaller presence of the positive label as was found in Section 4.1. This parameter, along with others as described in Section 3.2.6, are varied and an optimum is searched for with the `GridSearchCV` function, also from Sklearn. Experiments are performed in 5-fold.

3.2 Machine Learning Models

A method is called machine learning when it makes use of data in order to improve performance on a task, by adjusting it's prediction according to training data, without being explicitly programmed to do so. This may occur in a supervised or unsupervised manner. In the former technique, the machine learning algorithm is provided with an input and the desired output in order to find certain rules or parameters so it can generalize the input to match the required output. The latter is not provided with any label, so the algorithm has to find data patterns that provide structure in the input all by itself.

In both cases, the algorithm is trained on a set of data that is separated from the set that performance is tested on. Depending on the task at hand, this training uses different loss functions to process the difference between the given prediction and the expected prediction into the model, so it can improve for next predictions.

The tasks that are defined are a binary classification task, and a regression task.

Here binary classification is the task to label a subject on a given day as either BV-positive (1) or BV-negative (0). This is also the case for the predictions of BV occurring in the future.

The regression task is to find the number of days after which a subject will get BV on a given day.

In the next subsections the various ML models that were used are described, as well as the non-ML baseline model based on label probability. The latter will be used as a comparison for minimal performance required to the ML models. The choice of the ML models was made as to cover both unsupervised and supervised learners, linear models as well as more complex state-of-the-art models for smaller datasets.

3.2.1 Gradient Boosted Classifier & Regressor (gbc/gbr)

Gradient boosting makes use of an ensemble of weak predictors, called decision trees, to form a classification or regression model. The optimization in a boosting manner is powered by a gradient descent along a loss function that can be differentiated. Instead of bagging, which is used in Random Forest, which trains each model in the ensemble equally, boosting trains models in a way that allows new models to learn from the mistakes that older models made (T. Chen and Guestrin 2016). As a result, newer models continue to evolve and eventually create a powerful ensemble. The additive training method is modified in the XGBoost version used in this chapter. The successor tree's output integrates the results of all predecessors. This model makes use of supervised learning, using the correct answer to influence how it will classify datapoints in the future during training.

3.2.2 Support Vector Machine (SVM)

A SVM seeks to create a hyperplane between the two classes in a binary classification task to maximize the margins between the hyperplane and the two classes (Cortes and Vapnik 1995). This supervised learning model is linear by default, but the kernel can be changed to use a polynomial, sigmoid or radial basis function (rbf). These different kernels allow the model to separate data that is not linearly separable.

3.2.3 k-Nearest Neighbor (k-NN)

The k-nearest neighbors algorithm (k-NN) in statistics was created by Evelyn Fix and Joseph Hodges (Fix and Hodges 1989) and later improved by Thomas Cover (Cover and Hart 1967). It is a non-parametric supervised learning technique. Regression and classification are two uses for it. The input in both situations consists of a dataset's k closest training samples. Whether k-NN is applied for classification or regression determines the results.

The result of k-NN classification is a class membership. The class that an object is allocated to based on the majority vote of its k closest neighbors is determined by the item's neighbors. The result of k-NN regression is the object's property value. Which is determined as the average value of the k nearest neighbors.

3.2.4 Baseline Probabilistic Model

For the baseline model, the frequency of labels in the dataset for each CT-group, taken from a training set, is used as a BV probability parameter p. To perform a prediction in the test set, a number n is generated between 0 and 1. If the generated number n is smaller than p for the CT of that data point, a BV label 1 is outputted. Otherwise, 0 is outputted.

For predictions of BV in the future, CT group transition probabilities are taken into account for the calculation of future BV probability. This was done by counting the occurrences of CT-transitions for the training set. For each datapoint we can now find a vector v (1x4) with the probability of transitioning for each CT-group and multiply this with the vector w (4x1) of BV probabilities for each CT-group, to find the future BV probability f. Then, the same number generation is applied to determine the output of the model.

This classifier functions as a dummy classifier to determine a minimum performance that our other models should acquire to be better than guessing. The idea is based on the Markov Chain model made by Munoz et al. (2021), combined with the Dummy Classifier class from Sklearn (Pedregosa et al. 2011).

3.2.5 Labels to Predict

In this section the labels that are meant to be predicted by the ML models are defined.

BV in Future Binary Classification

For each data point we create new labels to indicate if the test subject gets BV in the future for 3 different time frames. In the following 10 days; in the following 30 days; and in the next month, meaning a gap of 30 days and the month that follows after is considered.
This is a binary label for classification, and was applied to every newly introduced BV type as well as the original SBV label.

Days Until BV Onsets

For a data point, we look ahead in time if the patient gets BV, and if she does, then the data point will get the label of the number of days until she gets BV. This value is zero if the patient has BV in the data point itself. In the regression task to predict the number of days until BV occurs some labels can not be given, since there may not be any case of BV anymore or at all. Therefore, the maximum duration of the study is used as a label value, which is 72 days.
This label was also created using each of the newly introduced BV types as well as the original

3.2.6 Parameter Grid Search

For the classification tasks, the parameters in the grid searches that were performed are shown for the gbc, k-NN, SVM and the linearSVM in Tables 3.1, 3.2, 3.4 and 3.3, respectively.

Table 3.1: The parameters that were grid-searched in the experiment for the gbc model.

Parameter	**Values**
Learning rate	[0.01,0.02,0.03]
Subsample	[0.9, 0.5, 0.2]
# estimators	[100, 500, 1000]
Max depth	[4, 6, 8]

For the regression tasks, the loss parameters for linearSVM regression models are different. All other models were grid searched as described previously, but the loss function grid search for the linearSVM was changed to the following two values: epsilon insensitive and squared epsilon insensitive.

Table 3.2: The parameters that were grid-searched in the experiment for the k-NN model.

Parameter	**Values**
n neighbors	[3, 5, 7, 11]
Weights	[uniform, distance]
Algorithm	[ball tree, kd tree]
Leaf size	[15, 20, 30, 40, 50, 60]
p	[1, 2]

Table 3.3: The parameters that were grid-searched in the experiment for the svm model.

Parameter	**Values**
Kernel	[rbf, poly, sigmoid]
Gamma	[scale, auto]
Max iterations	[10000, 20000]
C	[0.1, 1, 10, 100, 1000]
Class weight	[balanced]

Table 3.4: The parameters that were grid-searched in the experiment for the linearSVM model.

Parameter	**Values**
Loss	[hinge, squared hinge]
Dual	[True, False]
Max iterations	[10000, 20000]
C	[0.1, 1, 10, 100, 1000]
Class weight	[balanced]

3.2.7 Performance Metrics

To measure performance several metrics are used, which will be defined in this section. The abbreviations and visualization of the terms used in this section are shown in Table 3.5.

Table 3.5: A general example of a confusion matrix for diagnostic applications.

		Predicted	
		0	1
Actual	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

This confusion matrix shows the standard examples that can occur in terms of diagnostic performance, which will be used to calculate various metrics in the next subsections.

Sensitivity

Sensitivity, also referred to as recall or true positive rate, describes the probability of a positive prediction, given that the sample actually is positive. In this case, a patient is labelled as 'has BV', while she actually does have BV. Mathematically, this is described as shown in Equation 3.1, where true positive is TP and false negative is FN.

$$sensitivity = \frac{TP}{TP + FN} \tag{3.1}$$

This metric is most useful to increase performance for in this case, since the priority should be to find all the cases in advance, even if this results in some false positives. If performance in this metric can be increased while maintaining a low number of false positives, in other words, while maintaining a high precision, then this is even better.

Precision

The precision metric is used to indicate how many of the instances found are actually relevant. Precision is calculated as shown in Equation 3.2, where TP means true positive and FP stands for false positive.

$$precision = \frac{TP}{TP + FP} \tag{3.2}$$

Since being sensitive to finding BV cases is more important than being precisely sure that all the cases we find are actually BV, this metric is not the highest priority. However, a situation where we find all cases, but with a very high number of false positives should also be prevented. Therefore, this metric is also reported on, and the next metric, the F1-score, combines this precision metric with the sensitivity.

F1-score

The F1-score is defined as the harmonic mean of the sensitivity and the precision, and calculated as shown in Equation 3.3.

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3.3}$$

Since the F1-score captures the trade-off between recall and precision well, this metric can be considered as the main objective to perform high on. In a case where several models perform equally well in the F1-score, the one with the highest sensitivity should then be considered as the best one.

The F1-score can also be calculated with the negative label in mind. This metric, which will be referred to as the negative F1 (neg F1), will be used to evaluate how well healthy cases can be identified or predicted, together with the next metric: specificity.

Specificity

The specificity is essentially the same as sensitivity, where the negative sample is considered instead of the positive. Equation 3.4 describes how this is calculated.

$$specificity = \frac{TN}{TN + FP} \tag{3.4}$$

This metric shows how well a healthy person can be identified as such.

Mean Absolute Error (MAE)

The mean absolute error is the average taken for all data points of the absolute for the predicted value minus the actual value. This is defined as shown in Equation 3.5.

$$MAE = \frac{\sum_{n=1}^{n} |predicted - actual|}{n} \tag{3.5}$$

This value is reported on and optimized for in the regression task.

Mean Squared Error (MSE)

The mean squared error is the average taken for all data points of the square for the predicted value minus the actual value. This is defined as shown in Equation 3.6.

$$MSE = \frac{\sum_{n=1}^{n} (predicted - actual)^2}{n} \tag{3.6}$$

This value is also reported on and optimized for in the regression task.

R^2

R^2, or the coefficient of determination, is defined as the proportion of the output's variance. It describes how well a model is fit on the data.

R^2 is calculated using the residual sum of squares (SS_{res}) and the total sum of squares (SS_{tot}). The former being the sum of the squared errors made by the model and the latter being the sum of the squared difference between the actual values and the mean of all values.

Equations 3.7, 3.8 and 3.9 show the calculation for SS_{res}, SS_{tot}, and how these are used to calculate R^2. Actual value is indicated as y, the predicted value as f.

$$SS_{res} = \sum_{n=1}^{n}(y_n - f_n)^2 \tag{3.7}$$

$$SS_{tot} = \sum_{n=1}^{n}(y_n - \overline{y})^2 \tag{3.8}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{3.9}$$

This is the last value that is reported on and optimized for in the regression task.

3.3 Software & Data Availability

Requests for the code that was used in the conduction of the experiments can be sent through email to Stijn Oudshoorn. The same goes for the preprocessed data used during the experiments. The original data was made public by Ravel, Brotman, et al. (2013).

Chapter 4

Results

4.1 Data Exploration

During the exploratory phase of the thesis research, some gbc and gbr models were trained to predict the number of days where nugent BV occurred in the future in two ways: BV in the next month and BV the following 2 months. While the exploration was done without the same pre-processing steps as in the experiments that are described in the next section, some findings are still noteworthy.

Age was the most important feature in the gradient boosting regression models, this could be a form of overfitting since the model could learn which measurement belongs to which patient by their age. Therefore this feature was removed, which did not result in a decrease in model performance (R^2). Exploration and other experiments continued without age as a feature.

For the model predicting the days of nugent BV next month, that performed as shown in Figure 4.1, the feature importances stand out: Gardnerella vaginalis, expectedly scored highest, 0.27; and Bifidobacteriaceae, appears second unexpectedly with a score of 0.14.

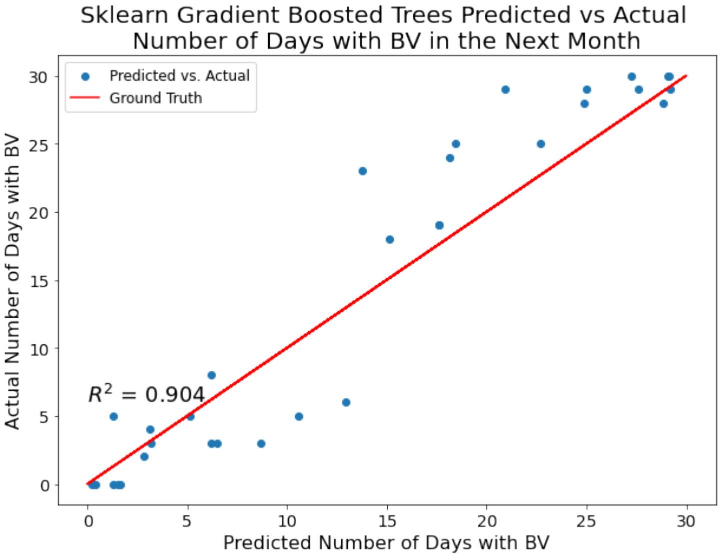


Figure 4.1: The predicted vs the actual number of days that nugent BV occurs in the next month.

The model predicting nugent BV in the following 60 days, the performance of which is shown in Figure 4.2, shared the Bifidobacteriaceae being second place in the feature importances with a value of 0.11, but more interestingly had Parvimonas micra as the most important feature with a value of 0.30. This is in line with earlier findings suggesting that Parvimonas micra could be an important biomarker for BV (Z. S. Ma and Ellison 2021).

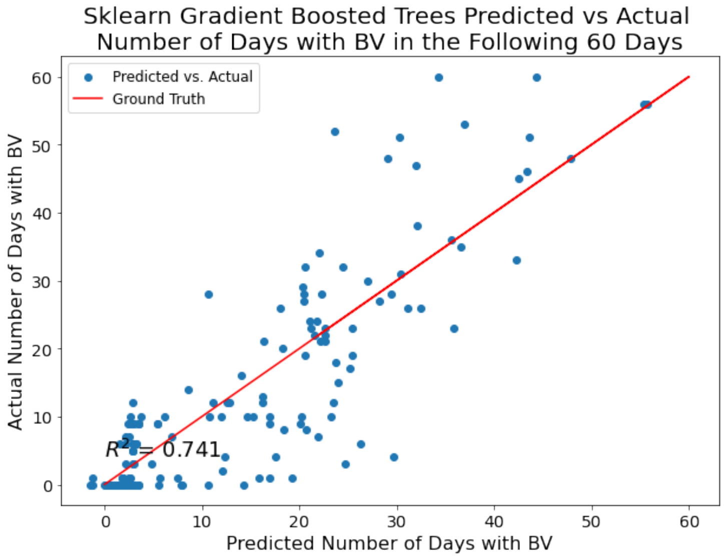


Figure 4.2: The predicted vs the actual number of days that nugent BV occurs in the next 2 months.

Although the NBV exploration does show an acceptable performance and some logical feature importances, the label itself is not strictly connected to symptomatic BV, which is what we attempt to predict in this thesis. Therefore, the other newly created BV labels are used in the remaining experiments.

To explore how well the new variations of BV labels and healthy data points are separable from each other, a t-SNE visualization (Maaten and Hinton 2008) is made using a Standardscalar from Sklearn (Pedregosa et al. 2011) with default settings on the preprocessed data. The results are depicted in Figure 4.3.

As could be expected, in this reduced dimensionality, the healthy and BV labels do not appear linearly separable. Different clusters are visible where similar data points are of the same label. However, other clusters also contain both positive and negative labels.

4.2 BV Diagnosis

The diagnosis experiment results are described here, for each of the label types as defined earlier. The data that was used from here on out is the preprocessed version. From the results of earlier research by Beck and Foster (2014), the hypothesized outcome is that the ML models outperform the guessing baseline classifier. Firstly, this results of the baseline model are presented. Then, the best machine learning model performance reported on, as well as how they were found in the grid search. Lastly the features that were most important to this model will be shown, if this model type allows for feature importance extraction.

The remaining results for the same experiment, that did not perform the best, are shown in Appendix A.

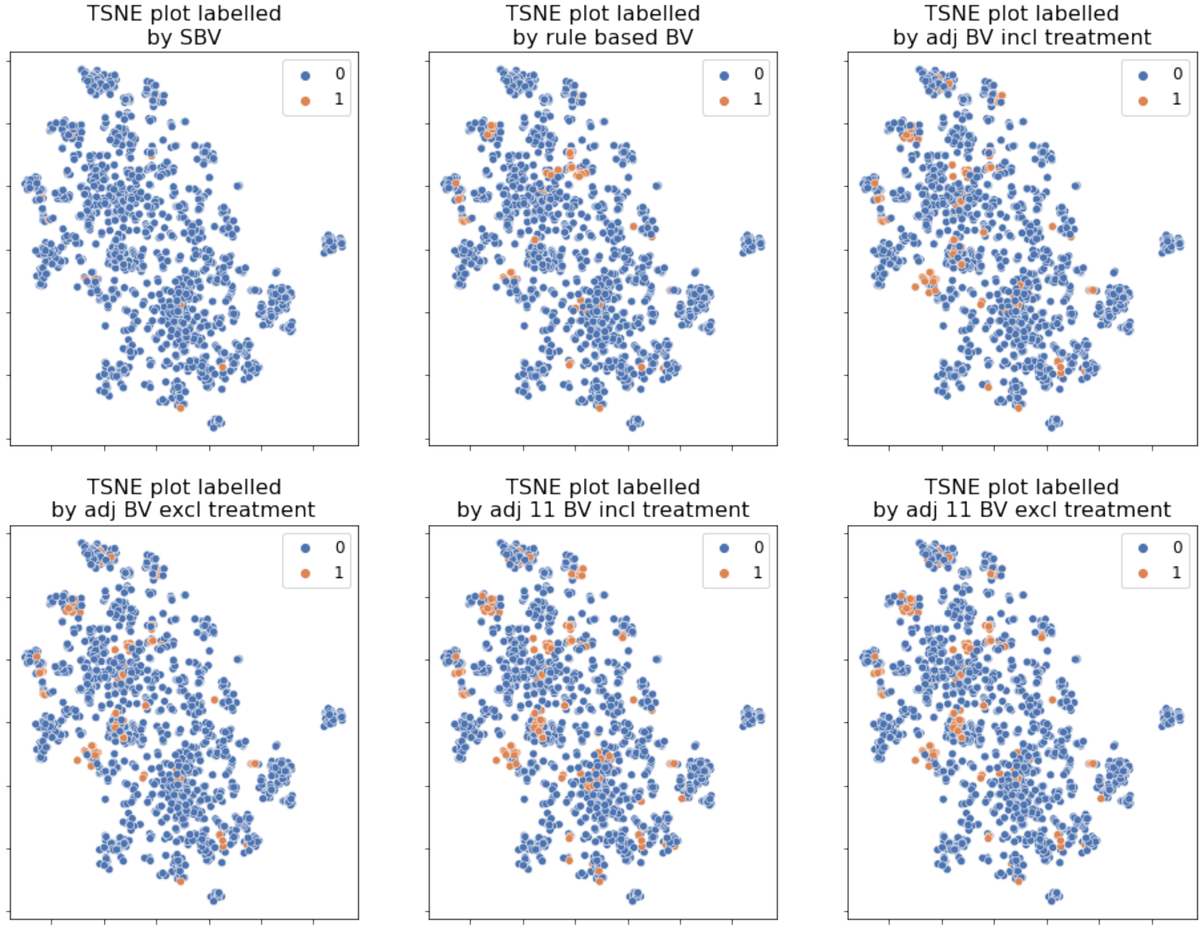


Figure 4.3: T-SNE dimensionality reduction plots of the standard scaled numerical data in 6-fold, where in each plot the hue of healthy and BV diagnosed women is based on a different labeling rule.

4.2.1 Baselines

In Table 4.1 the highest sensitivity performance results of the BV-probability CT-group association classifier in a 5-fold experiment are depicted for each label type.

Table 4.1: The best baseline performance metrics on the diagnosis task.

label	F1	neg_F1	Sensitivity	Precision	Specificity	Test_size
SBV	0.0	0.99	0.0	0.0	0.99	350
RBV	0.08	0.96	0.07	0.08	0.97	350
7 ABVin	0.15	0.95	0.13	0.19	0.96	350
7 ABVex	0.06	0.95	0.05	0.06	0.95	350
11 ABVin	0.11	0.90	0.11	0.11	0.90	350
11 ABVex	0.11	0.92	0.11	0.11	0.92	350

Intuitively, the more frequent a label is, the higher the probability based guesser should score on it. This is not the case for the label of 11 adjacent BV days including and excluding treatment, when compared to it's counterpart with just 7 adjacent days where treatment days are included. The F1-score is 0.04 lower for both where it was expected to be higher.

4.2.2 Model Performance

The highest performing model and label combination on this task is the gbc on 11 adjacent days of SBV including the treatment days. The results, among other models on the same task, are shown in Figure 4.4.

All models score a higher F1 score than the baseline model does on the same label. Though the linear and non-linear support vector machine both sacrifice F1 score performance on the healthy label, resulting in a lower score than the baseline. The best results by the gbc model were achieved using the parameters as shown in Table 4.2.

Table 4.2: The parameters with which the gbc performed best of all models on the task of diagnosing BV.

Learning Rate	Max Depth	# estimators	subsample
0.03	8	500	0.9

Other label performance results are shown in Figures A.1 and A.2 of Appendix A for the baselines and the machine learning models, respectively.

Feature Importances

Since the best model was a gbc, the feature importances were extracted and the 5 most important ones their values are visualized in Figure 4.5.

The remaining feature importances from the gbc models trained on the other label types are depicted in Appendix A, Figure A.3.

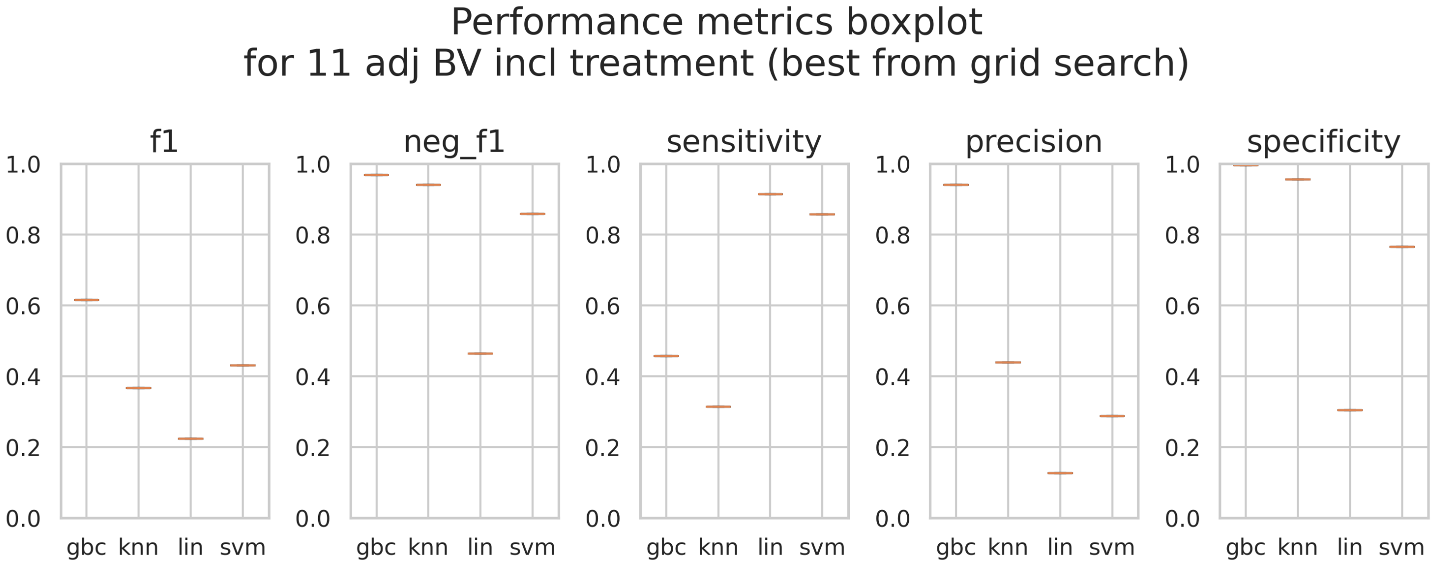


Figure 4.4: The performance metrics for the best performing models on the task of diagnosing BV where the surrounding 11 days of a SBV diagnosis are labelled as BV, including days where treatment takes place.

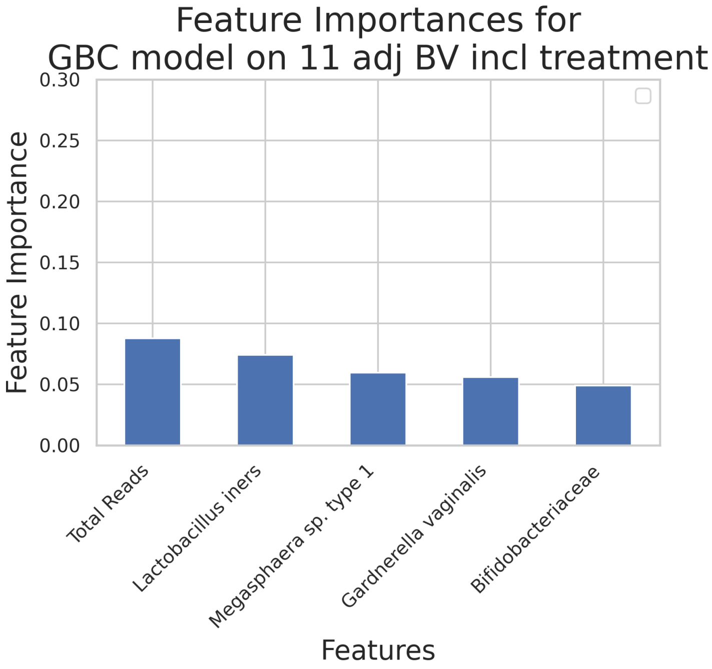


Figure 4.5: The feature importances for the best performing gbc model and label combination which scored as shown in Figure 4.4 on the diagnosis task.

4.3 BV Prediction

In this section, the results of the experiments where prediction of BV is done are described. The approaches are split in a classification and a regression task. The classifiers attempt to predict the presence of BV in various time ranges, while the regressors do the same for the continuous integer values of the number of days until BV onsets. Both are described in more detail in Section 3.2.5. Here too the baseline model results are depicted first, followed by the ML models performances. Then, the best parameters found in the grid search are given, as well as the corresponding feature importances if available for the model type.
Again, the remaining results for this experiment, that were not the best performing, can be found in Appendix A.

4.3.1 BV Prediction 10 days in the future

Baselines

In Table 4.3 the highest sensitivity performance results of the BV-probability CT-group association classifier in a 5-fold experiment are depicted for each label type.

Table 4.3: The best baseline performance metrics on the 10 days ahead prediction task.

label	F1	neg_F1	Sensitivity	Precision	Specificity	Test_size
SBV	0.00	0.99	0.00	0.00	0.99	350
rule_based_BV	0.07	0.96	0.07	0.06	0.96	350
7 ABVin	0.09	0.94	0.09	0.09	0.94	350
7 ABVex	0.10	0.94	0.11	0.09	0.94	350
11 ABVin	0.25	0.92	0.26	0.25	0.91	350
11 ABVex	0.21	0.93	0.21	0.20	0.93	350

Intuitively, the more frequent a label is, the higher the probability based guesser should score on it. This is not the case for the label of 7 adjacent BV days including treatment, when compared to it's counterpart where treatment days are excluded. The F1-score is 0.01 lower in where it was expected to be higher.

Model Performance

The highest performing model and label combination on this task is the gbc on the rule based BV label, regarding F1-score. The results, as well as that of the other models on the same task, are shown in Figure 4.6.
All models, except the linear support vector machine, score a higher F1 score than the baseline model does on the same label. A noteworthy outlier is the high sensitivity found by the svm, though it scores significantly lower in precision while achieving this. The best results by the gbc model were achieved using the parameters as shown in Table 4.2.
Other label performance results are shown in Figures A.4 and A.5 of Appendix A for the baselines and the machine learning models, respectively.

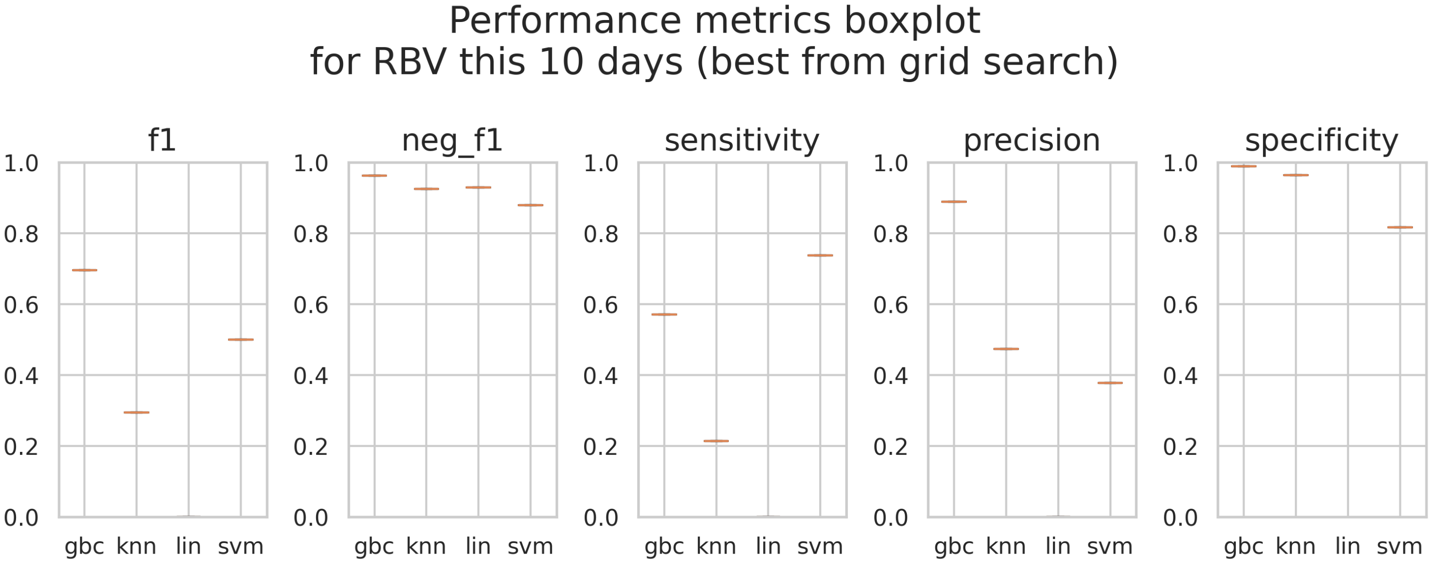


Figure 4.6: The performance metrics for the best performing model F1-score wise, the gbc, and the other ML models on the task of predicting BV where the rule based BV diagnoses are labelled as BV.

Learning Rate	**Max Depth**	**# estimators**	**subsample**
0.03	6	1000	0.9

Table 4.4: The parameters with which the gbc performed best of all models on the task of predicting BV in the next 10 days.

Feature Importances

Since the best model was a gbc, the feature importances were extracted and the 5 most important ones their values are visualized in Figure 4.7.

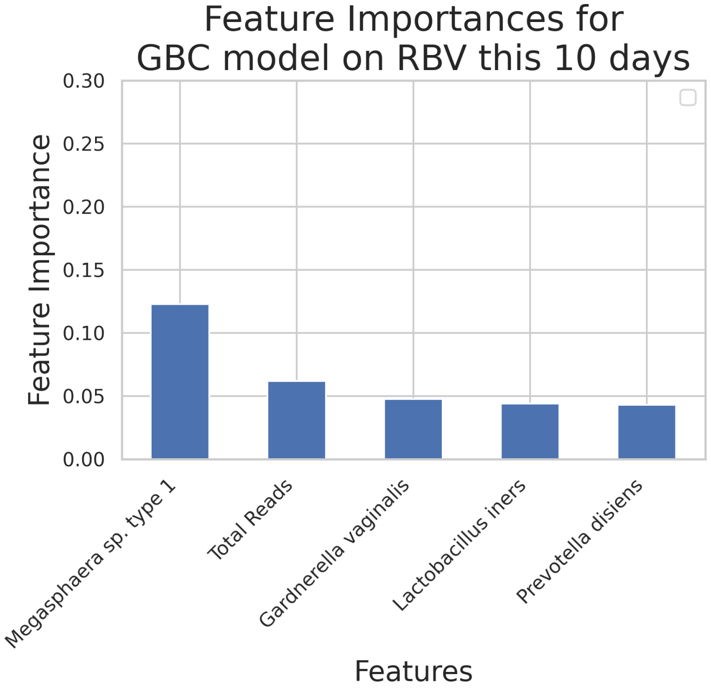


Figure 4.7: The feature importances for the best performing gbc model and label combination which scored as shown in Figure 4.6 on the BV prediction task for the next 10 days.

The remaining feature importances from the gbc models trained on the other label types are depicted in Appendix A, Figure A.6.

4.3.2 BV Prediction this month

Baselines

In Table 4.5 the highest sensitivity performance results of the BV-probability CT-group association classifier in a 5-fold experiment are depicted for each label type.
The results are in line of the expectations considering the frequency of each label.

Model Performance

The highest performing model and label combination on this task is the gbc on the label for 11 adjacent days of BV excluding the treatment days, which goes for all the considered metrics. The results, as well as that of the other models on the same task, are shown in Figure 4.8.
Only the gbc model performs higher in the F1-score, while maintaining a higher negative F1-score than the baseline model. The parameters found in the grid search that resulted in this performance are shown in Table 4.6.
Other label performance results are shown in Figures A.7 and A.8 of Appendix A for the baselines and the machine learning models, respectively.

Table 4.5: The best baseline performance metrics on the one month ahead prediction task.

label	F1	neg_F1	Sensitivity	Precision	Specificity	Test_size
SBV	0.00	0.99	0.00	0.00	1.00	350
rule_based_BV	0.06	0.96	0.07	0.06	0.95	350
7 ABVin	0.12	0.93	0.13	0.10	0.92	350
7 ABVex	0.06	0.95	0.05	0.06	0.95	350
11 ABVin	0.16	0.90	0.17	0.15	0.89	350
11 ABVex	0.15	0.91	0.18	0.13	0.89	350

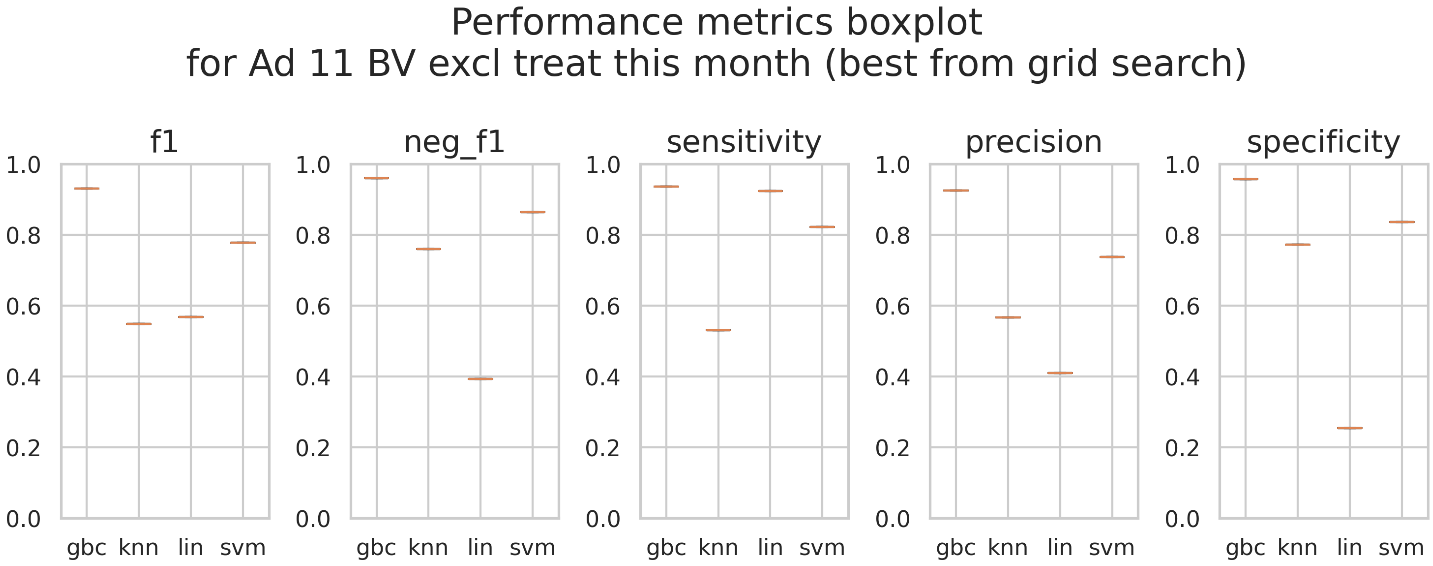


Figure 4.8: The performance metrics for the best performing models on the task of predicting BV in the following month where the surrounding 11 days of a SBV diagnosis are labelled as BV, including days where treatment takes place.

Learning Rate	Max Depth	# estimators	subsample
0.03	8	500	0.9

Table 4.6: The parameters with which the gbc performed best of all models on the task of predicting BV in the following month.

Feature Importances

Since the best model was a gbc, the feature importances were extracted and the 5 most important ones their values are visualized in Figure 4.9.

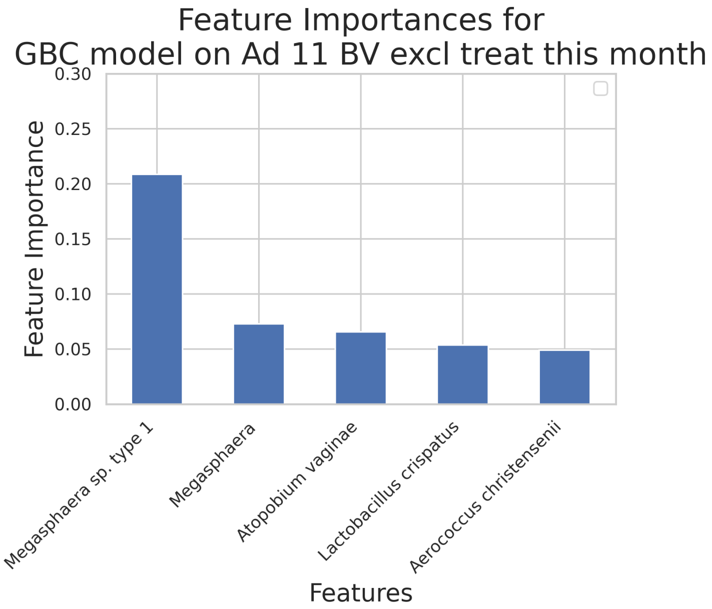


Figure 4.9: The feature importances for the best performing gbc model and label combination which scored as shown in Figure 4.8 on the BV prediction task for the following month.

The remaining feature importances from the gbc models trained on the other label types are depicted in Appendix A, Figure A.9.

4.3.3 Classification next month

Baselines

The same baseline is used as for the previous classification task on the following month, results of which are therefore depicted in the previous subsection in Table 4.5.

Model Performance

The best performing model and label combination on this task is once again the gbc model on the label for 7 adjacent days of BV excluding the treatment days, which again goes for all the considered metrics. The results, as well as that of the other models on the same task, are shown in Figure 4.10.

Again the gbc model performs higher than the baseline in the F1-score, while also maintaining a higher negative F1-score. The parameters found in the grid search that resulted in this performance are shown in Table 4.7.

Other label performance results are shown in Figures A.7 and A.10 of Appendix A for the baselines and the machine learning models, respectively.

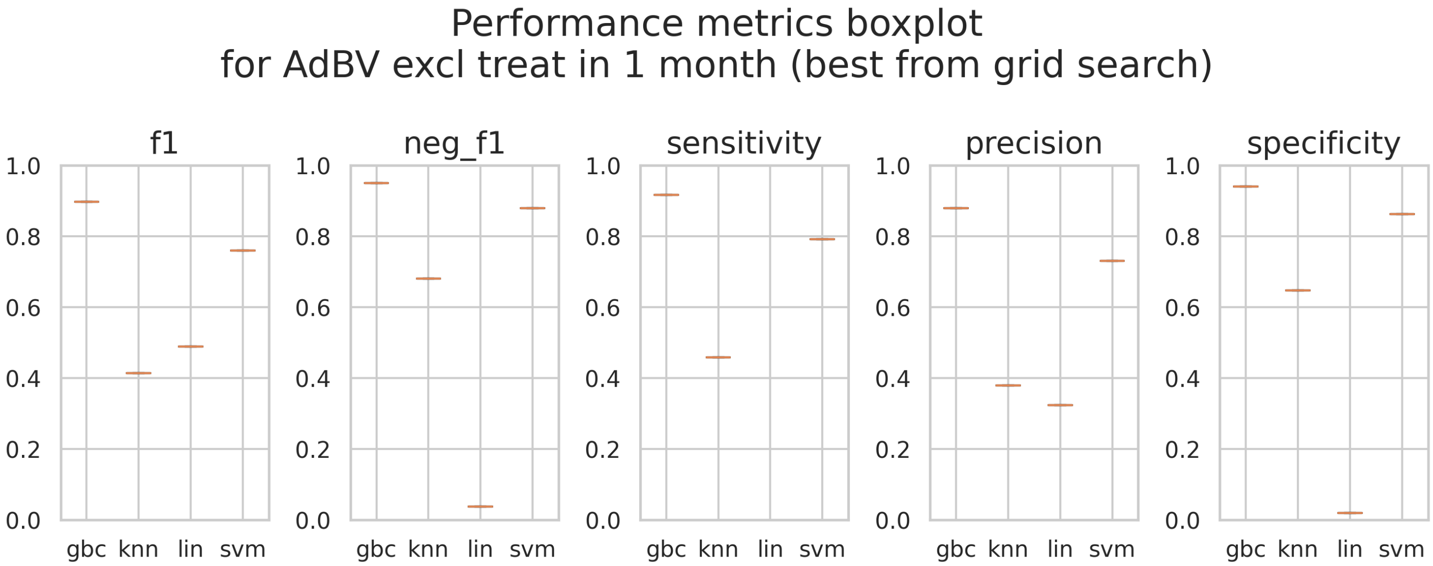


Figure 4.10: The performance metrics for the best performing models on the task of predicting BV where the surrounding 7 days of a SBV diagnosis are labelled as BV, excluding days where treatment takes place.

Table 4.7: The parameters with which the gbc performed best of all models on the task of predicting BV in the next month.

Learning Rate	**Max Depth**	**# estimators**	**subsample**
0.02	6	1000	0.9

Feature Importances

Since the best model was a gbc, the feature importances were extracted and the 5 most important ones their values are visualized in Figure 4.11.

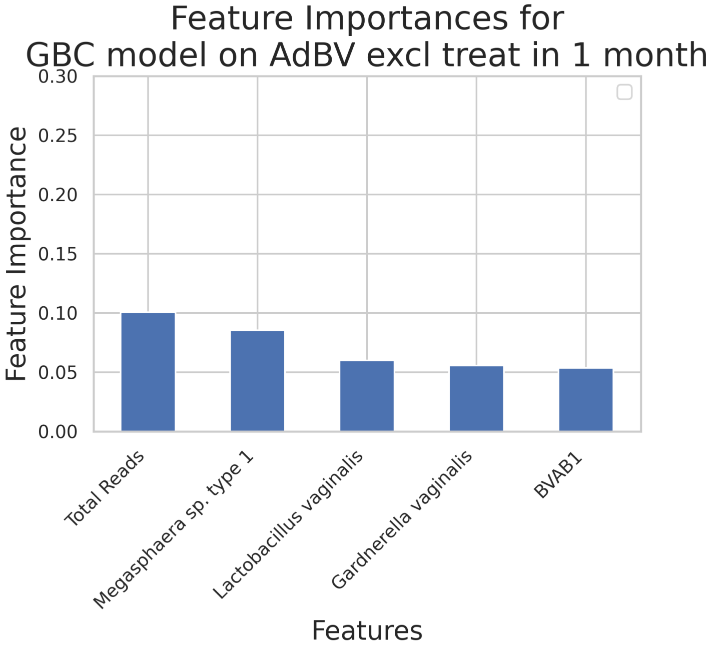


Figure 4.11: The feature importances for the best performing gbc model and label combination which scored as shown in Figure 4.10 on the BV prediction task for the next month.

The remaining feature importances from the gbc models trained on the other label types are depicted in Appendix A, Figure A.11.

4.3.4 Regression Task: days until BV onsets

The regression task was trained using one of the three regression performance metrics (MAE, MSE or R^2) as the one that should be optimized for. Only the best one will be reported on in this case. No baseline is included in this experiment since the same guesser could not be used for the continuous integer values that are predicted in this task.

Model Performance

The best performance on all of the regression metrics is achieved by the gbr model on 7 adjacent days of SBV excluding the treatment days. The results, among those of other models on the same task, are shown in Figure 4.12. This best performance was found while optimizing for the R^2 metric.

The gbr model outperforms all others. The parameters found in the grid search that resulted in this performance are shown in Table 4.8.

Other label performance results are shown in Figures A.14, A.16 and A.12 of Appendix A for the machine learning models. It should be noted that the remaining labels were predicted nearly as well as this one by the gbr model.

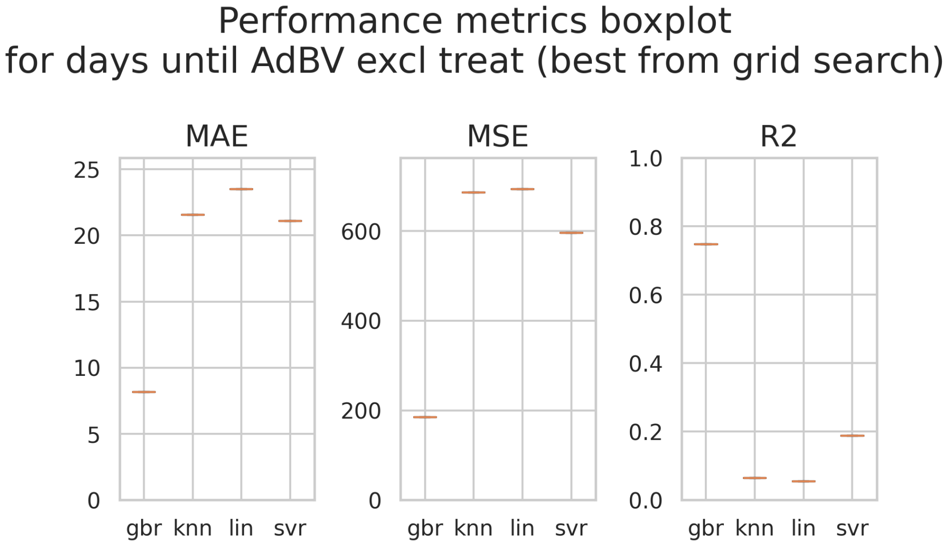


Figure 4.12: The performance metrics for the best performing models on the task of predicting the amount of days until BV onsets where the surrounding 7 days of a SBV diagnosis are labelled as BV, excluding days where treatment takes place. This performance was achieved by optimizing for the R^2 metric.

Table 4.8: The parameters with which the gbr performed best of all models on the regression task of predicting the number of days until BV onsets.

Learning Rate	**Max Depth**	**# estimators**	**subsample**
0.02	8	1000	0.5

Feature Importances

As the best performing model was a gbr, the feature importances were extracted and the 5 most important ones and their values are visualized in Figure 4.13.

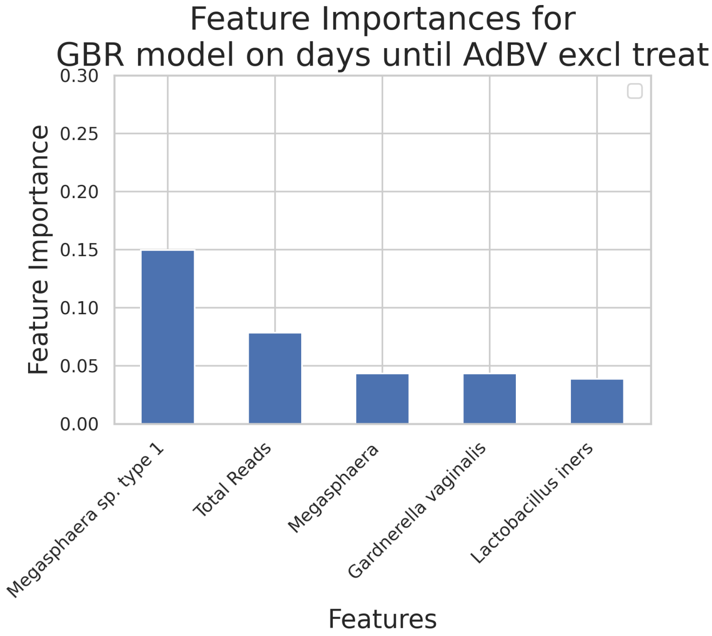


Figure 4.13: The feature importances for the best performing gbr model and label combination which scored as shown in Figure 4.12 on the task of predicting the number of days until BV onsets.

The remaining feature importances from the gbr models trained on the other label types are depicted in Appendix A, Figures A.15, A.17 and A.13.

Chapter 5

Discussion

In order to find to what extend vaginal microbiota taxa abundance data has predictive qualities for BV, ML models were trained on various classification and regression tasks. These tasks included BV diagnosis, prediction in the following 10 days, 1 month or the next month, as well as the number of days until BV onsets. The longitudinal dataset that was used only held labels for the moment of diagnosis. Therefore, various labels were created to represent the likely presence of BV. In order to confirm that the models were learning how to classify a datapoint as BV positive or healthy, feature importances were extracted. An indicator that this is successful could be that a model performs well on a task and gives a high importance to features that align with the relevant bacteria indicated in the literature.

The hypothesis resulting from the t-SNE visualizations, that the datapoints would not be easily separated, at least not linearly, appears proven. Since the linear svm performed worse than the baseline model in most cases. The gradient boosting trees based model outperforming in every task at hand was therefore to be expected, as they have been proven among the best algorithms in machine learning (T. Chen and Guestrin 2016). Since the amount of data was limited, the choice was made not to make use of deep learning techniques. However, this could be an interesting follow-up to this thesis. More so if the amount of data available increases significantly.

Even though the number of available datapoints was just 1748, the best performance of the ML models on the test set on the diagnosis task was an F1-score of 0.62, with a negative F1-score of 0.97. The most important features for this model align with findings in the literature (X. Chen et al. 2021). As expected, Lactobacillus iners is important in the diagnosis classification task, as this is associated with a lower risk of BV short term. Bifidobacteriaceae was also found among the most important features. The perianal presence of at least one type of this bacterium is associated with BV (Swidsinski et al. 2010), and therefore its presence can be expected. As was hypothesized, the SBV label in the diagnosis task was particularly hard to predict compared to the newly created labels. With the svm being the sole exception that outperformed the baseline model.

In the short term future of a prediction 10 days ahead, one could expect similar important features as with the diagnosis. All but one of the best model features do indeed overlap with the diagnosis best model. Prevotella disiens, associated with a healthy vagina, was found instead of Bifidobacteriaceae.

For longer term, Lactobacillus crispatus was expected to be more important. This was indeed found in the prediction of BV in the following month. Also noteworthy, is that the same BVAB1 species was found to be important in following month prediction, which is in line with the findings by X. Chen et al. (2021). Atopium vaginae was also found to be an important feature, aligning with findings that the co-presence of this bacterium and Gardnerella vaginalis plays a role in BV onset. Another finding that was also done in other research is the importance of Aerococcus christensenii presence for BV after one month. More precisely, this is associated with recurrent BV after 1 month (Xiao et al. 2019).

The same hypothesis of the L. crispatus importance for longer term prediction does not hold for the next month prediction in the best performing model. Lactobacillus vaginalis was found to be important for this prediction task, while other important features overlapped with the other findings. This Lactobacillus genus accompanies L. crispatus dominated microbiota in 79 percent of cases (Jespers et al. 2017). Therefore, this encounter could also be expected in longer term BV prediction, and might explain the absence of L. crispatus in the feature importance top 5.

In the regression task, the substantial performance of 0.75 for the R^2 metric was reached. No new feature importances were found in this experiment, as they overlap with the findings in the future BV prediction experiments.

As a follow-up to this thesis to further confirm the findings of the feature importances, the same features could be used in another dataset to perform similar predictions. When the non-important features are omitted, and the same performance can be reached, this would further confirm that the models used these features for their learning.

A more general observation is that the tasks with more positive labels are performed better by the models, even if there are fewer data points in total for these tasks, such as BV prediction in this month or the next month. Also noteworthy, is that the best performing newly created label type was different for each task.

Another approach that could result in more informative data, rather than more abundant data, is to switch from genetic data, which expresses functional potential, to transcriptomics data (France et al. 2022).

In future work, new approaches could be searched for to effectively predict BV in a materially cheap way, like using questionnaire data (Noyes et al. 2018). This could provide women with a low effort way to see if they are likely to have BV.

Another topic to research could be to predict the recurrence of BV after treatment. This information could be used to compare new treatments, so that current plans of treatment can be improved upon with a clear window of reference. Some data to research this topic was provided by Deng et al. (2018).

In conclusion, depending on the task at hand, models will rely more on certain bacterial features than others. This is in line with findings from previous work. For the field of BV research, a clear consensus should be found on which BV type is most relevant to research to ideally help women prevent BV, or otherwise cure it with lower risk of recurrence. Once this consensus is reached, large scale research, like the one resulting in the dataset used in this thesis, should provide the field with more information to work with. This could help prevent cases of damaging STI's, risks during pregnancy and general inconvenience in the lives of millions of affected women. Making sure the BV label is provided on every day a patient has it, for example, could already get rid of a lot of uncertainty in similar future work. In the expansion of the available data, it is critical to find a diverse group to participate in the study, to ensure that any resulting tools are able to generalize well across the whole population.

Chapter 6

References

Allsworth, Jenifer E. and Jeffrey F. Peipert (Aug. 2011). "Severity of bacterial vaginosis and the risk of sexually transmitted infection". en. In: *American Journal of Obstetrics and Gynecology* 205.2, 113.e1–113.e6. ISSN: 00029378. DOI: 10.1016/j.ajog.2011.02.060. URL: https://linkinghub.elsevier.com/retrieve/pii/S000293781100250X.

Amsel, Richard et al. (1983). "Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations". In: *The American journal of medicine* 74.1, pp. 14–22.

Beck, Daniel and James A. Foster (Feb. 2014). "Machine Learning Techniques Accurately Classify Microbial Communities by Bacterial Vaginosis Characteristics". In: *PLoS ONE* 9.2. Ed. by Bryan A. White, e87830. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0087830. URL: https://dx.plos.org/10.1371/journal.pone.0087830.

Burmølle, Mette et al. (2006). "Enhanced biofilm formation and increased resistance to antimicrobial agents and bacterial invasion are caused by synergistic interactions in multispecies biofilms". In: *Applied and environmental microbiology* 72.6, pp. 3916–3923.

Chen, Chen et al. (Dec. 2017). "The microbiota continuum along the female reproductive tract and its relation to uterine-related diseases". en. In: *Nature Communications* 8.1, p. 875. ISSN: 2041-1723. DOI: 10.1038/s41467-017-00901-0. URL: http://www.nature.com/articles/s41467-017-00901-0.

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

Chen, Xiaodi et al. (Apr. 2021). "The Female Vaginal Microbiome in Health and Bacterial Vaginosis". In: *Frontiers in Cellular and Infection Microbiology* 11, p. 631972. ISSN: 2235-2988. DOI: 10.3389/fcimb.2021.631972. URL: https://www.frontiersin.org/articles/10.3389/fcimb.2021.631972/full.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

Costello, Elizabeth K et al. (2009). "Bacterial community variation in human body habitats across space and time". In: *science* 326.5960, pp. 1694–1697.

Cover, T. and P. Hart (1967). "Nearest neighbor pattern classification". In: *IEEE Transactions on Information Theory* 13.1, pp. 21–27. DOI: 10.1109/TIT.1967.1053964.

Deng, Zhi-Luo et al. (2018). "Metatranscriptome Analysis of the Vaginal Microbiota Reveals Potential Mechanisms for Protection against Metronidazole in Bacterial Vaginosis".

In: *mSphere* 3.3, e00262–18. DOI: 10.1128/mSphereDirect.00262-18. eprint: https://journals.asm.org/doi/pdf/10.1128/mSphereDirect.00262-18. URL: https://journals.asm.org/doi/abs/10.1128/mSphereDirect.00262-18.

Dobell, Clifford (1920). "The discovery of the intestinal protozoa of man." In: *Proceedings of the Royal Society of Medicine* 13.Sect_Hist_Med, pp. 1–15.

Dobzhansky, Theodosius and Theodosius Grigorievich Dobzhansky (1971). *Genetics of the evolutionary process*. Vol. 139. Columbia University Press.

Donders, Gilbert GG, Jana Zodzika, and Dace Rezeberga (2014). "Treatment of bacterial vaginosis: what we have and what we miss". In: *Expert Opinion on Pharmacotherapy* 15.5, pp. 645–657. DOI: 10.1517/14656566.2014.881800. eprint: https://doi.org/10.1517/14656566.2014.881800. URL: https://doi.org/10.1517/14656566.2014.881800.

Elgantri, R, Alhadi Mohamed, and Fatma Ibrahim (2010). "Diagnosis of Bacterial Vaginosis by Amsel Criteria and Gram Stain Method". In: *Sebha MedicalJournal* 9.1, pp. 20–27.

Eschenbach, David A et al. (1988). "Diagnosis and clinical manifestations of bacterial vaginosis". In: *American journal of obstetrics and gynecology* 158.4, pp. 819–828.

Fix, Evelyn and Joseph Lawson Hodges (1989). "Discriminatory analysis. Nonparametric discrimination: Consistency properties". In: *International Statistical Review/Revue Internationale de Statistique* 57.3, pp. 238–247.

France, Michael T. et al. (2022). "Insight into the ecology of vaginal bacteria through integrative analyses of metagenomic and metatranscriptomic data". In: *Genome Biology* 23.1. DOI: 10.1186/s13059-022-02635-9.

Goldenberg, Robert L. et al. (1996). "Bacterial colonization of the vagina during pregnancy in four ethnic groups". In: *American Journal of Obstetrics and Gynecology* 174.5, pp. 1618–1621. ISSN: 0002-9378. DOI: https://doi.org/10.1016/S0002-9378(96)70617-8. URL: https://www.sciencedirect.com/science/article/pii/S0002937896706178.

Hay, P. E. et al. (1994). "A longitudinal study of bacterial vaginosis during pregnancy". In: *BJOG: An International Journal of Obstetrics & Gynaecology* 101.12, pp. 1048–1053. DOI: https://doi.org/10.1111/j.1471-0528.1994.tb13580.x. eprint: https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1471-0528.1994.tb13580.x. URL: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1111/j.1471-0528.1994.tb13580.x.

HL, EAAR (2006). "National guidelines on prevention, management and control of reproductive tract infections including sexually transmitted infections". In.

Jespers, Vicky et al. (2017). "A longitudinal analysis of the vaginal microbiota and vaginal immune mediators in women from sub-Saharan Africa". In: *Scientific Reports* 7.1, pp. 1–13.

Linhares, Iara M. et al. (2011). "Contemporary perspectives on vaginal pH and lactobacilli". In: *American Journal of Obstetrics and Gynecology* 204.2, 120.e1–120.e5. ISSN: 0002-9378. DOI: https://doi.org/10.1016/j.ajog.2010.07.010. URL: https://www.sciencedirect.com/science/article/pii/S0002937810008768.

Luckey, T. D. (Dec. 1972). "Introduction to intestinal microecology". In: *The American Journal of Clinical Nutrition* 25.12, pp. 1292–1294. ISSN: 0002-9165. DOI: 10.1093/ajcn/25.12.1292. eprint: https://academic.oup.com/ajcn/article-pdf/25/12/1292/24187652/1292.pdf. URL: https://doi.org/10.1093/ajcn/25.12.1292.

Ma, Zhanshan S. and Aaron M. Ellison (2021). "In silico trio biomarkers for bacterial vaginosis revealed by species dominance network analysis". In: *Computational and Structural Biotechnology Journal* 19, pp. 2979–2989. ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2021.05.020. URL: https://www.sciencedirect.com/science/article/pii/S2001037021002014.

Maaten, Laurens van der and Geoffrey Hinton (Nov. 2008). "Viualizing data using t-SNE". In: *Journal of Machine Learning Research* 9, pp. 2579–2605.

Marchesi, Julian R. and Jacques Ravel (Dec. 2015). "The vocabulary of microbiome research: a proposal". en. In: *Microbiome* 3.1, 31, s40168-015-0094–5. ISSN: 2049-2618. DOI: 10.1186/s40168-015-0094-5. URL: https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-015-0094-5.

Martius, JOACHIM et al. (1988). "Relationships of vaginal Lactobacillus species, cervical Chlamydia trachomatis, and bacterial vaginosis to preterm birth." In: *Obstetrics and gynecology* 71.1, pp. 89–95.

Mehta, Supriya D. et al. (Aug. 2020). "The Microbiome Composition of a Man's Penis Predicts Incident Bacterial Vaginosis in His Female Sex Partner With High Accuracy". In: *Frontiers in Cellular and Infection Microbiology* 10, p. 433. ISSN: 2235-2988. DOI: 10.3389/fcimb.2020.00433. URL: https://www.frontiersin.org/article/10.3389/fcimb.2020.00433/full.

Mehta, Supriya Dinesh et al. (Dec. 2020). "Characteristics of Women and Their Male Sex Partners Predict Bacterial Vaginosis Among a Prospective Cohort of Kenyan Women With Nonoptimal Vaginal Microbiota". en. In: *Sexually Transmitted Diseases* 47.12, pp. 840–850. ISSN: 1537-4521, 0148-5717. DOI: 10.1097/OLQ.0000000000001259. URL: https://journals.lww.com/10.1097/OLQ.0000000000001259.

Al-Memar, M et al. (Jan. 2020). "The association between vaginal bacterial composition and miscarriage: a nested case–control study". In: *BJOG: An International Journal of Obstetrics & Gynaecology* 127.2, pp. 264–274. ISSN: 1470-0328, 1471-0528. DOI: 10.1111/1471-0528.15972. URL: https://onlinelibrary.wiley.com/doi/10.1111/1471-0528.15972.

Munoz, Alexander et al. (Dec. 2021). "Modeling the temporal dynamics of cervicovaginal microbiota identifies targets that may promote reproductive health". In: *Microbiome* 9.1, p. 163. ISSN: 2049-2618. DOI: 10.1186/s40168-021-01096-9. URL: https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-021-01096-9.

Muzny, Christina A. et al. (Feb. 2020). "Host–vaginal microbiota interactions in the pathogenesis of bacterial vaginosis". In: *Current Opinion in Infectious Diseases* 33.1, pp. 59–65. ISSN: 0951-7375. DOI: 10.1097/QCO.0000000000000620. URL: http://journals.lww.com/10.1097/QCO.0000000000000620.

Noyes, Noelle et al. (Jan. 2018). "Associations between sexual habits, menstrual hygiene practices, demographics and the vaginal microbiome as revealed by Bayesian network analysis". In: *PLOS ONE* 13.1, pp. 1–25. DOI: 10.1371/journal.pone.0191625. URL: https://doi.org/10.1371/journal.pone.0191625.

Nugent, Robert P, Marijane A Krohn, and Sharon L Hillier (1991). "Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation". In: *Journal of clinical microbiology* 29.2, pp. 297–301.

O'Hanlon, Deirdre E, Thomas R Moench, and Richard A Cone (Dec. 2011). "In vaginal fluid, bacteria associated with bacterial vaginosis can be suppressed with lactic acid but not hydrogen peroxide". en. In: *BMC Infectious Diseases* 11.1, p. 200. ISSN: 1471-2334. DOI: 10.1186/1471-2334-11-200. URL: https://bmcinfectdis.biomedcentral.com/articles/10.1186/1471-2334-11-200.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peebles, Kathryn et al. (May 2019). "High Global Burden and Costs of Bacterial Vaginosis: A Systematic Review and Meta-Analysis". en. In: *Sexually Transmitted Diseases* 46.5, pp. 304–311. ISSN: 1537-4521, 0148-5717. DOI: 10.1097/OLQ.0000000000000972. URL: https://journals.lww.com/00007435-201905000-00005.

Philip, Preethi S, Anoop I Benjamin, and Paramita Sengupta (2013). "Prevalence of symptoms suggestive of reproductive tract infections/sexually transmitted infections in women in an urban area of Ludhiana". In: *Indian journal of sexually transmitted diseases and AIDS* 34.2, p. 83.

Ravel, Jacques, Rebecca M Brotman, et al. (2013). "Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis". In: *Microbiome* 1.1, pp. 1–6.

Ravel, Jacques, Pawel Gajer, et al. (2010). "Vaginal microbiome of reproductive-age women". In: *Proceedings of the National Academy of Sciences* 108.supplement 1, pp. 4680–4687. DOI: 10.1073/pnas.1002611107.

Ravel, Jacques, Inmaculada Moreno, and Carlos Simón (Mar. 2021). "Bacterial vaginosis and its association with infertility, endometritis, and pelvic inflammatory disease". In: *American Journal of Obstetrics and Gynecology* 224.3, pp. 251–257. ISSN: 00029378. DOI: 10.1016/j.ajog.2020.10.019. URL: https://linkinghub.elsevier.com/retrieve/pii/S0002937820311935.

Savage, D. C. (1977). "MICROBIAL ECOLOGY OF THE GASTROINTESTINAL TRACT". In: *Annual Review of Microbiology* 31.1. PMID: 334036, pp. 107–133. DOI: 10.1146/annurev.mi.31.100177.000543. eprint: https://doi.org/10.1146/annurev.mi.31.100177.000543. URL: https://doi.org/10.1146/annurev.mi.31.100177.000543.

Sharma, Mahima et al. (Nov. 2021). "An Insight into Vaginal Microbiome Techniques". In: *Life* 11.11, p. 1229. ISSN: 2075-1729. DOI: 10.3390/life11111229. URL: https://www.mdpi.com/2075-1729/11/11/1229.

Spiegel, Carol A, R Amsel, and KK Holmes (1983). "Diagnosis of bacterial vaginosis by direct Gram stain of vaginal fluid". In: *Journal of clinical microbiology* 18.1, pp. 170–177.

Spiegel, Carol A. et al. (1980). "Anaerobic Bacteria in Nonspecific Vaginitis". In: *New England Journal of Medicine* 303.11. PMID: 6967562, pp. 601–607. DOI: 10.1056/NEJM198009113031102. eprint: https://doi.org/10.1056/NEJM198009113031102. URL: https://doi.org/10.1056/NEJM198009113031102.

Swidsinski, Alexander et al. (Oct. 2010). "Dissimilarity in the occurrence of Bifidobacteriaceae in vaginal and perianal microbiota in women with bacterial vaginosis". en. In: *Anaerobe* 16.5, pp. 478–482. ISSN: 10759964. DOI: 10.1016/j.anaerobe.2010.06.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S1075996410001046.

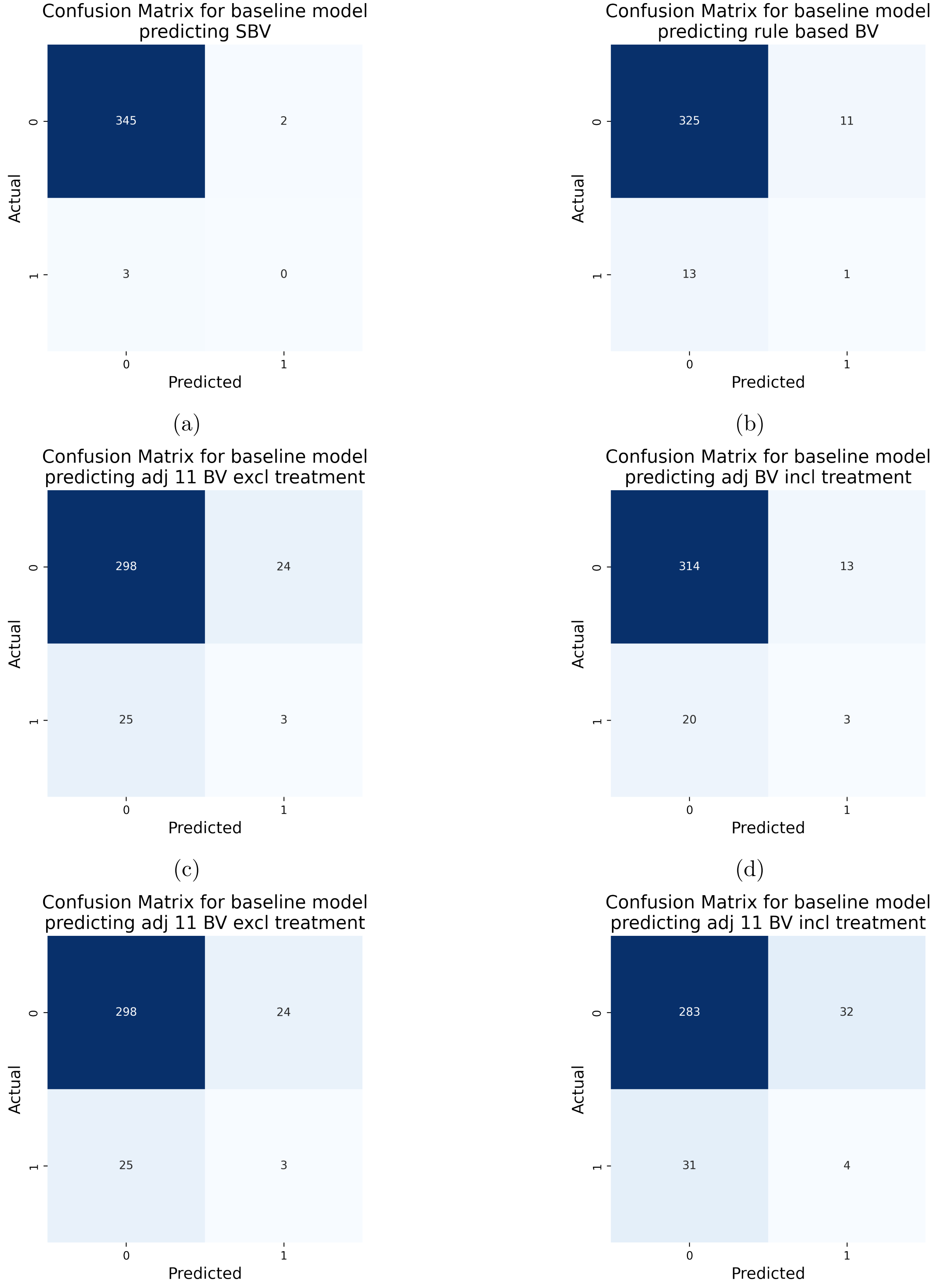
Twin, Jimmy et al. (Sept. 2013). "The Potential of Metatranscriptomics for Identifying Screening Targets for Bacterial Vaginosis". In: *PLoS ONE* 8.9. Ed. by Jacques Ravel, e76892. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0076892. URL: https://dx.plos.org/10.1371/journal.pone.0076892.

Xiao, Bingbing et al. (2019). “Association analysis on recurrence of bacterial vaginosis revealed microbes and clinical variables important for treatment outcome”. In: *Frontiers in cellular and infection microbiology* 9, p. 189.

Appendix A

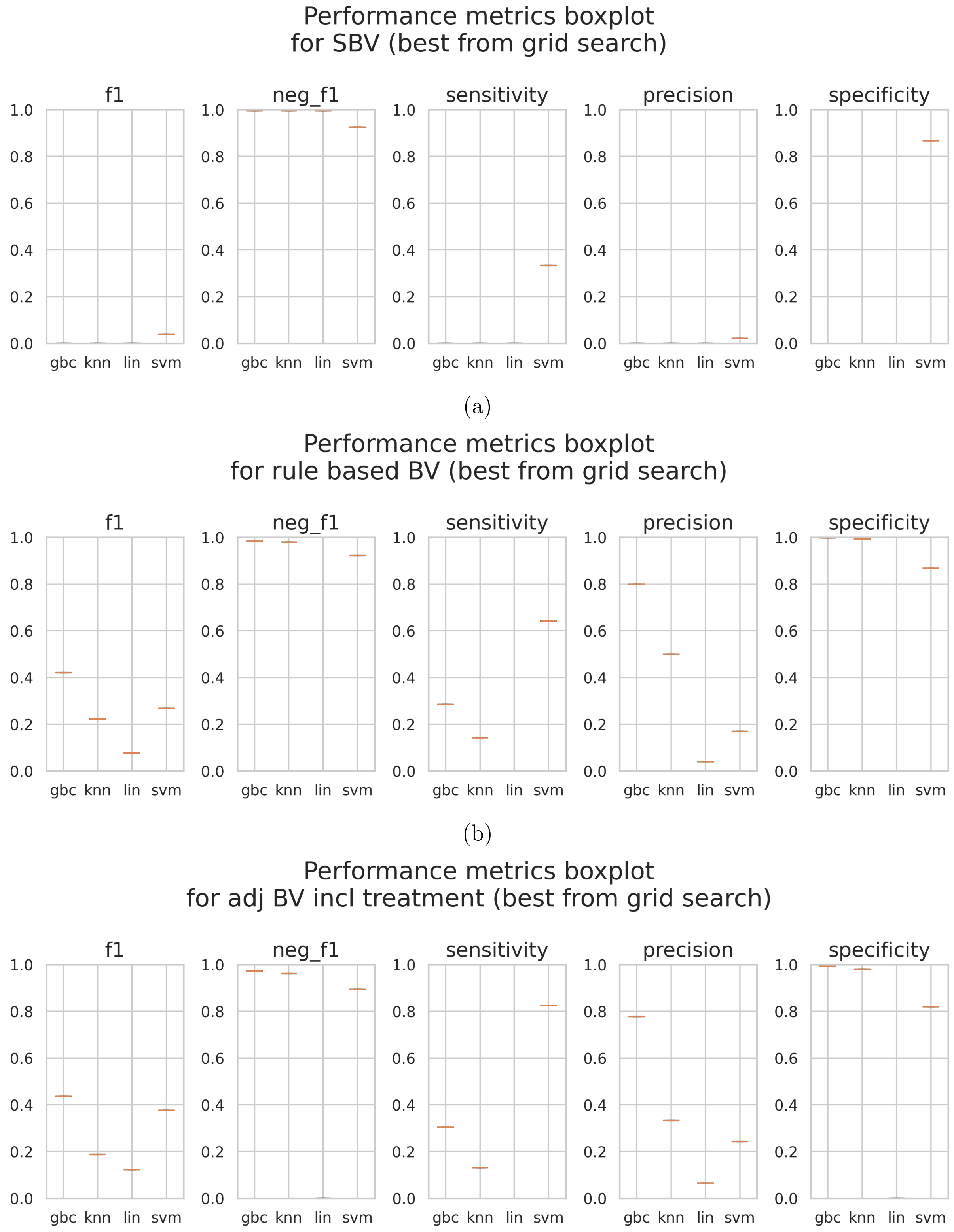
Supplemental Figures

A.1 Diagnosis



(a) (b) (c) (d) (e) (f)

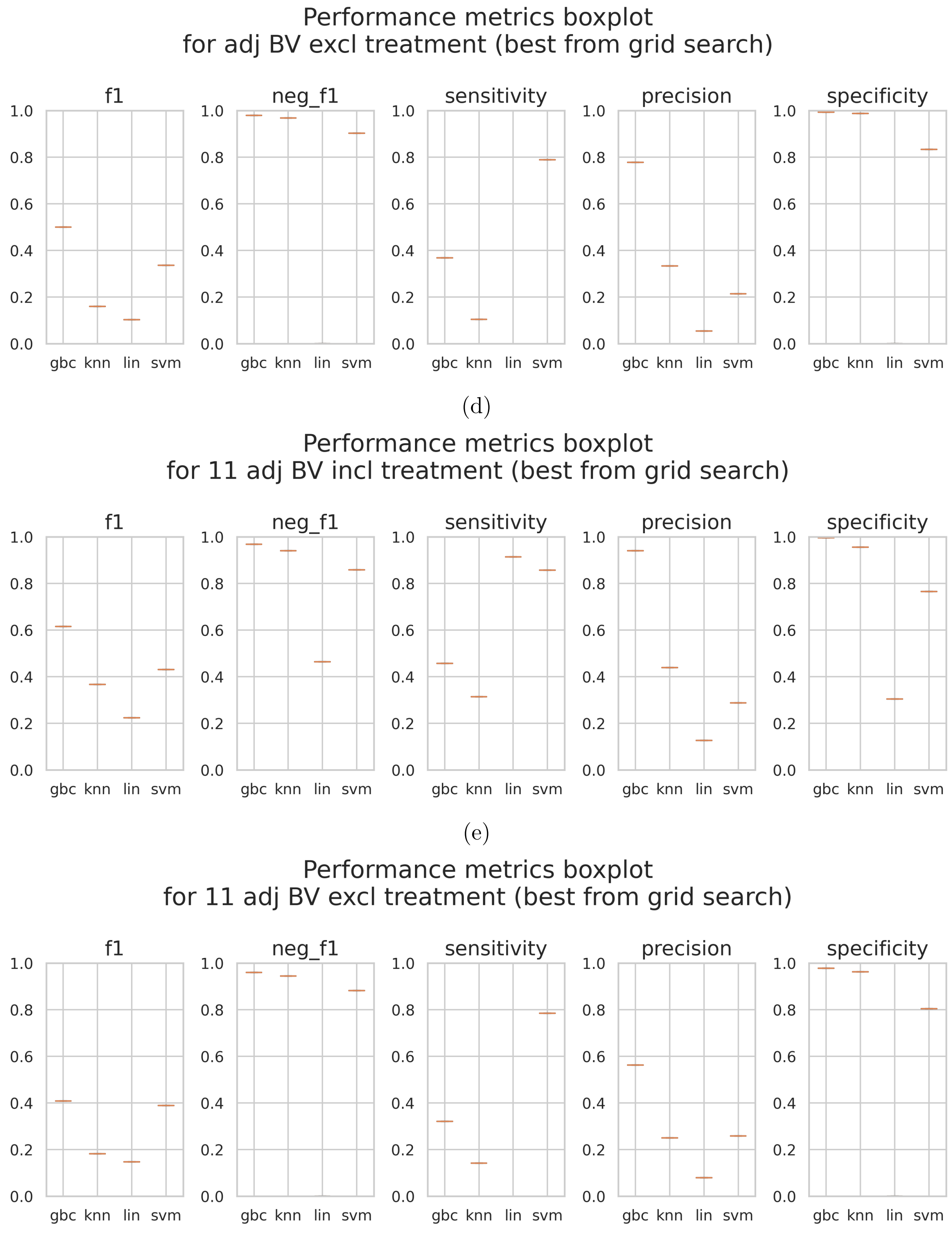
Figure A.1: The confusion matrices for the best found performance on the test data in a 5-fold repetition for the baseline CT to BV probability model on the diagnosis task of each BV label type.


Performance metrics boxplot
for SBV (best from grid search)
Performance metrics boxplot
for rule based BV (best from grid search)
Performance metrics boxplot
for adj BV incl treatment (best from grid search)
f1
neg_f1
sensitivity
precision
specificity
1.0
0.8
0.6
0.4
0.2
0.0
gbc
knn
lin
svm

(a)

(b)

(c)

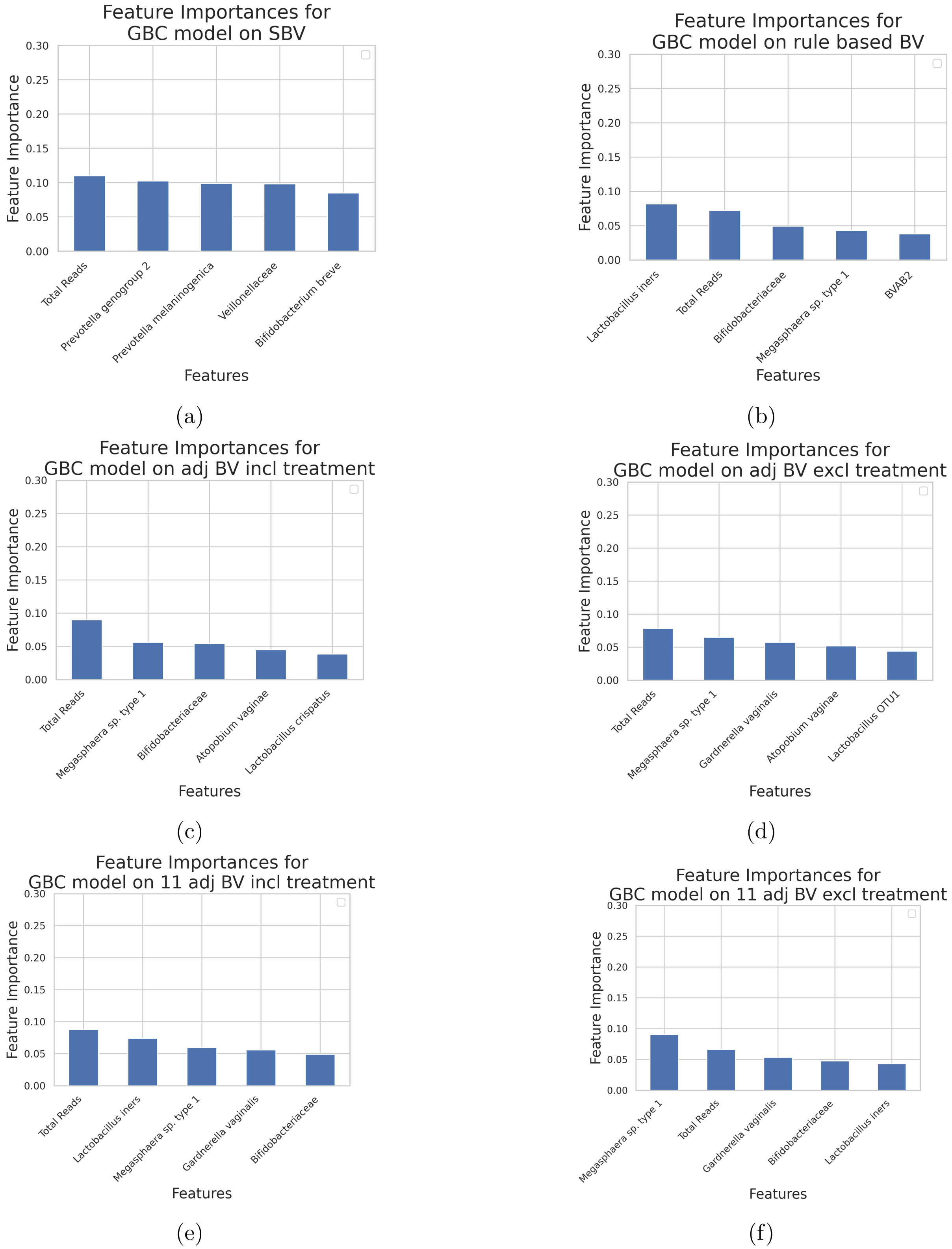


(d)

(e)

(f)

Figure A.2: The model performances for the best found ML models on the test data in a 5-fold repetition of the diagnosis task for each BV label type.



(a)

(b)

(c)

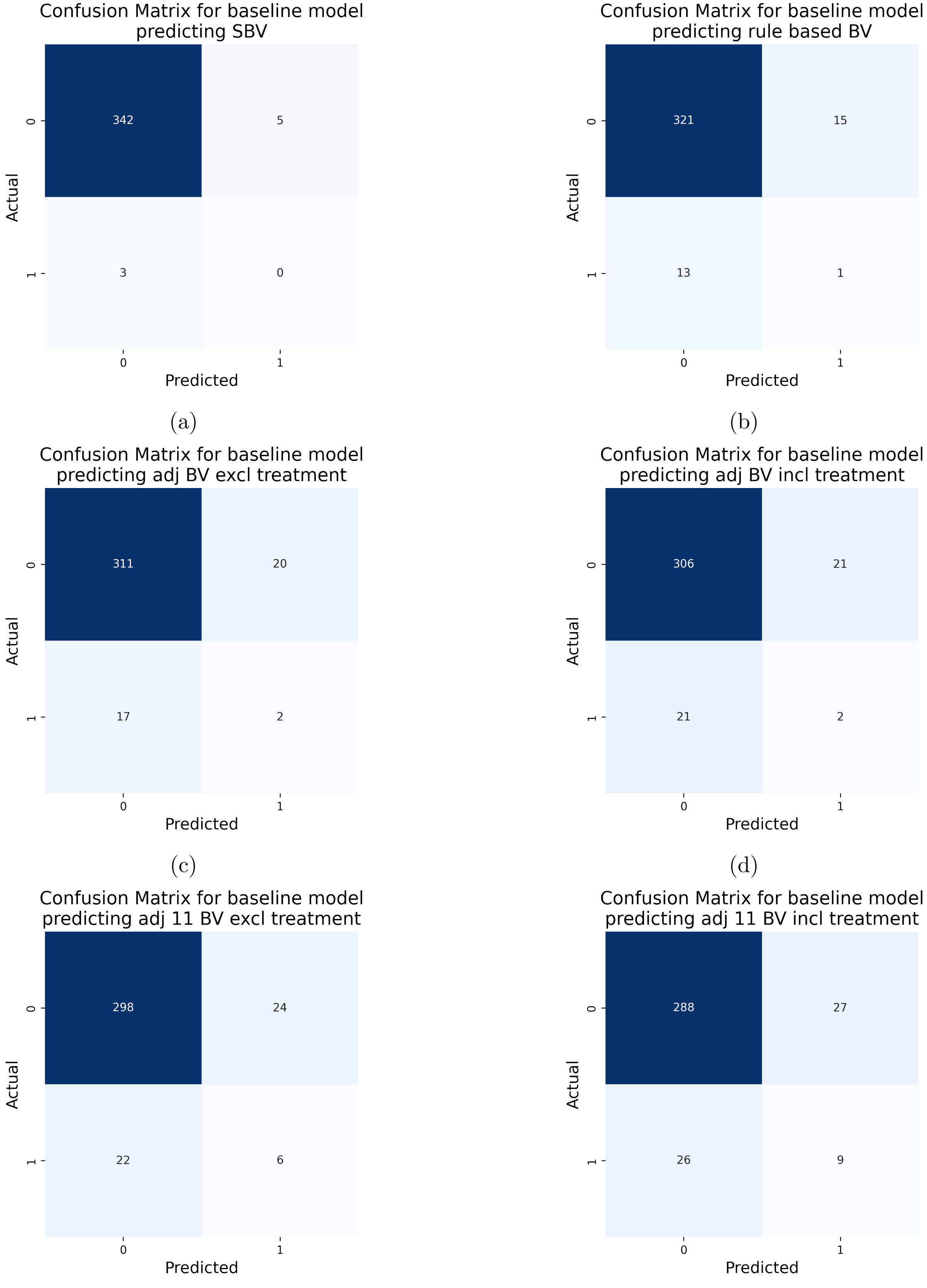
(d)

(e)

(f)

Figure A.3: The feature importances for the best found performances by the gbc model on the test data in a 5-fold repetition of the diagnosis task for each BV label type.

A.2 Prediction: 10 Days



Confusion Matrix for baseline model predicting SBV

(a)

Confusion Matrix for baseline model predicting rule based BV

(b)

Confusion Matrix for baseline model predicting adj BV excl treatment

(c)

Confusion Matrix for baseline model predicting adj BV incl treatment

(d)

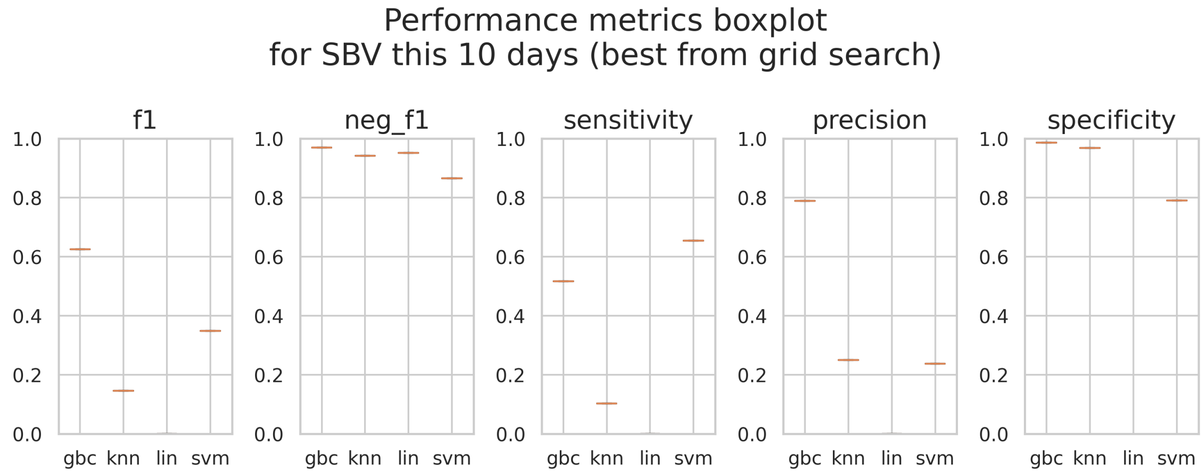
Confusion Matrix for baseline model predicting adj 11 BV excl treatment

(e)

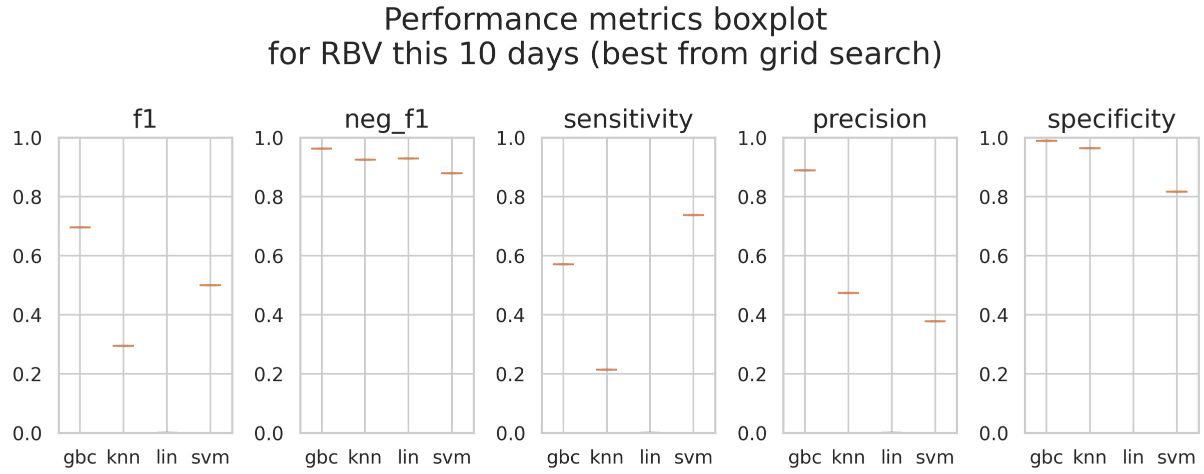
Confusion Matrix for baseline model predicting adj 11 BV incl treatment

(f)

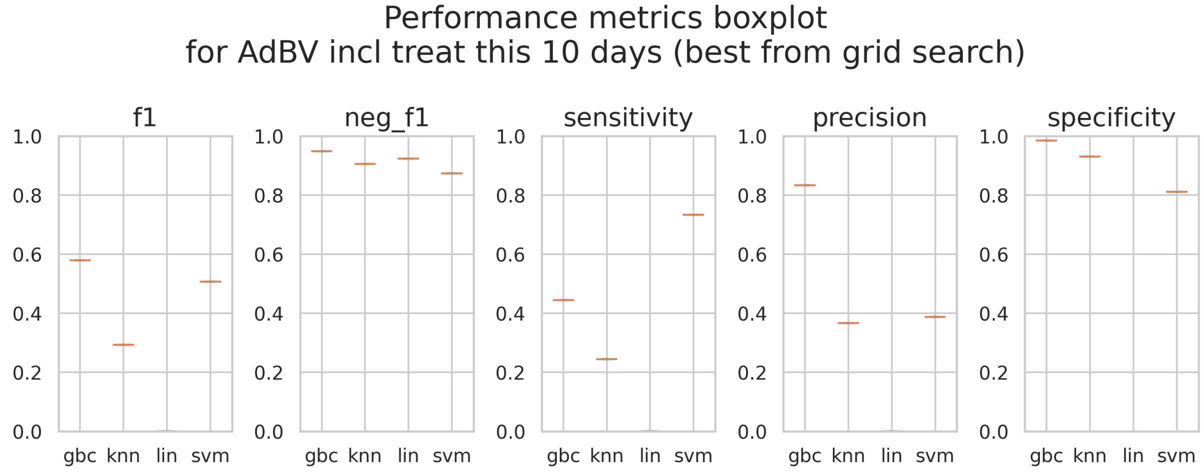
Figure A.4: The confusion matrices for the best found performance on the test data in a 5-fold repetition for the baseline CT to BV probability model on the 10 day ahead prediction task of each BV label type.


Performance metrics boxplot
for SBV this 10 days (best from grid search)
f1
neg_f1
sensitivity
precision
specificity
1.0
0.8
0.6
0.4
0.2
0.0
gbc
knn
lin
svm

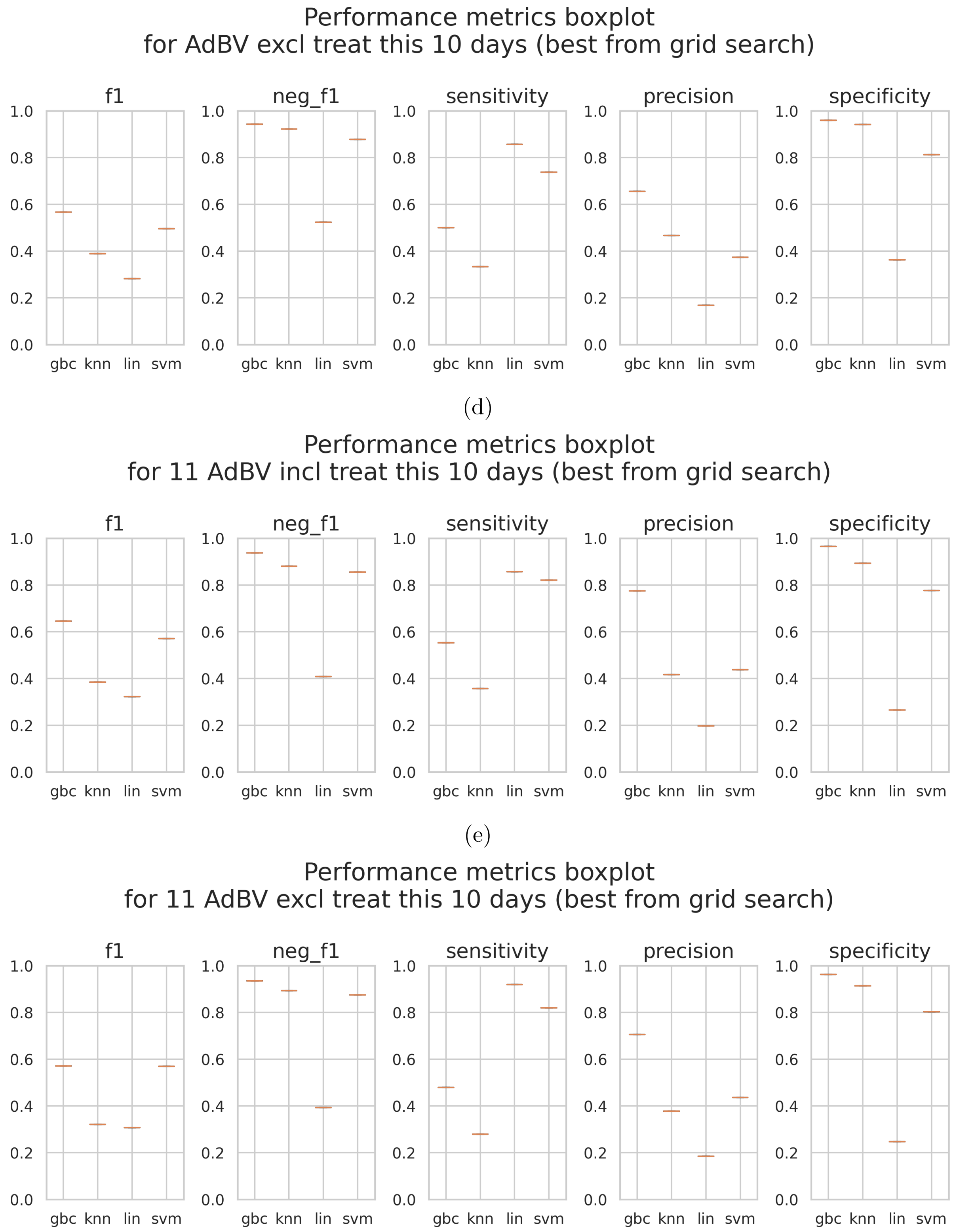
(a)


Performance metrics boxplot
for RBV this 10 days (best from grid search)
f1
neg_f1
sensitivity
precision
specificity
1.0
0.8
0.6
0.4
0.2
0.0
gbc
knn
lin
svm

(b)


Performance metrics boxplot
for AdBV incl treat this 10 days (best from grid search)
f1
neg_f1
sensitivity
precision
specificity
1.0
0.8
0.6
0.4
0.2
0.0
gbc
knn
lin
svm

(c)

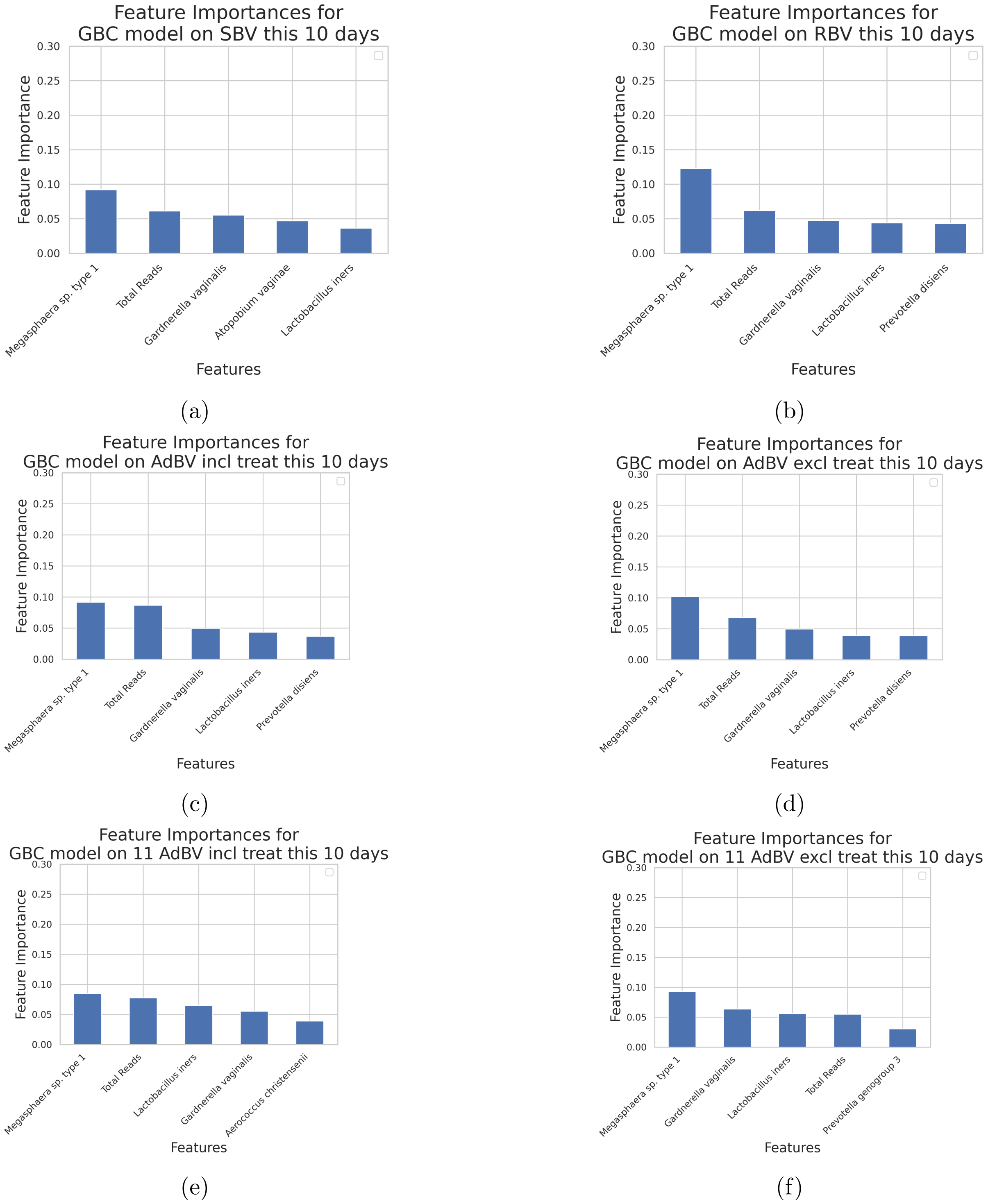


(d)

(e)

(f)

Figure A.5: The model performances for the best found ML models on the test data in a 5-fold repetition of the future 10 days task for each BV label type.



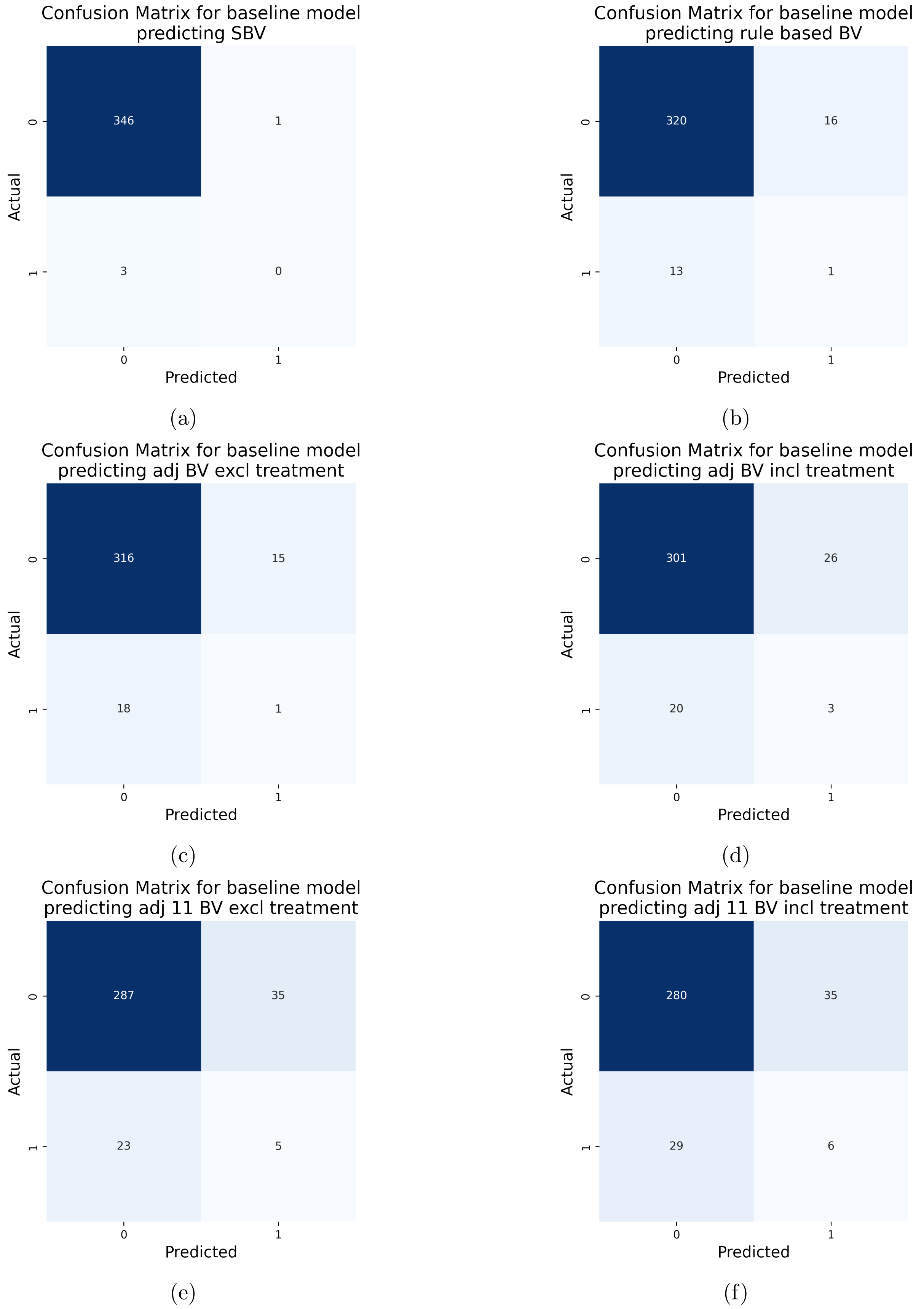
(a) (b)

(c) (d)

(e) (f)

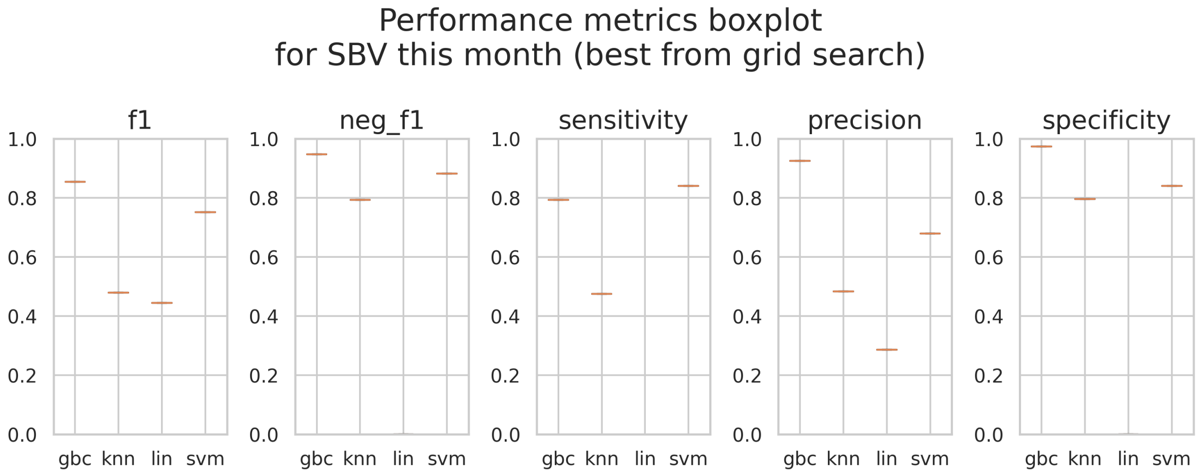
Figure A.6: The feature importances for the best found performances by the gbc model on the test data in a 5-fold repetition of the future 10 days task for each BV label type.

A.3 Prediction: This Month

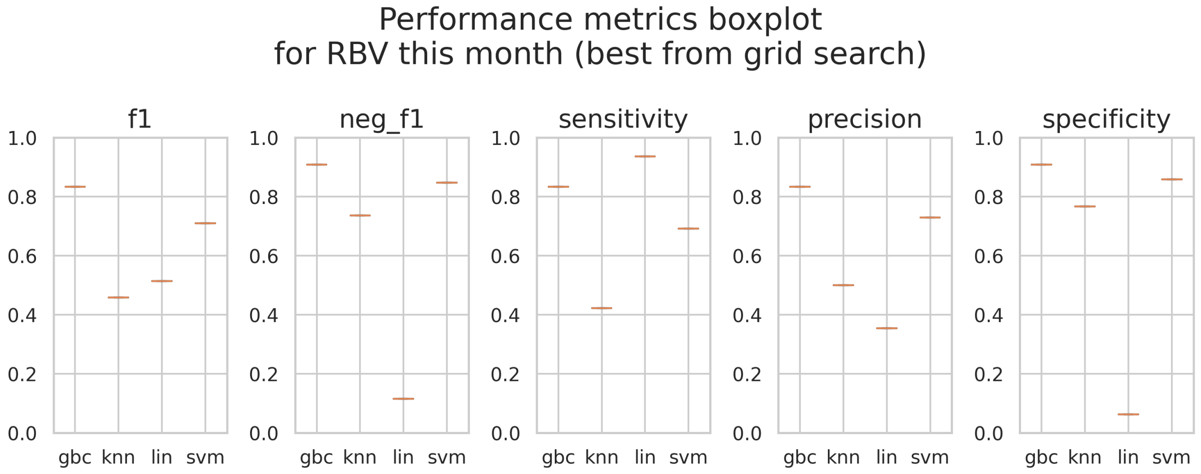


(a) (b) (c) (d) (e) (f)

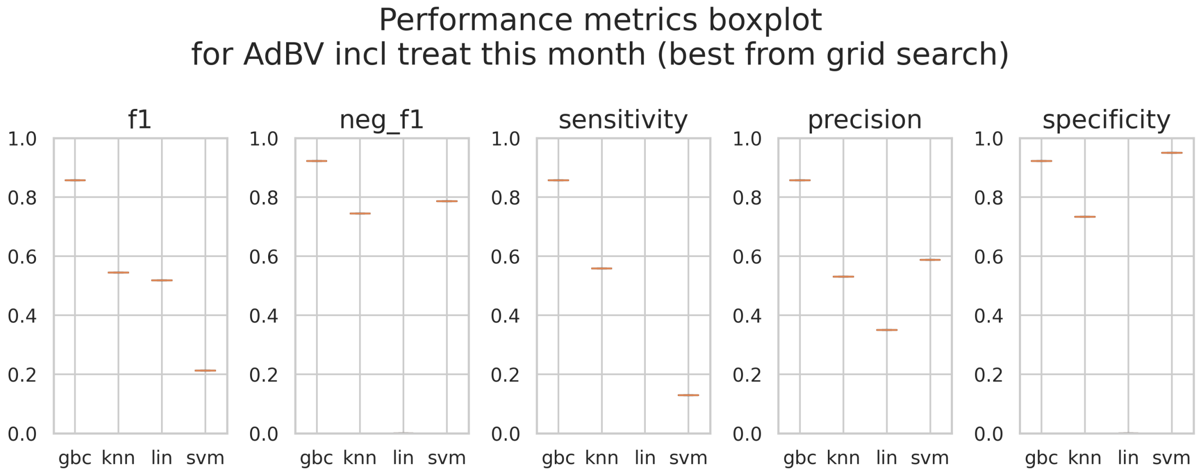
Figure A.7: The confusion matrices for the best found performance on the test data in a 5-fold repetition for the baseline CT to BV probability model on the 1 month ahead prediction task of each BV label type.


Performance metrics boxplot
for SBV this month (best from grid search)
f1
neg_f1
sensitivity
precision
specificity
1.0
0.8
0.6
0.4
0.2
0.0
gbc
knn
lin
svm

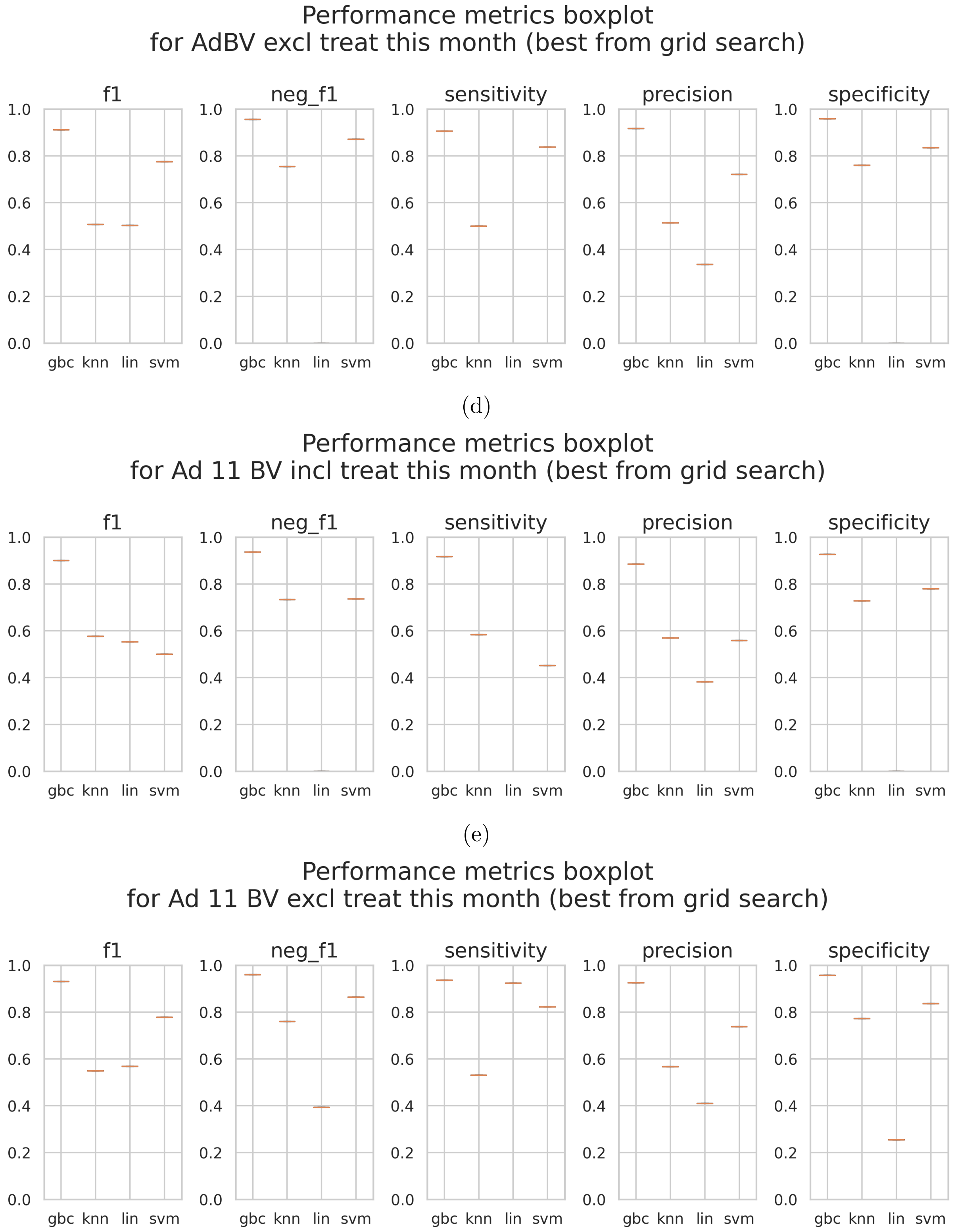
(a)


Performance metrics boxplot
for RBV this month (best from grid search)
f1
neg_f1
sensitivity
precision
specificity
1.0
0.8
0.6
0.4
0.2
0.0
gbc
knn
lin
svm

(b)


Performance metrics boxplot
for AdBV incl treat this month (best from grid search)
f1
neg_f1
sensitivity
precision
specificity
1.0
0.8
0.6
0.4
0.2
0.0
gbc
knn
lin
svm

(c)

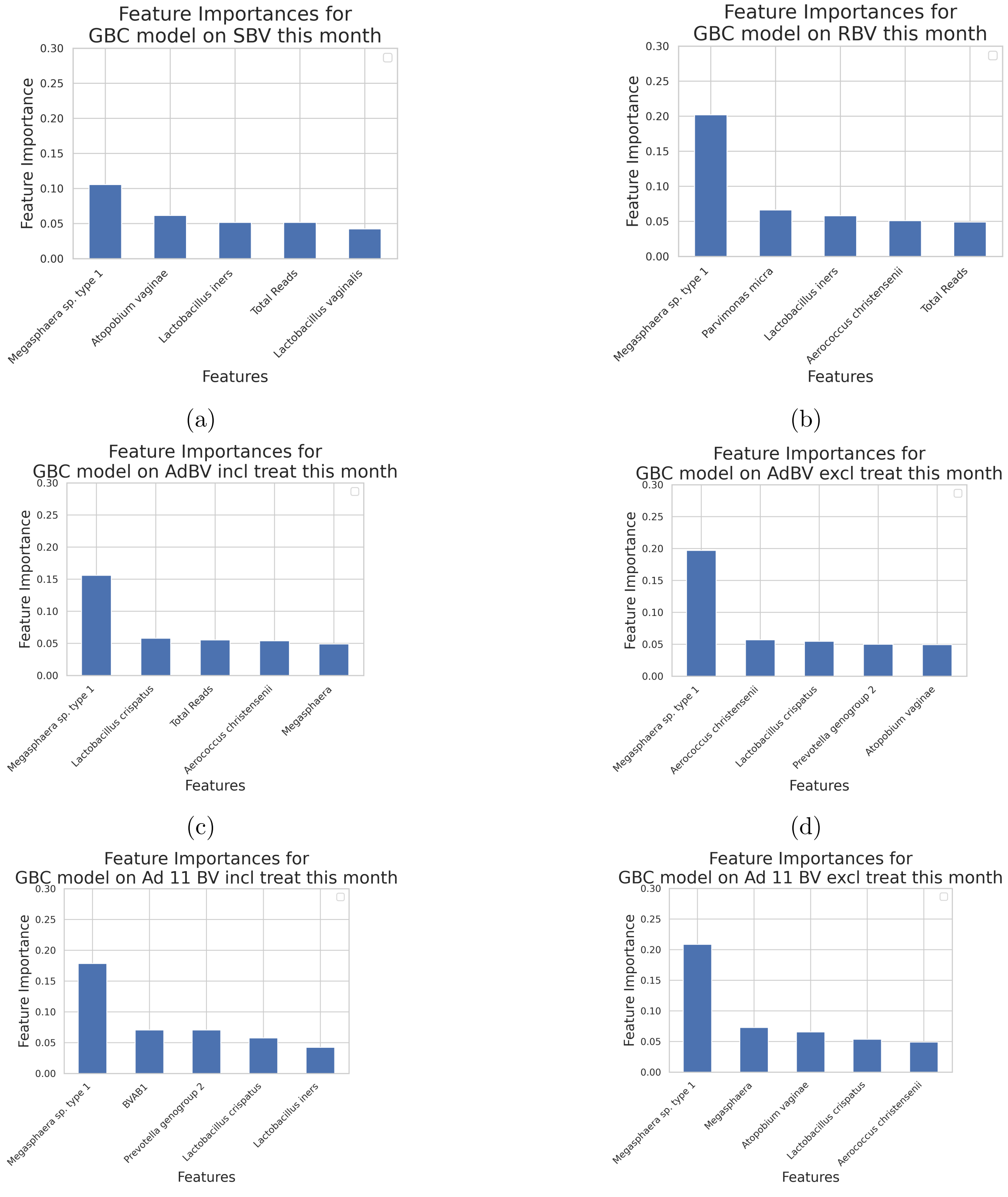


(d)

(e)

(f)

Figure A.8: The model performances for the best found ML models on the test data in a 5-fold repetition of this month prediction task for each BV label type.



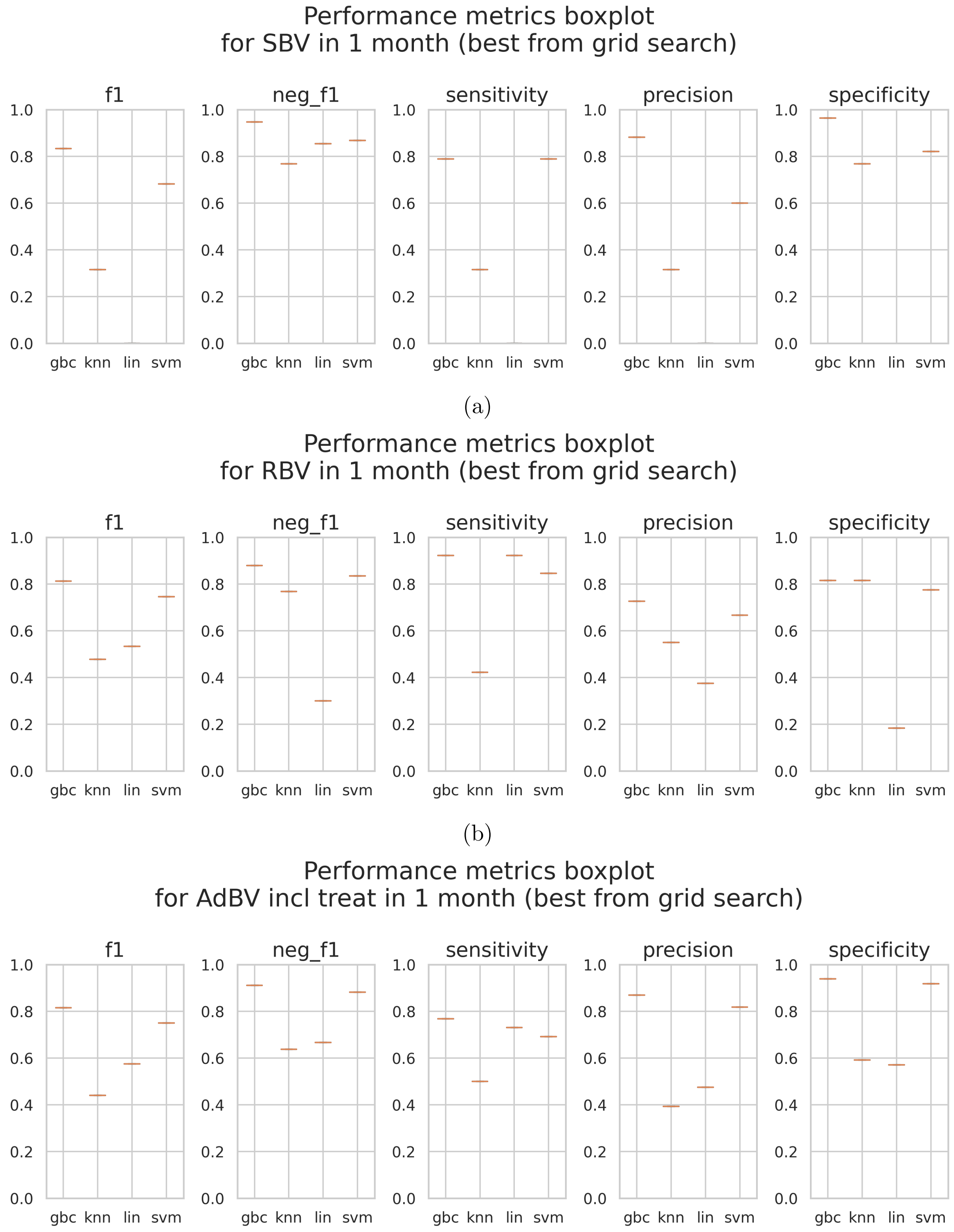
(a) (b)

(c) (d)

(e) (f)

Figure A.9: The feature importances for the best found performances by the gbc model on the test data in a 5-fold repetition of this month BV prediction task for each BV label type.

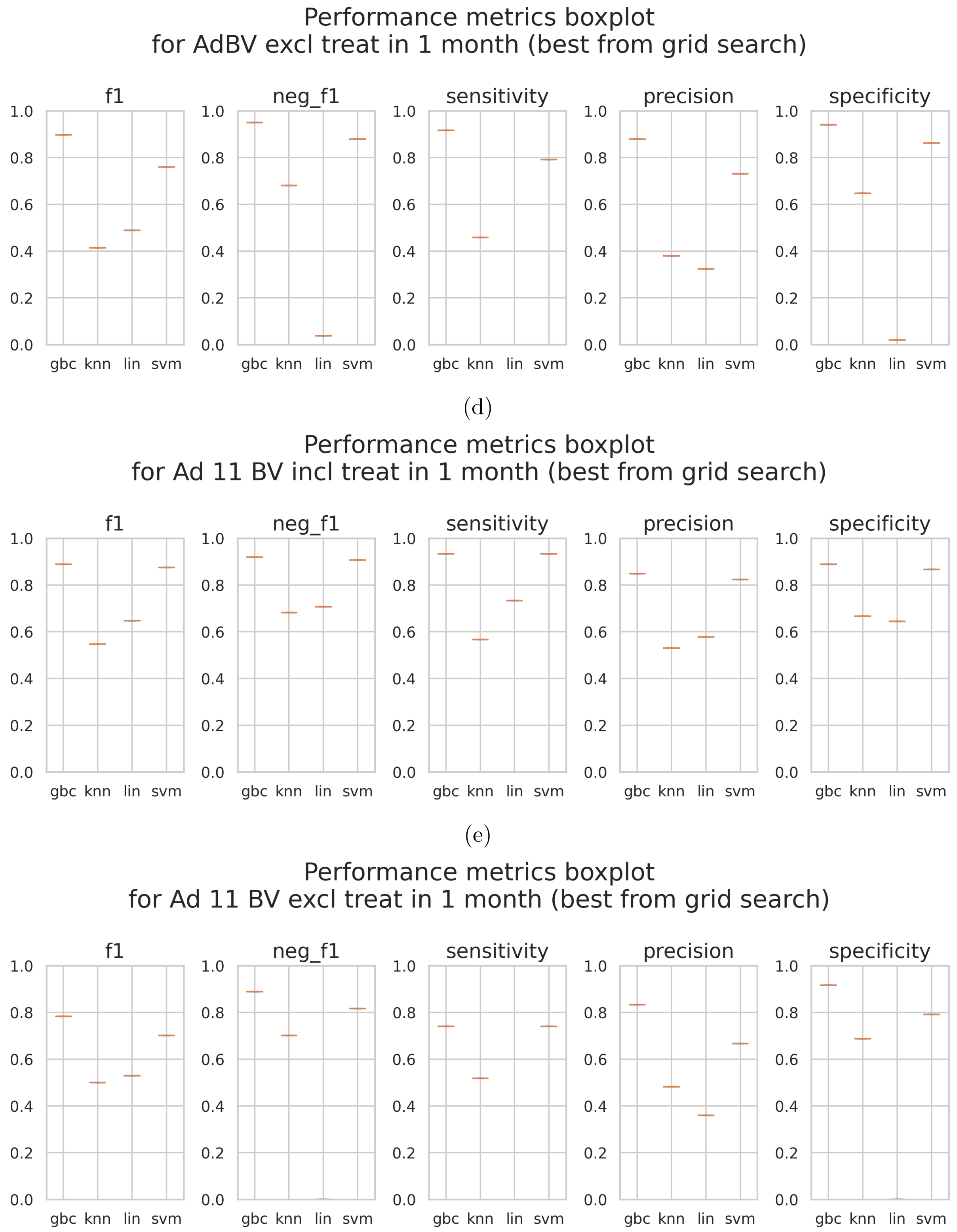
A.4 Prediction: Next Month


Performance metrics boxplot
for SBV in 1 month (best from grid search)
Performance metrics boxplot
for RBV in 1 month (best from grid search)
Performance metrics boxplot
for AdBV incl treat in 1 month (best from grid search)
f1
neg_f1
sensitivity
precision
specificity
0.0
0.2
0.4
0.6
0.8
1.0
gbc
knn
lin
svm

(a)

(b)

(c)

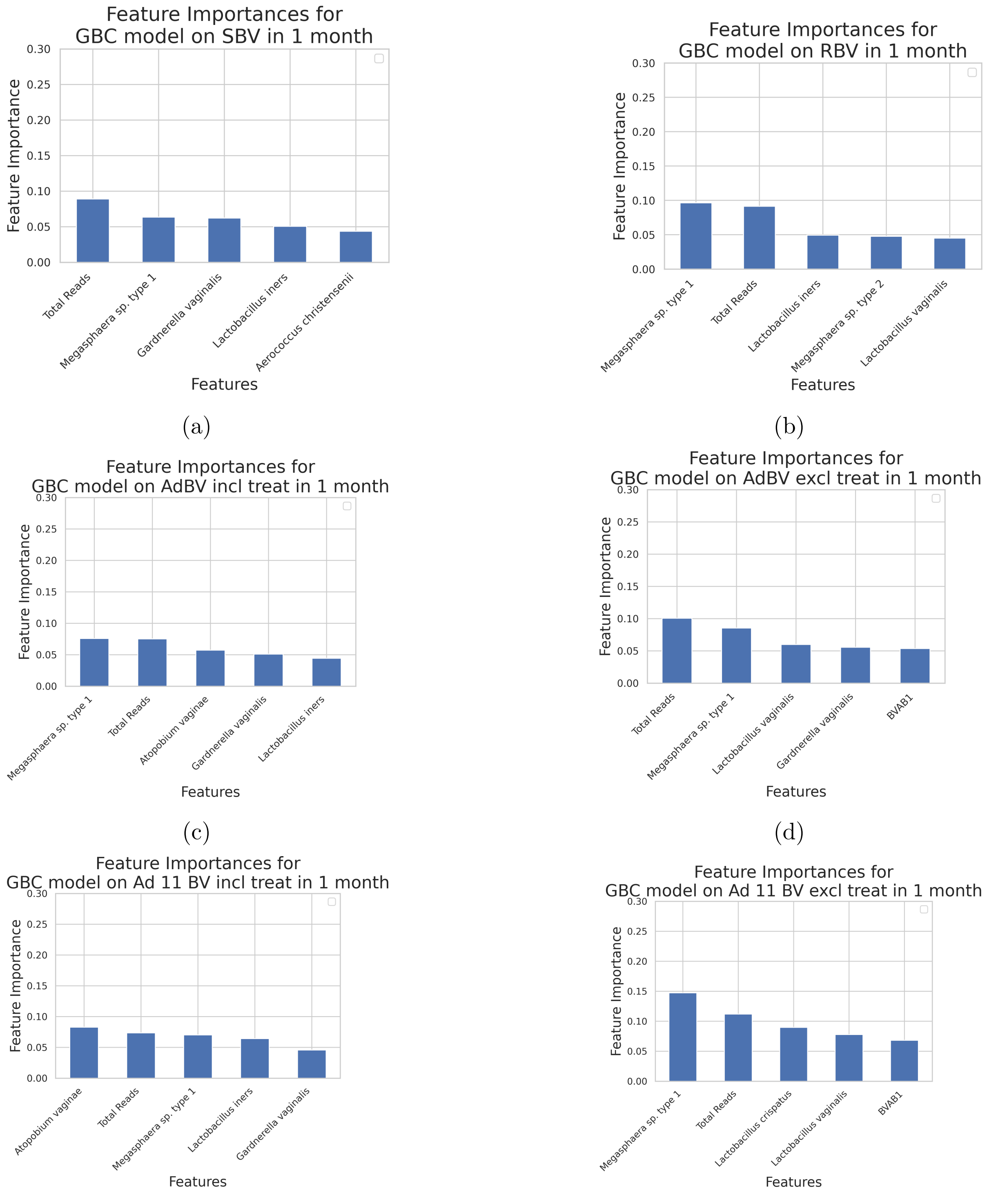


(d)

(e)

(f)

Figure A.10: The model performances for the best found ML models on the test data in a 5-fold repetition of the next month prediction task for each BV label type.



(a)

(b)

(c)

(d)

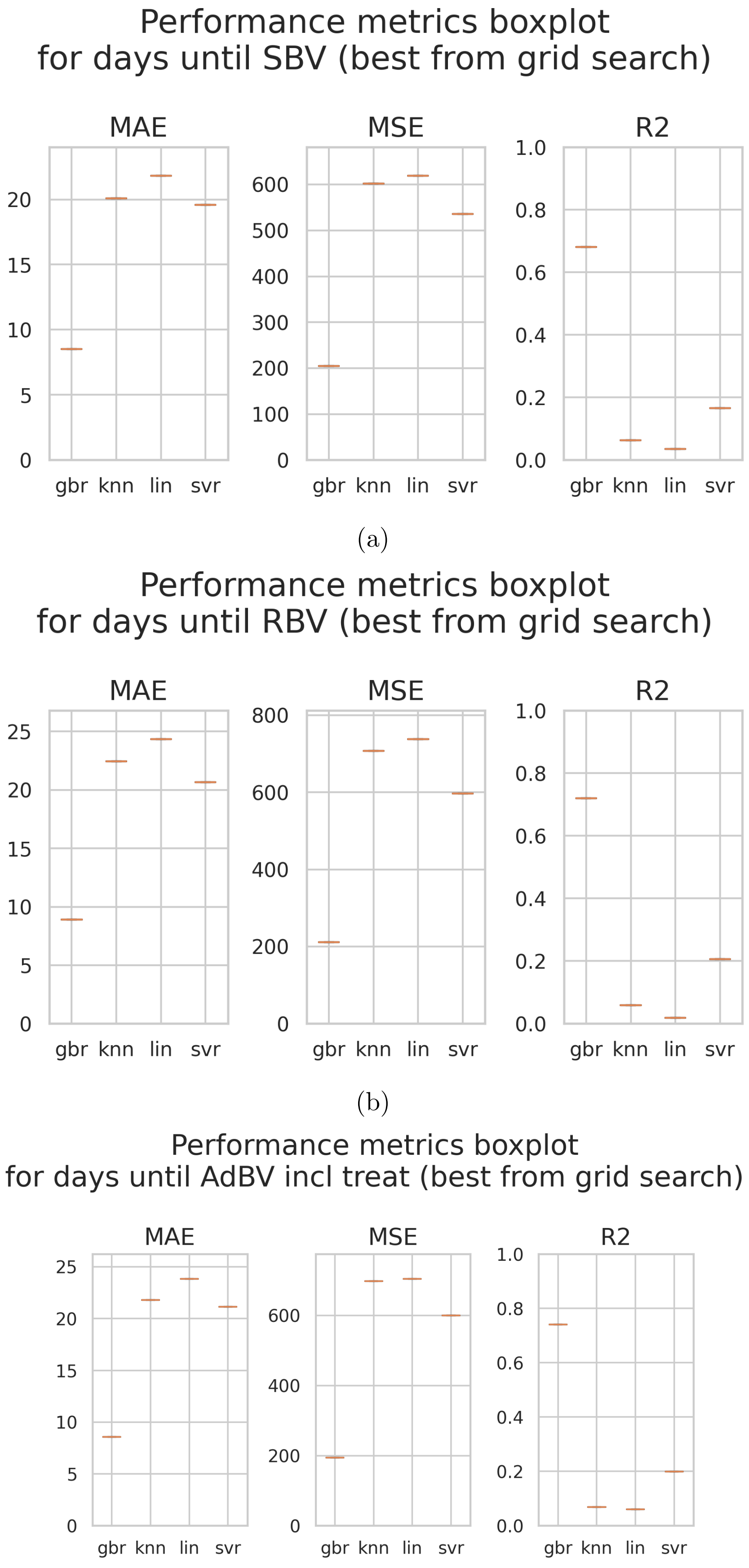
(e)

(f)

Figure A.11: The feature importances for the best found performances by the gbc model on the test data in a 5-fold repetition of the next month BV prediction task for each BV label type.

A.5 Regression: Days Until BV Onsets

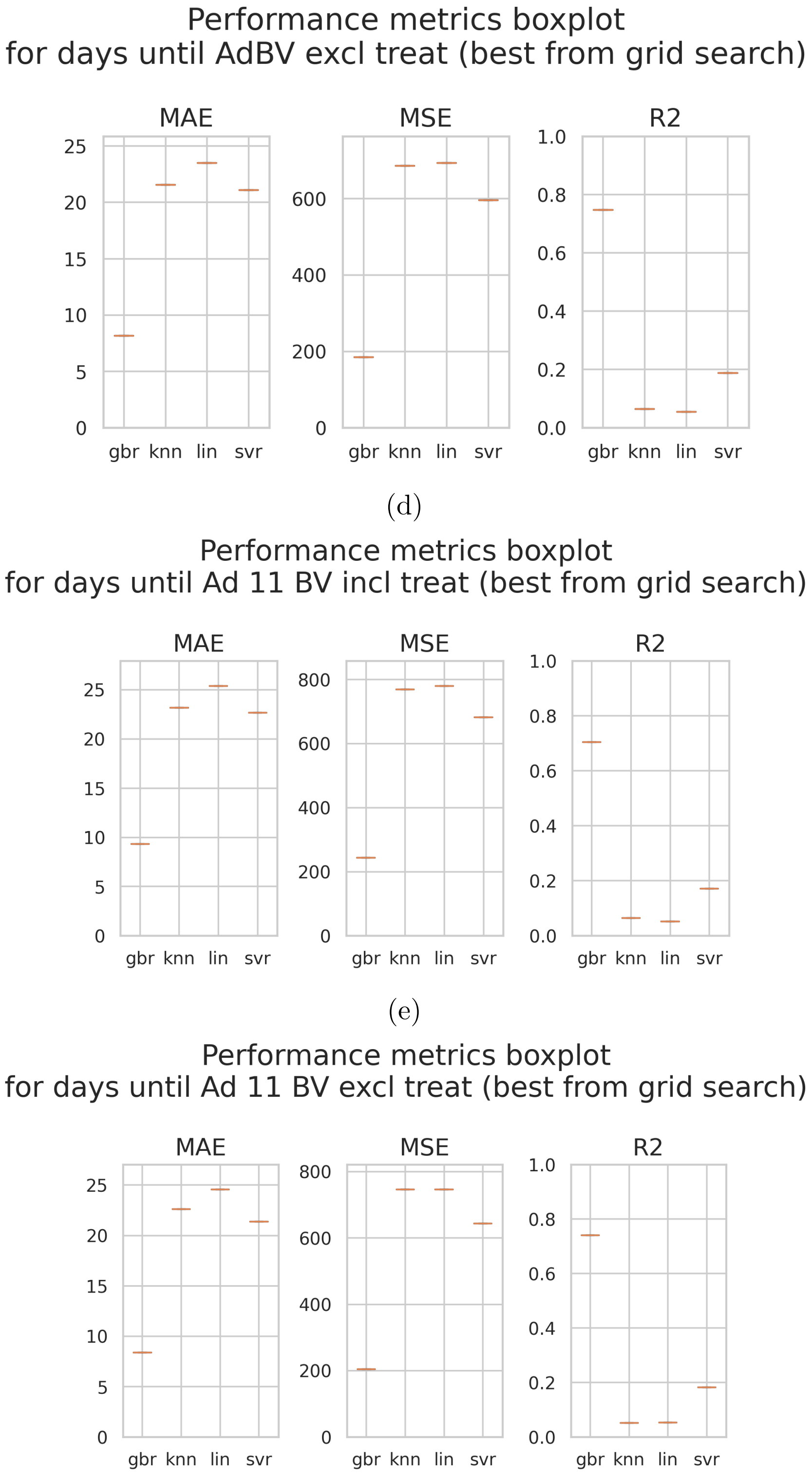
A.5.1 R^2 Optimized


Performance metrics boxplot
for days until SBV (best from grid search)
MAE
MSE
R2
0
5
10
15
20
0
100
200
300
400
500
600
0.0
0.2
0.4
0.6
0.8
1.0
gbr
knn
lin
svr
Performance metrics boxplot
for days until RBV (best from grid search)
MAE
MSE
R2
0
5
10
15
20
25
0
200
400
600
800
0.0
0.2
0.4
0.6
0.8
1.0
gbr
knn
lin
svr
Performance metrics boxplot
for days until AdBV incl treat (best from grid search)
MAE
MSE
R2
0
5
10
15
20
25
0
200
400
600
0.0
0.2
0.4
0.6
0.8
1.0
gbr
knn
lin
svr

(a)

(b)

(c)

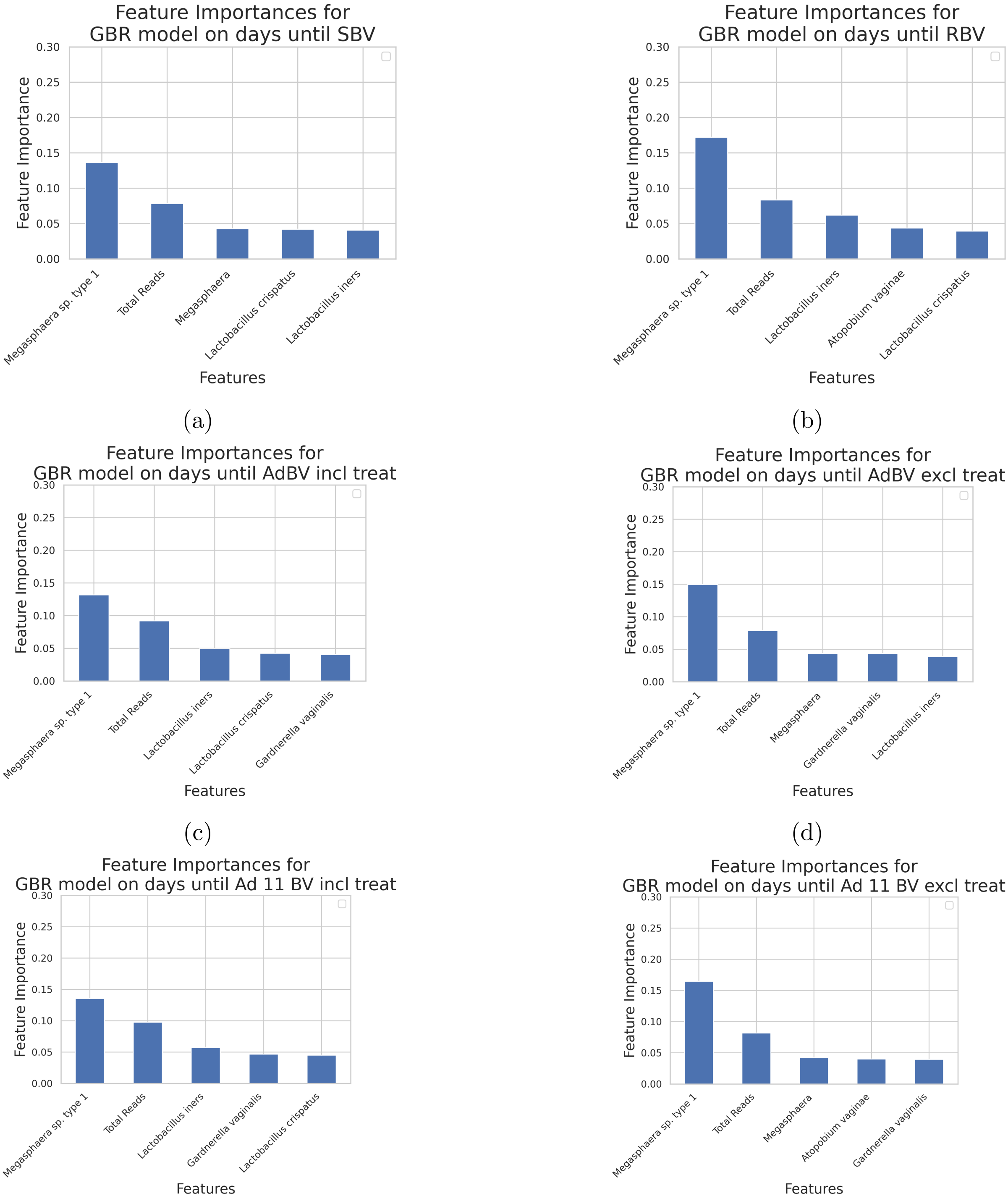


(d)

(e)

(f)

Figure A.12: The model performances for the best found ML models on the test data in a 5-fold repetition of this month prediction task for each BV label type.



(a)

(b)

(c)

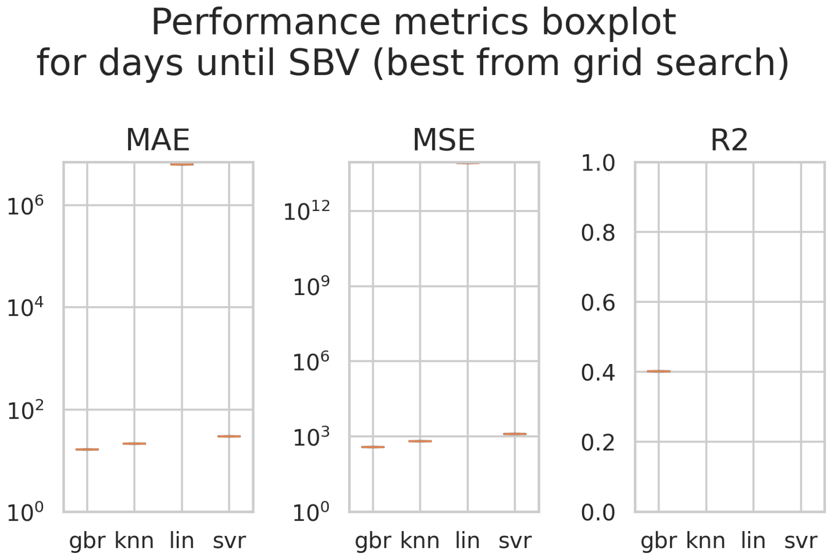
(d)

(e)

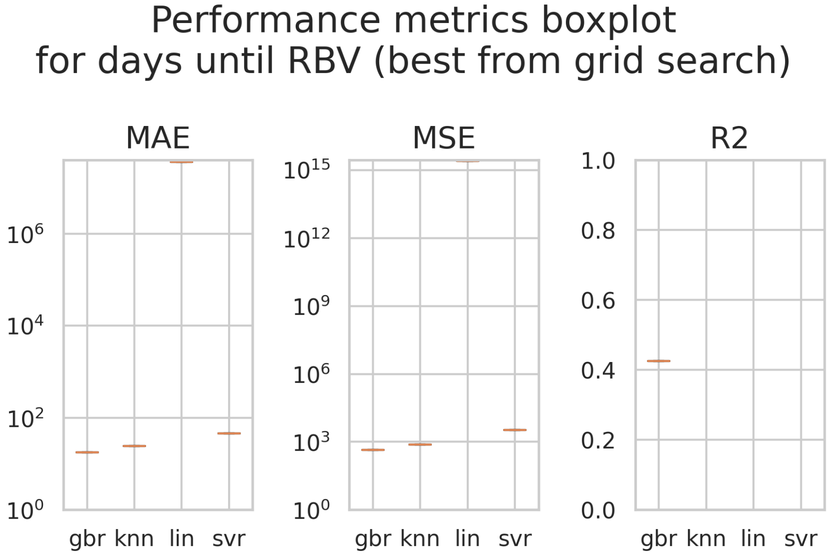
(f)

Figure A.13: The feature importances for the best found performances by the gbr model on the test data in a 5-fold repetition of this month BV prediction task for each BV label type.

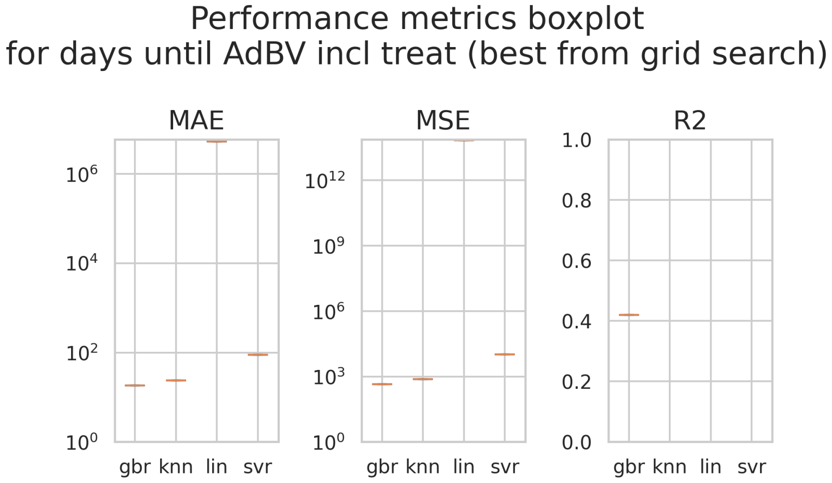
A.5.2 MAE Optimized


Performance metrics boxplot
for days until SBV (best from grid search)
MAE
10^{6}
10^{4}
10^{2}
10^{0}
gbr knn lin svr
MSE
10^{12}
10^{9}
10^{6}
10^{3}
10^{0}
gbr knn lin svr
R2
1.0
0.8
0.6
0.4
0.2
0.0
gbr knn lin svr

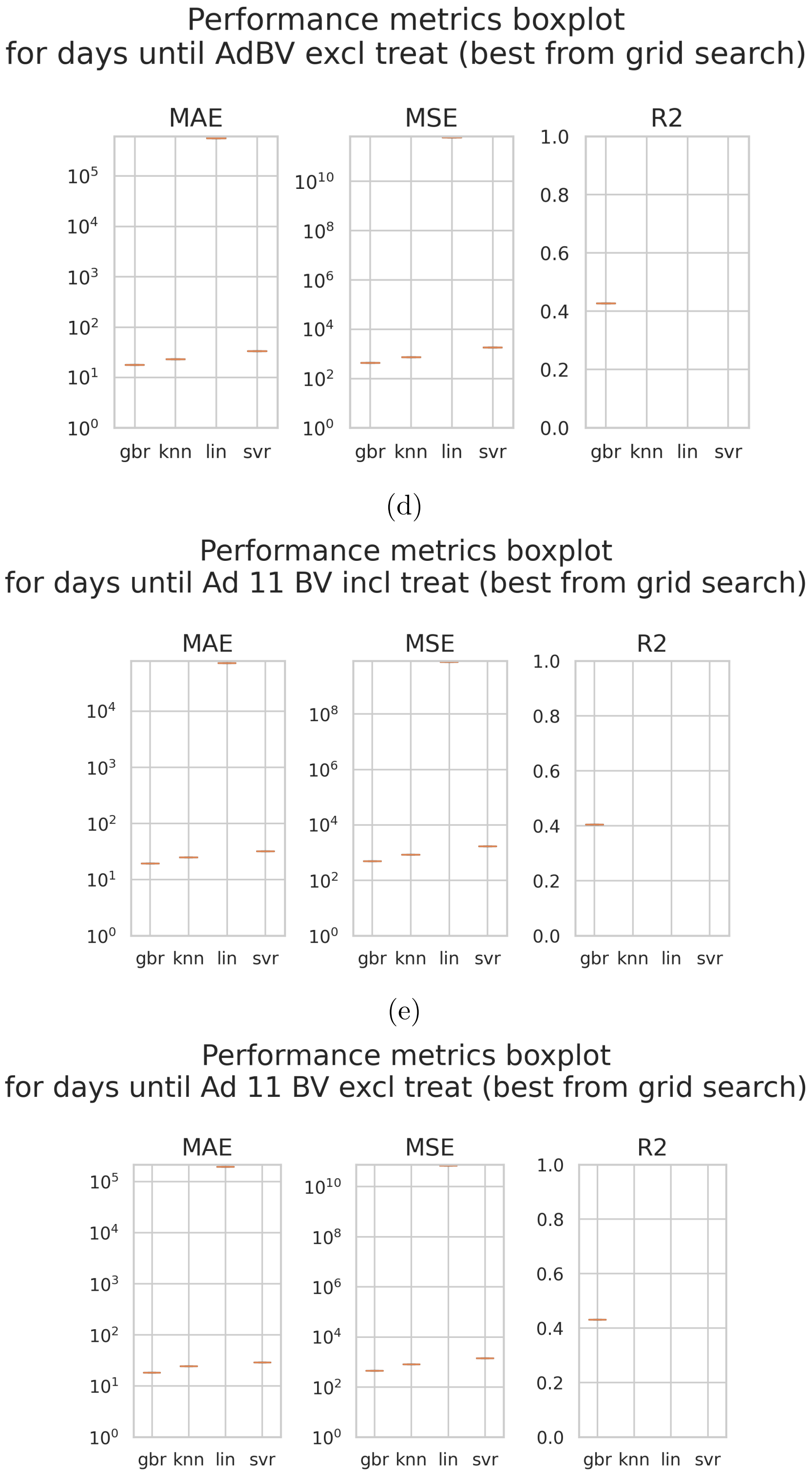
(a)


Performance metrics boxplot
for days until RBV (best from grid search)
MAE
10^{6}
10^{4}
10^{2}
10^{0}
gbr knn lin svr
MSE
10^{15}
10^{12}
10^{9}
10^{6}
10^{3}
10^{0}
gbr knn lin svr
R2
1.0
0.8
0.6
0.4
0.2
0.0
gbr knn lin svr

(b)


Performance metrics boxplot
for days until AdBV incl treat (best from grid search)
MAE
10^{6}
10^{4}
10^{2}
10^{0}
gbr knn lin svr
MSE
10^{12}
10^{9}
10^{6}
10^{3}
10^{0}
gbr knn lin svr
R2
1.0
0.8
0.6
0.4
0.2
0.0
gbr knn lin svr

(c)

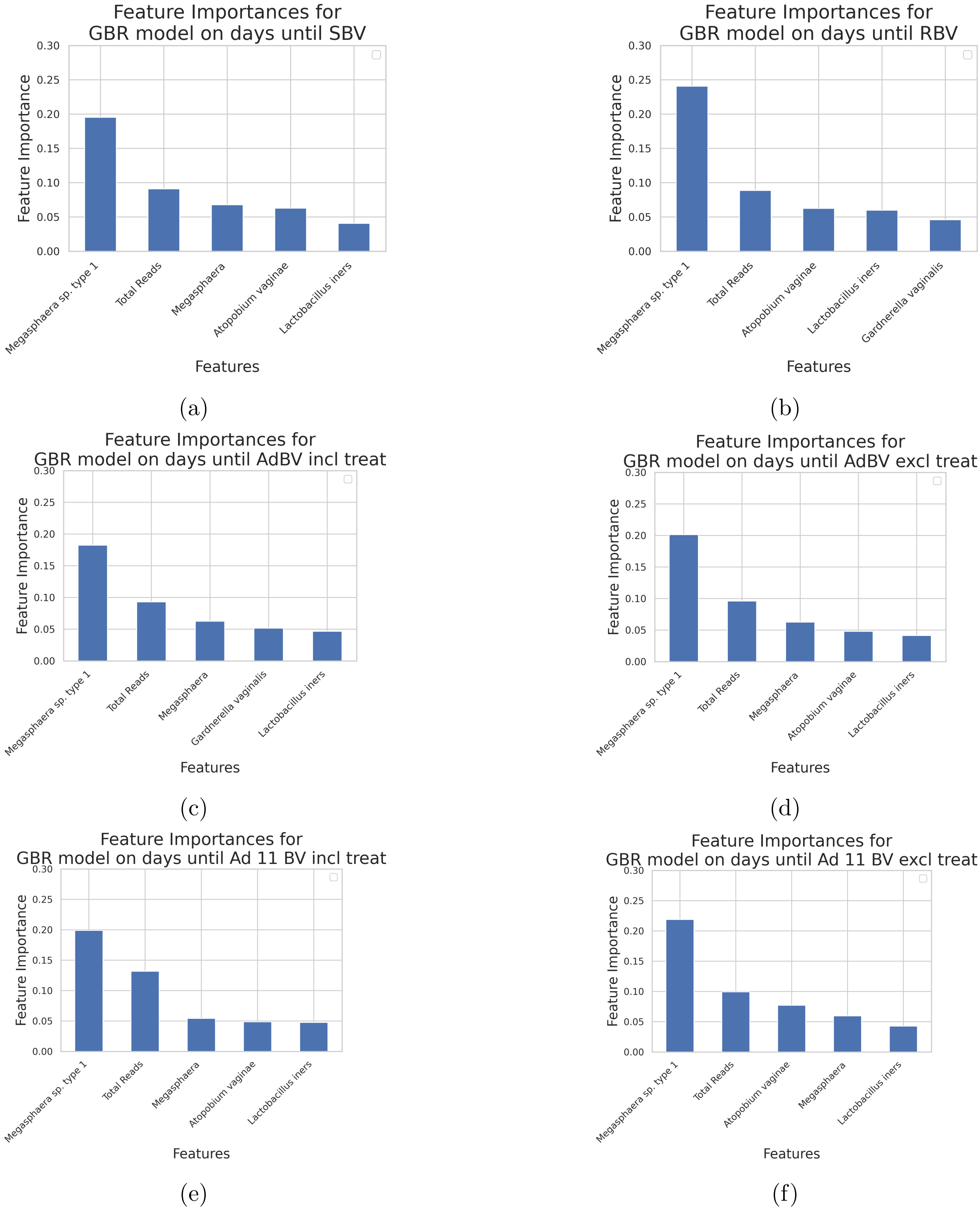


(d)

(e)

(f)

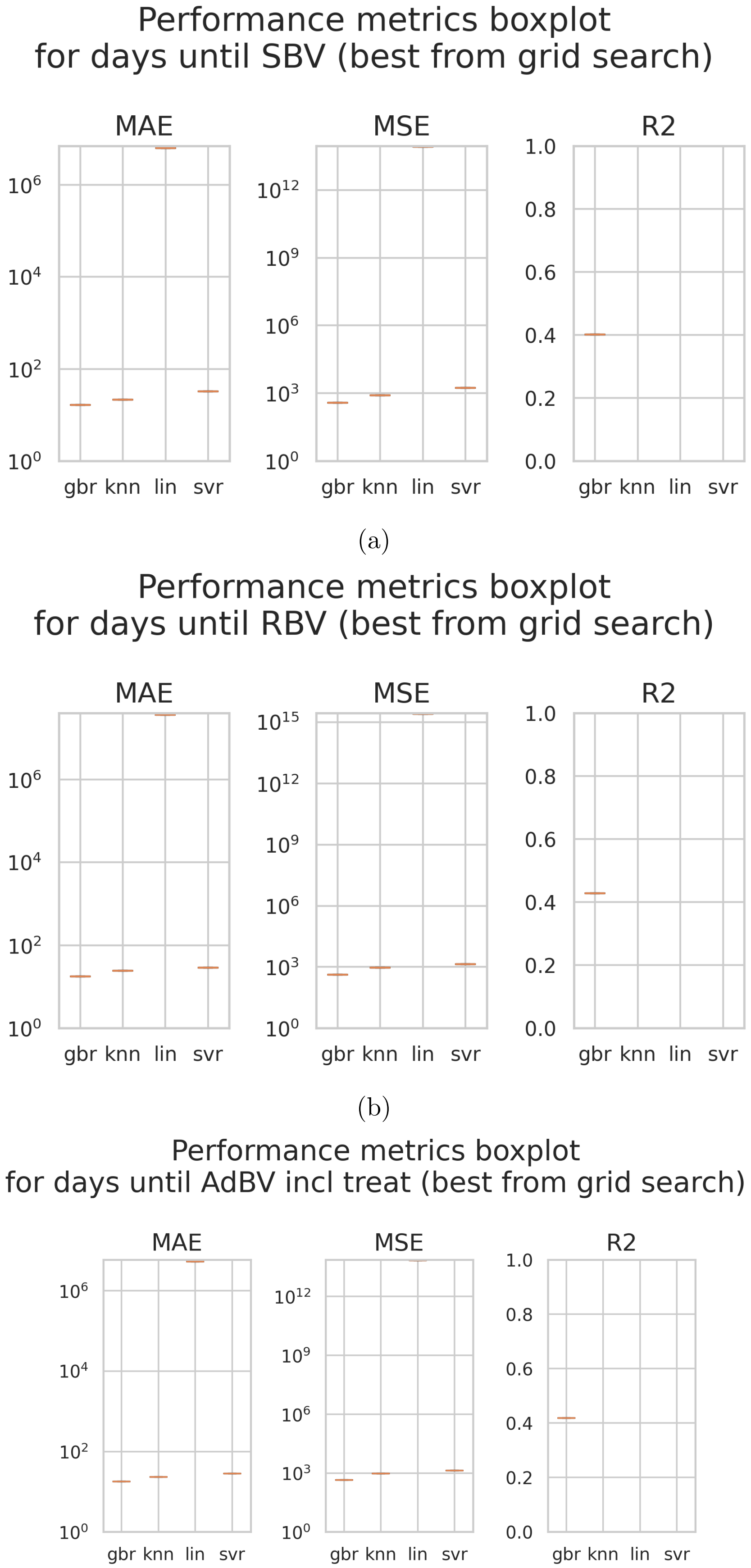
Figure A.14: The model performances for the best found ML models on the test data in a 5-fold repetition of the regression task optimized on the MAE for each BV label type.



(a) (b) (c) (d) (e) (f)

Figure A.15: The feature importances for the best found performances by the gbr model on the test data in a 5-fold repetition of the MAE optimized days until BV regression task for each BV label type.

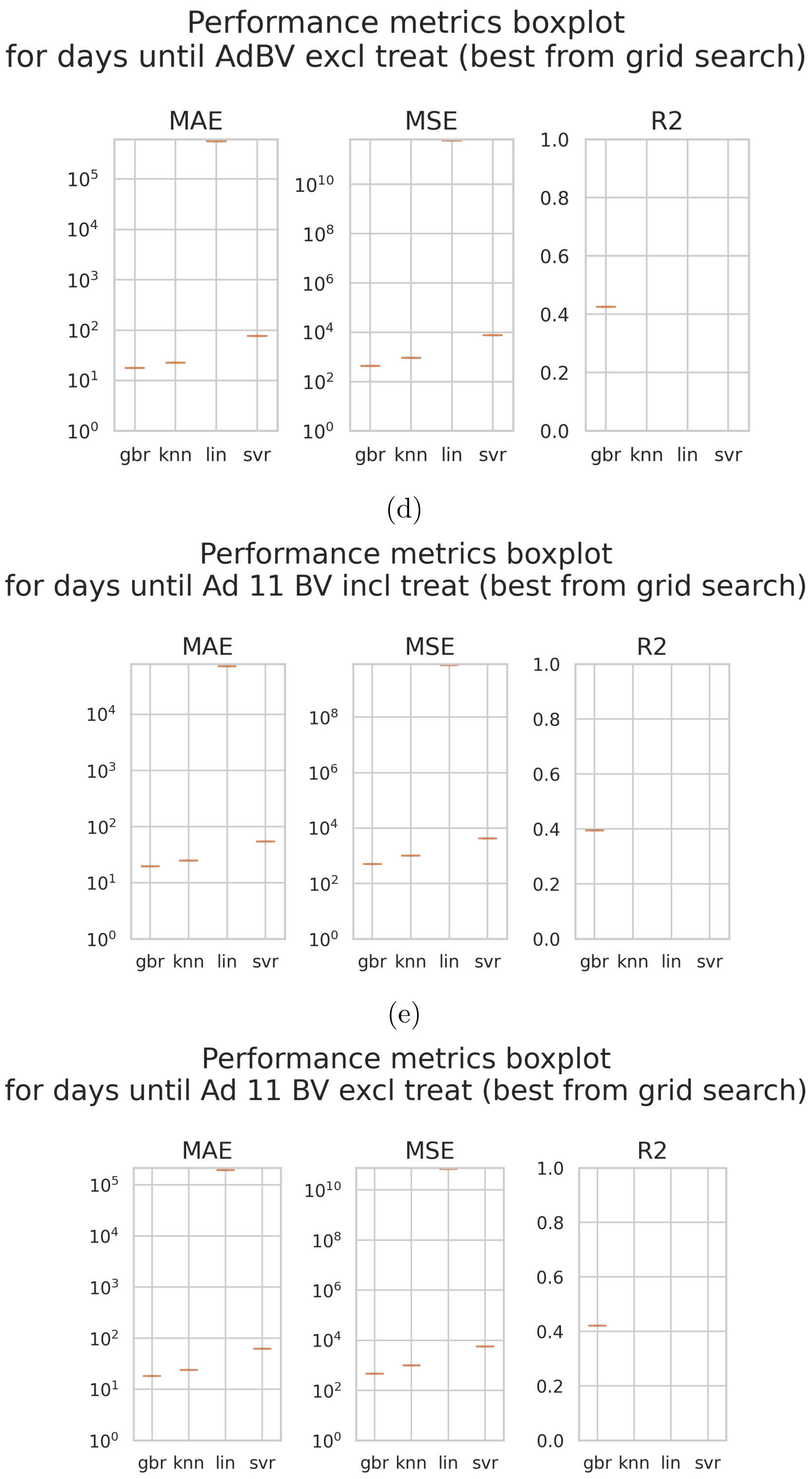
A.5.3 MSE Optimized


Performance metrics boxplot
for days until SBV (best from grid search)
MAE
MSE
R2
10^6
10^4
10^2
10^0
10^{12}
10^9
10^6
10^3
10^0
1.0
0.8
0.6
0.4
0.2
0.0
gbr knn lin svr
Performance metrics boxplot
for days until RBV (best from grid search)
MAE
MSE
R2
10^6
10^4
10^2
10^0
10^{15}
10^{12}
10^9
10^6
10^3
10^0
1.0
0.8
0.6
0.4
0.2
0.0
gbr knn lin svr
Performance metrics boxplot
for days until AdBV incl treat (best from grid search)
MAE
MSE
R2
10^6
10^4
10^2
10^0
10^{12}
10^9
10^6
10^3
10^0
1.0
0.8
0.6
0.4
0.2
0.0
gbr knn lin svr

(a)

(b)

(c)



Performance metrics boxplot
for days until AdBV excl treat (best from grid search)

(d)

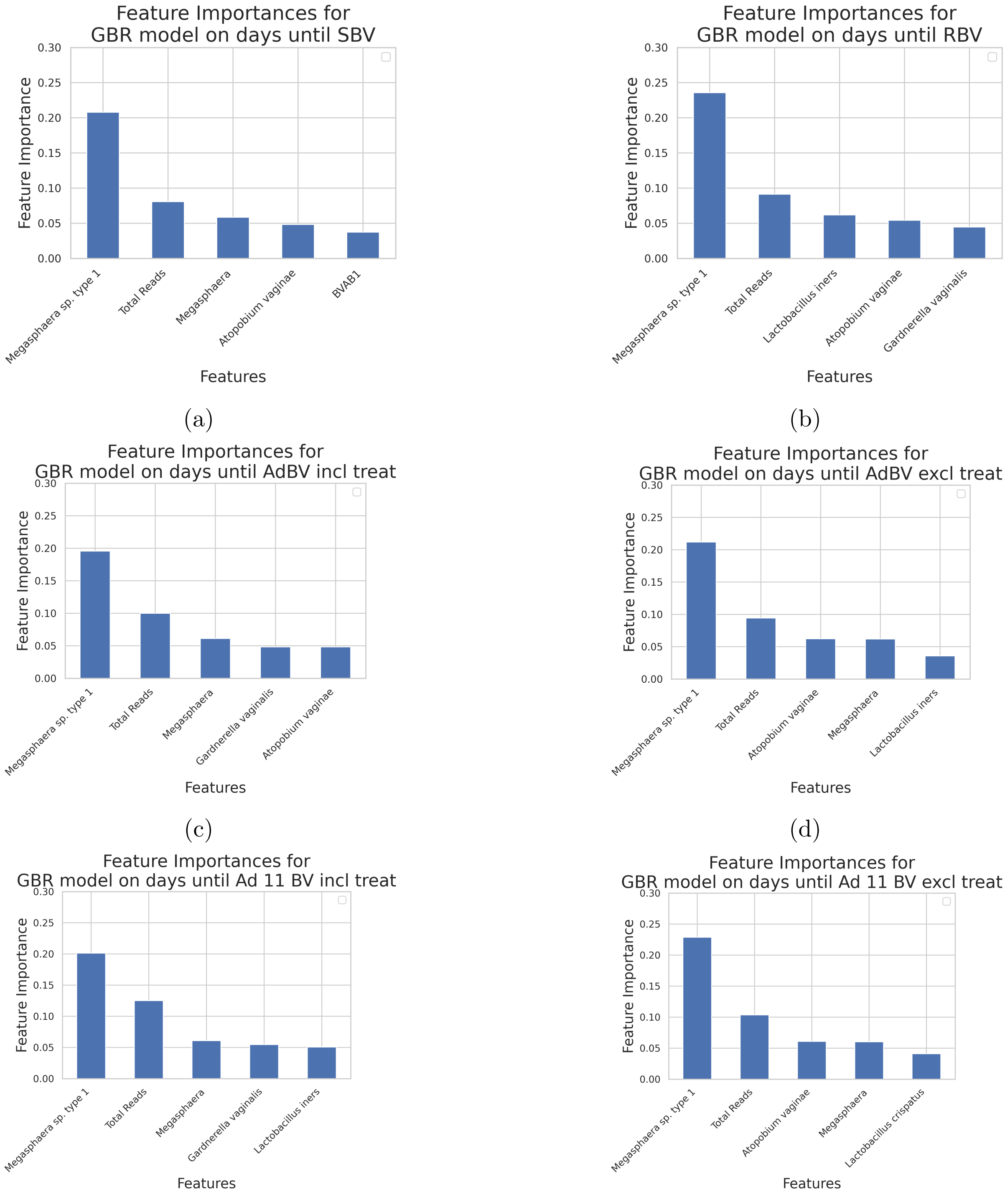
Performance metrics boxplot
for days until Ad 11 BV incl treat (best from grid search)

(e)

Performance metrics boxplot
for days until Ad 11 BV excl treat (best from grid search)

(f)

Figure A.16: The model performances for the best found ML models on the test data in a 5-fold repetition of the MSE optimized regression task for each BV label type.



(a)

(b)

(c)

(d)

(e)

(f)

Figure A.17: The feature importances for the best found performances by the gbr model on the test data in a 5-fold repetition of the MSE optimized regression task for each BV label type.