UNIVERSITÀ
DI PAVIA

SCUOLA DI ALTA FORMAZIONE DOTTORALE

MACRO-AREA SCIENZE DELLA VITA

PhD in Psychology, Neuroscience and Data Science
Department of Brain and Behavioral Sciences

Leveraging Machine Learning for Innovation in Public and Medical Healthcare

---

Academic year 2024-2025
Cycle XXXVIII

Coordinator
Prof.ssa Gabriella Bottini

Doctoral candidate
Dr. Federico Fassio

Tutor
Prof. Simona Villani

Summary

1. Introduction

In recent decades, the public health sector has undergone a significant transformation thanks to technological advancement, especially with the introduction of artificial intelligence models and advanced algorithms. As computational capabilities have increased and enormous amounts of health data have become available, it has become possible to address and solve complex problems in ways that, until recently, were unthinkable. In this context, data analysis is taking on a crucial role in improving health policies, identifying new disease patterns, improving prevention and improving the management of health resources[1,2].

Public health, understood as the field of study that deals with the prevention of diseases, the promotion of well-being and the management of social and economic factors that influence the health of the population, has always had as its main objective the improvement of living conditions of the community. However, the challenges of the XXI century, such as an aging population, the increase in chronic diseases, global pandemics and health inequalities, require new analytical methodologies. Machine learning algorithms represent potentially revolutionary solutions capable of addressing these issues, but their application requires careful consideration of both benefits and limitations[3,4].

Among the enormous amount of innovative approaches, community detection algorithm and novelty/anomaly detection models could have an important role in this field, with different applications[5–7].

This thesis examines these two different machine learning approaches, still relatively unexplored in healthcare and global health, across two different contexts defined by the target and the nature of the data (aggregated vs. individual ones). To facilitate readability, the thesis is structured into two mirrored sections, followed by a single concluding section at the end.

2. Community detection

2.1. Background

Community detection is a network analysis technique that identifies groups of nodes densely connected to one another but sparsely connected to the rest of the network. Originally developed within social network studies, it has since been applied across numerous fields, including biology, psychology, political science, and public health[8–10].

Although there are similarities with clustering, important differences exist. Clustering is a machine learning technique that groups similar data points based on their attributes and is widely used in unsupervised learning across various data types. While clustering can be applied to network data, it is a more general approach not specifically designed for graph structures. In contrast, community detection is tailored for network analysis and relies on a single type of attribute (edges) to identify groups of densely connected nodes. Additionally, traditional clustering algorithms may isolate peripheral nodes, failing to assign them to the communities they logically belong to within a network. Despite these differences, both methods are valuable in network analysis and offer distinct advantages and limitations depending on the application domain[11].

In medical settings, community detection models could be particularly useful in several areas, for example:

- Identifying patient subgroups with shared characteristics, such as genetic variants or therapeutic responses, to facilitate personalized treatment plans.
- Detecting clusters of related pathologies across hospital networks to map disease transmission pathways and pinpoint outbreak epicentres.
- Uncovering functional modules in molecular networks (e.g., protein–protein or gene co-expression networks) to reveal novel drug targets and biomarkers.
- Grouping patients by imaging-feature similarity in radiology or histopathology to improve diagnostic classification of lesions and tumour subtypes.
- Segmenting clinical-note co-occurrence networks to discover hidden patient cohorts based on symptomatology or treatment narratives.

In public health, community detection can inform:

- Mapping mobility flow networks by clustering regions according to travel patterns, thereby guiding targeted vaccination or quarantine measures.
- Segmenting social media networks to isolate communities sharing specific health content, enabling precision behaviour-change campaigns and misinformation countermeasures.
- Locating psychological and social support networks among patients with chronic illnesses or psychiatric disorders to optimize peer-support programs.
- Identifying at-risk communities for tailored disease-prevention outreach ─ such as cardiovascular, diabetes, or obesity interventions ─ in defined geographic areas.
- Monitoring dynamic transmission networks in real time to track emerging epidemic clusters and allocate resources for containment.
- Analysing provider-referral networks to detect collaboration hubs and streamline public health service delivery.

Early approaches to understanding network structure often drew on methods from classical graph theory and social network analysis, where hierarchical clustering and block-modelling were used to uncover cohesive subgroups. Among these, hierarchical clustering was widely applied to uncover community divisions. However, network clustering fundamentally differs from traditional clustering methods applied to feature vectors. While classical clustering algorithms minimize intra-cluster variance by grouping data points with similar attributes in some metric space, network community detection focuses on connectivity patterns, seeking dense subgraphs where nodes are more connected to each other than to the rest of the network[11–13].

In network hierarchical clustering, a numerical similarity measure or edge weight is assigned between each pair of vertices to quantify relatedness. This is computed through various approaches, including connectivity-based measures derived from vertex- or edge-disjoint paths and max-flow ideas, or path-based measures that sum contributions from all possible paths while exponentially down-weighting longer routes so that short paths dominate[14,15]. Using agglomerative clustering, the process begins with isolated vertices and iteratively merges the most similar pairs, producing a nested hierarchy of

components represented as a dendrogram. Different cuts through this tree reveal community structure at varying resolutions.

While conceptually appealing and occasionally effective, these hierarchical approaches exhibit notable limitations specific to network structure. They may isolate degree-1 peripheral nodes from the communities to which they naturally belong and, in networks with known ground-truth communities, often fail to accurately recover the true partitions, in part due to sensitivity to weak inter-community links and the lack of a global notion of community quality[11–13].

In response, community detection has evolved beyond simple hierarchical methods, developing approaches that explicitly leverage network-specific properties. Modularity-based algorithms seek partitions that maximize the difference between observed and expected intra-community connections under a null model. Spectral clustering methods use eigenvectors of graph Laplacians to embed nodes into spaces where community structure becomes geometrically separable. Probabilistic models such as the stochastic block model formulate community detection as statistical inference on a generative model for community-structured networks. More recently, deep learning approaches employing graph neural networks can learn complex, hierarchical, or overlapping communities by capturing multi-scale connectivity patterns[14–18].

Each paradigm brings distinct assumptions about community structure, from the modularity optimization's focus on density contrasts to spectral methods' emphasis on cut minimization, and probabilistic models' explicit generative assumptions. This diversity enables adaptation to the growing complexity and scale of real-world networks while addressing the fundamental difference between clustering nodes based on connectivity versus clustering feature vectors based on attribute similarity.

The application of community detection methods to occupational health surveillance represents an emerging but promising direction. In healthcare systems analysis, community detection has proven effective in delineating health service areas and care regions based on patient flow patterns. Modularity optimization methods have successfully identified regional healthcare networks that more accurately capture actual care patterns compared to traditional geographic boundaries, showing superior performance in metrics such as localization index, market share index, and connectivity[19,20].

In occupational health specifically, network-based approaches have been developed to analyse disease-exposure associations in occupational surveillance systems. Faisandier et al. (2009) proposed a network-based approach for surveillance of occupational diseases that considers composite occupational exposures in their entirety when describing disease-exposure associations, generating hypotheses about complex exposure patterns. This method differs from traditional pharmacovigilance approaches by accounting for the interconnected nature of occupational exposures, where workers are typically exposed to multiple agents simultaneously rather than single isolated factors[21].

Hierarchical clustering has been specifically applied to systematically identify groups of jobs with similar occupational exposure patterns. In a case study using the New England Bladder Cancer Study, hierarchical cluster models were applied to diesel-related exposure variables, demonstrating that clustering can serve as a data reduction step to identify jobs with similar response patterns prior to expert exposure assessment. This approach has the potential to aid rule-based assessment by systematically reducing the number of exposure decisions needed while maintaining exposure estimate homogeneity within clusters[22].

For musculoskeletal health in industrial settings, hierarchical cluster analysis has been employed to identify similarities between workstations in terms of musculoskeletal stress factors and their impact on workers' bodies. In cardboard manufacturing, agglomerative clustering methods considering multiple risk variables and ergonomic factors have been used to group workstations with similar risk profiles[23].

Regarding petrochemical workers specifically, while direct applications of community detection algorithms are limited in the published literature, related network approaches have been applied to operational risk assessment in petrochemical plants. These methods focus on monitoring worker behavioural patterns and safety status, though they employ different analytical frameworks such as micro-Doppler monitoring rather than classical community detection[24].

The translation of community detection methods from healthcare network analysis to occupational health surveillance holds significant potential. Just as modularity-based methods have successfully identified emergency surgery care regions and patient referral communities, similar approaches could delineate occupational health surveillance regions based on shared exposure patterns, industrial processes, or disease occurrence networks. The ability of these methods to be scale-flexible, automated, and responsive to changes

over time makes them particularly suitable for dynamic occupational health surveillance systems where exposure profiles and industrial practices evolve. However, further research is needed to adapt these community detection frameworks specifically to the unique characteristics of petrochemical and heavy industry occupational health data, where exposure complexity, temporal dynamics, and regulatory structures differ substantially from healthcare delivery networks[20–22].

2.2. Objective

This first section aims to assess whether community clustering can reliably identify macro-groups in data that has been aggregated from the outset. The central question is whether meaningful community structure remains detectable when individual-level information is partially obscured for privacy reasons, as is often the case in occupational health and corporate surveillance settings. In this context, the analysis focuses not on individuals but on production sites as nodes in a network, linked by similarity in their employees' diagnostic-pathway engagement. By treating sites as vertices and their screening profiles as the basis for edge weights, community detection allows the emergence of higher-order organization to be studied without compromising worker confidentiality.

To reach this aim, community-detection models are applied to an aggregated dataset from a major Italian petrochemical company, compiled specifically to preserve company privacy while retaining essential structural information on health service use. The dataset summarizes employee participation in on-site health screening programs and downstream diagnostic examinations across multiple production sites, yielding a site-by-site similarity structure in terms of how workers use preventive and diagnostic services. Community detection on this network is used to uncover natural groupings of production sites that share similar patterns of engagement along diagnostic pathways, capturing both intensity and composition of participation rather than only simple uptake rates.

Within this framework, well-defined clusters of production sites are expected to emerge, representing macro-groups with comparable preventive-health and diagnostic profiles. Such clusters can be interpreted as latent "occupational health communities" within the company, potentially reflecting shared production processes, workforce characteristics,

organizational practices, or local health-management cultures that are not explicitly encoded in the data. Identifying these macro-groups may highlight specific subsets of sites where participation in screening or follow-up diagnostics is systematically lower or more fragmented, suggesting opportunities for targeted public health interventions, harmonization of health promotion strategies, or tailored communication campaigns to improve access and adherence. Conversely, clusters with consistently high and coordinated engagement can serve as benchmarks or best-practice references for internal occupational health policy development.

2.3. Materials

The dataset covers health-promotion initiatives, screening programs, and wellness tests offered by Versalis (a subsidiary of ENI). Data are provided in aggregated form ─ the total number of participants in each initiative or diagnostic test ─ broken down by four participant types: employees, contractors, in-house workers, and family members. The analysis spans two years (2022–2023), with 29 sites in 2022 and 35 in 2023. The total number of clinical examinations or screening initiatives is 126, for both years. They included screening for infectious diseases (e.g., tuberculosis, pertussis), routine medical examinations, imaging studies, audiometry, cardiovascular, dermatological, pulmonary, and ophthalmological evaluations, as well as alcohol screening tests and vaccinations.

This is the list of company sites and examinations are reported in Tables 1 and 2.

Sites	***2022***	***2023***
Brindisi Servizi Generali S. c. a r. l.(BSG)	✓	✓
Dunastyr Polisztirolgyarto Zartkruen Mukodo Részvénytarsasag - Stabilimento di Szazhalombatta	✓	✓
FINPROJECT S.p.A. - Ancarano/Castorano		✓
FINPROJECT S.p.A. – Morrovalle		✓
FINPROJECT S.p.A. - Roccabianca		✓
FINPROJECT S.p.A. - Ascoli Piceno		✓
Ravenna Servizi Industriali S.C.p.A. (RSI)	✓	✓
Servizi Porto Marghera S.c. a r. l. (SPM)	✓	✓
Versalis Americas Inc	✓	✓
Versalis Congo Sarlu	✓	
Versalis Deutschland GmbH - Stabilimento di Oberhausen	✓	✓
Versalis France SAS - Stabilimento di Dunkerque	✓	✓
Versalis International SA - Sedi commerciali	✓	✓
Versalis Kimya Ticaret Limited Sirketi		✓

Sites	*2022*	*2023*
Versalis México S. de R.L. de C.V.	✓	✓
Versalis Pacific (India) Private Limited	✓	✓
Versalis Pacific Trading (Shanghai) Co Ltd	✓	✓
Versalis Singapore Pte Ltd	✓	✓
Versalis SpA - Oilfield	✓	✓
Versalis SpA - R&D and Green Chemistry Research Center Novara	✓	✓
Versalis SpA - R&D and Green Chemistry Research Center Rivalta Scrivia	✓	✓
Versalis SpA - San Donato Milanese + Sedi Commerciali	✓	✓
Versalis SpA - Stabilimento di Brindisi	✓	✓
Versalis SpA - Stabilimento di Crescentino	✓	✓
Versalis SpA - Stabilimento di Ferrara	✓	✓
Versalis SpA - Stabilimento di Mantova	✓	✓
Versalis SpA - Stabilimento di Porto Marghera	✓	✓
Versalis SpA - Stabilimento di Porto Torres	✓	✓
Versalis SpA - Stabilimento di Priolo	✓	✓
Versalis SpA - Stabilimento di Ragusa	✓	✓
Versalis SpA - Stabilimento di Ravenna	✓	✓
Versalis SpA - Stabilimento di Sarroch	✓	✓
Versalis UK Ltd - Stabilimento di Grangemouth	✓	✓
Versalis Zeal Ltd	✓	✓

Table 1. Company sites included in the dataset, stratified by year. A tick indicates the presence of the site in the dataset for the corresponding year.

Examination	*2022*	*2023*
Other diagnostic images (CT, MRI, ultrasound, angiography, etc)	✓	✓
Diagnostics: other instrumental examinations	✓	✓
X-rays	✓	✓
Laboratory analysis	✓	
Medical Referrals - Inpatient		✓
Medical Referrals - Outpatient		✓
Medical consultation in company	✓	✓
Medical consultation in outsourcing	✓	✓
A00-B99.I Certain infectious and parasitic diseases	✓	✓
C00-D48.II Neoplasms	✓	
E00-E90.IV Endocrine, nutritional and metabolic diseases	✓	✓
F00-F99.V Mental and behavioural disorders	✓	✓
G00-G99.VI Diseases of the nervous system	✓	✓

Examination	*2022*	*2023*
H00-H59.VII Diseases of the eye and anexa	✓	✓
H60-H95.VIII Diseases of the ear and mastoid process	✓	✓
I00-I99.IX Diseases of the circulatory system	✓	✓
J00-J99.X Diseases of the respiratory system	✓	✓
K00-K93.XI Diseases of the digestive system	✓	✓
L00-L99.XII Diseases of the skin and subcutaneous tissue	✓	✓
M00-M99.XIII Diseases of the musculoskeletal system and connective tissue	✓	✓
N00-N99.XIV Diseases of the genitourinary system	✓	✓
O00-O99.XV Pregnancy, childbirth and the puerperium		✓
Q00-Q99.XVII Congenital malformations, deformations and chromosomal abnormalities	✓	
R00-R99.XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	✓	✓
S00-T98.XIX Injury, poisoning and certain other consequences of external causes	✓	✓
Z00-Z99.XXI Factors influencing health status and contact with health services	✓	✓
A02-A09.Infections, poisoning and food-borne intoxications		✓
A09.Infective Diarrhoea	✓	
A31.Infection due to other mycobacteria (different than TB and Leprosy)	✓	
A38.Scarlet fever		✓
Other diseases (report in notes)	✓	
U07.1 Covid-19, virus identified	✓	✓
How many for occupational health including travel medicine	✓	✓
Vaccinations executed for health promotion	✓	✓
Negative FtW	✓	✓
Positive FtW	✓	✓
Positive FtW with prescriptions	✓	✓
Positive FtW with restrictions	✓	✓
Audiometries	✓	✓
Electrocardiogram (ECG)	✓	✓
N. ophtalmology exams (ergo vision)	✓	✓
Other cardiological instrumental examinations (echocardiography, stress test, etc.)	✓	✓
Other_instrum_exam	✓	✓
Spirometries	✓	✓
X-rays exams	✓	✓
Alcohol test with a negative result	✓	✓
Biological exposure indicators (from urine, blood, air, other - creatinineuria excluded)	✓	✓
CDT (Desalted Transferrin) normal	✓	✓
CDT (Desragate Transferrin) altered	✓	✓
Drug test with a negative result	✓	✓
Drug test with a positive result	✓	✓
Alcohol screening: 2nd level	✓	✓
Alcohol screening: 1st level	✓	✓
Hematourochemical examinations	✓	✓
Extraordinary visits	✓	✓
Preventive visits in pre-assumption phase	✓	✓

Examination	*2022*	*2023*
Visits at the request of the worker	✓	✓
Visits before the resumption of work	✓	✓
Visits for change of work task	✓	✓
Visits prior to termination of employment	✓	✓
Visits for trip mission	✓	✓
Periodical medical examinations	✓	✓
Medical Examinations for Regulatory Compliance (pre-shift and post shift, etc.)	✓	✓
Cardiological visits	✓	✓
Dermatological visits	✓	
Eye examinations	✓	✓
Neurological visits	✓	✓
Orthopedic visits	✓	✓
Other specialist visits	✓	✓
Otolaryngology visits	✓	✓
Other Work prescriptions	✓	✓
Physical agents (noise)	✓	✓
Videoterminals	✓	✓
Biological agents	✓	✓
Expatriation - international travel	✓	✓
Hazardous substances	✓	✓
Manual handling of loads	✓	✓
N. work at height - confined environments	✓	✓
Other Work restrictions	✓	✓
Physical agents	✓	✓
Working conditions	✓	✓
Works in shifts	✓	✓

Table 2. Diagnostic and clinical tests and examinations performed at the various sites, presented by year. A tick indicates the presence of the corresponding feature in the dataset for that year.

2.4. Methods

To facilitate the analyses, participation in initiatives and diagnostic tests was considered only for workers employed by the company. Contractors and workers' families were excluded due to missing or unrepresentative data. The five sites added in 2023 that were not present in 2022 ─ FINPROJECT S.p.A. Ancarano/Castorano, FINPROJECT S.p.A. Morrovalle, FINPROJECT S.p.A. Roccabianca, FINPROJECT S.p.A. Ascoli Piceno, Versalis Kimya Ticaret Limited Sirketi ─ were also excluded to ensure comparability of results across both years. The data were analysed separately for each of the two years.

Since participation in health examinations was reported in absolute terms, it was normalized by dividing it by the average annual number of employees at each work site. As some initiatives were repeated over time without indicating their frequency, any resulting rate above 100% (i.e., values greater than 1) was arbitrarily capped at 100%. To reduce computational time and eliminate redundant, non-informative data ─ within each year considered ─ rows corresponding to diagnostic tests with a participation rate of zero across all industrial locations were removed.

For 2022, a matrix of 77 rows (diagnostic exams) and 29 columns (industrial sites) was constructed. For the community analysis, the 29 sites were treated as the nodes of a graph, with the edges representing the similarity in participation patterns across diagnostic exams. For 2023, diagnostic procedures included in the analysis were 75.

To employ a community detection algorithm, it is necessary to build a similarity matrix that serves as the basis for graph construction.

In this structure, each node corresponds to an entity, and each edge is weighted to reflect how similar or dissimilar two nodes are. Stronger connections indicate greater similarity in the chosen feature space. Calculating those weights typically begins with constructing a distance (or dissimilarity) matrix from all pairwise measurements, which is then converted into a weighted adjacency matrix, the key input for most community detection algorithms.

A variety of distance metrics can be used to quantify the relationships between nodes, each with its own assumptions and suitability for different types of data.

In this analysis, to quantify these relationships, eight similarity matrices were generated using different estimation methods: Jaccard, Weighted Jaccard, Dice coefficient, Cosine, Kendall' τ, Normalized Euclidean, Normalized Mutual Information, Overlap coefficient and the Gram matrix.

The choice of these matrices was arbitrary, aiming to explore a range of methods with distinct conceptual approaches, each offering different advantages and disadvantages. Here, we briefly explain the main concept of the selected methods.

2.4.1. Euclidean distance

It's defined as the distance between two points in Euclidean space and aim to find the straightest and shortest path between two points:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

More generally, it can be extended to calculate the distance between two vectors:

$$D_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

Where n is the length of the vector (i.e. the number of elements in each vector. In a matrix, this represents the number of features or dimensions in a data frame) and k represents the sum (it represents each component in the vectors x_i and x_j.

It has the advantage of being intuitive, computationally efficient, and suitable for data with large magnitude differences. However, it is sensitive to scale across dimensions (requiring normalization), suffers from the curse of dimensionality in high-dimensional spaces, and is strongly affected by outliers due to the influence of squared terms.

To convert distances into similarities, different approaches may be employed. For simplicity, we transform Euclidean distance into similarity using a simple inversion:

$$S_{ij} = \frac{1}{1 + D_{ij}}$$

This formulation bounds similarities in $(0,1](0,1]$, with $S_{ij} = 1$ if $D_{ij} = 0$, so that "zero distance" becomes perfect affinity. As distances increase, similarity values decrease smoothly toward zero.

2.4.2. Cosine similarity

It focuses on the angle (proportional patterns) instead of captures straight-line magnitude differences (as Euclidean distance does). In clustering approach, it could be useful to estimate instead cosine dissimilarity: 1- cosine similarity.

$$S_{ij} = \frac{x_i x_j}{\|x_i\| \, \|x_j\|} = \frac{\sum_{k=1}^{n} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{n} x_{ik}^2} \sqrt{\sum_{k=1}^{n} x_{jk}^2}}$$

Where $\|x_i\|$ $\|x_j\|$ are the Euclidean norms (magnitudes) of the vectors x_i and x_j.

Among the advantages of this method, there is the mitigation of the "curse of dimensionality" compared to Euclidean distance. In very high dimensions, Euclidean distances tend to concentrate, making all pairs look similarly distant. Cosine dissimilarity preserves angular separation, improving contrast between truly similar and dissimilar vectors[25].

The cosine dissimilarity (1- cosine similarity) ranges from 0 to 2.

2.4.3. Jaccard index

It measures the similarity (or dissimilarity) between two sets by comparing the size of their intersection to their union, focusing on whether elements are present or absent in vectors, ignoring any notion of magnitude or frequency.

$$S_{ij} = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} = \frac{\sum_{k=1}^{n} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{n} \max(x_{ik}, x_{jk})}$$

Where n is the length of the vector (i.e. the number of elements in each vector. In a matrix, this represents the number of features or dimensions in a data frame) and k represents the sum (it represents each component in the vectors x_i and x_j. In case of zeros, these values are ignored, and only the presence of non-zero counts is considered.

It is a simple and interpretable similarity measure, particularly effective for binary or set-based data. It naturally handles sparse data by ignoring joint absences and is bounded between 0 and 1, facilitating clear interpretation. Additionally, it is robust to dataset size differences, focusing on relative overlap rather than absolute size. However, the Jaccard index has notable limitations. It ignores frequency information, treating all non-zero values equally, which can be problematic when feature prevalence matters. It is also sensitive to rare features, which may disproportionately influence the similarity score. Furthermore, the Jaccard index does not account for the magnitude of overlap, potentially failing to capture nuanced differences between datasets[25–28].

2.4.4. Weighted Jaccard

Weighted Jaccard extends the standard Jaccard index by incorporating weights or adjusting for the prevalence of zeros in sparse datasets. This modification allows

consideration of feature importance or frequency information while maintaining the set-theoretic foundation.

$$J_{\text{weighted}}(x, y) = \frac{\sum_{i=1}^{n} w_i \cdot \min(x_i, y_i)}{\sum_{i=1}^{n} w_i \cdot \max(x_i, y_i)}$$

Where $w_i$represents the weight assigned to feature i. Specifically, if weights are given to account for excesses of zero present in a dataset, it can be considered as a zero-adjusted index

It addresses limitations of standard Jaccard by incorporating magnitude or importance, it's more flexible for datasets with varying feature relevance and maintains interpretability. However, it requires additional step of determining appropriate weights and inappropriate weighting schemes can introduce bias or distort similarity measurements[27,28].

In out tests, we simply ensures that two all-zero rows are treated as identical (similarity = 1) instead of undefined.

2.4.5. Dice Coefficient

The Dice coefficient, also known as the Sørensen-Dice coefficient, is a statistic used to measure the similarity between two samples or sets. It was independently developed by two botanists: Lee Raymond Dice (1945) and Thorvald Sørensen (1948).

The coefficient is defined for two sets $X$and Yas:

$$DSC = \frac{2 \mid X \cap Y \mid}{\mid X \mid + \mid Y \mid}$$

where $\mid X \mid$and $\mid Y \mid$denote the cardinalities (number of elements) of the two sets. The Sørensen index equals twice the number of elements common to both sets divided by the sum of the number of elements in each set, which can equivalently be viewed as the size of the intersection as a fraction of the average size of the two sets.

The Dice coefficient is closely related to the Jaccard index, and the two are mathematically convertible: given a Dice coefficient value S, the corresponding Jaccard index:

$$J = S/(2 - S)$$

However, the Dice coefficient consistently produces higher similarity values than Jaccard for the same data because it gives double weight to the intersection term.

The Sørensen-Dice coefficient is particularly useful for ecological community data and has been widely adopted in this field. Compared to Euclidean distance, the Sørensen distance retains sensitivity in more heterogeneous datasets and gives less weight to outliers. It has also become a preferred metric in image segmentation evaluation, where it is widely used in challenges and benchmarks to assess and rank model performance[29,30].

2.4.6. Kendall correlation

Kendall's τ quantifies the strength of a monotonic association by directly counting how often pairs of observations agree or disagree in their order. Like other two previous methods, it ranges from −1 to +1 — where $\tau = +1$ indicates a perfect concordance among pairs (same order), $\tau = -1$ a perfect discordance (opposite order) and when $\tau = 0$, concordant and discordant pairs balance out, indicating no overall monotonic trend

Kendall's τ is calculated as:

$$\tau(x_i, x_j) = \frac{1}{\binom{n}{2}} \sum_{1 \le k < \ell \le n} \text{sign}(R_{ik} - R_{i\ell})\text{sign}(R_{jk} - R_{j\ell})$$

where $R_{ik}$is the ranks of the k-th observation for variables x_i and x_j and the total number of observation pairs is $\binom{n}{2} = \frac{n(n-1)}{2}$.

In includes greater robustness to outliers than Spearman's correlation. It also provides a direct probabilistic interpretation, reflecting the probability of concordance versus discordance between pairs. Additionally, Kendall's τ has better statistical properties for small sample sizes.

However, it has notable disadvantages. It is computationally expensive for large datasets, with an $O(n^2)$ complexity, and generally has lower statistical power than Spearman's correlation for detecting associations. Furthermore, it is more complex to calculate and interpret compared to simpler methods[31,32].

2.4.7. Mutual Information and Normalized Mutual Information

Unlike traditional distance measures, such as Euclidean or Hamming, which are based on geometric or numerical differences, mutual information quantifies how much information two variables share. It is widely used in the analysis of statistical dependence. In the context of distance estimation, mutual information is used as a measure of similarity between two variables. When used as a distance, we can define it as the inverse of the mutual information:

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

Where *H(X)* is the entropy of *X,* which measures the uncertainty of *X*, *H(Y)* is the entropy of *Y*, *H(X,Y)* is the joint entropy of *X* and *Y.* To transform mutual information into a distance (which is usually a measure of dissimilarity):

$$D_{ij} = 1 - \frac{I(x_i; x_j)}{\max\left(I(x_i; x_j)\right)}$$

where the max operation ensures that the dissimilarity is normalized between 0 and 1[33,34].

Normalized Mutual Information

Normalized Mutual Information (NMI) is a standardized version of mutual information that accounts for the individual entropies of both variables. It normalizes mutual information to range between 0 (no shared information) and 1 (perfect dependency):

$$\text{NMI}(X,Y) = \frac{2 \cdot I(X;Y)}{H(X) + H(Y)}$$

Where $I(X;Y)$is the mutual information between $X$and Y, and $H(X)$and $H(Y)$are their respective entropies.

The NMI captures non-linear relationships that correlation-based methods may miss and is a scale-invariant and symmetric measure. It is particularly useful for categorical or discrete data and provides an information-theoretic interpretation of variable dependency. Additionally, it is bounded between 0 and 1, making it easy to interpret across different datasets.

However, it requires discretization for continuous variables, which can introduce bias. The method is computationally intensive, especially for high-dimensional data, and its performance is sensitive to the choice of binning strategy for continuous data.

Furthermore, it can be difficult to interpret without context about the entropy values involved[35].

2.4.8. Overlap coefficient

Also known as the Szymkiewicz-Simpson index, it represents a specialized similarity measure that quantifies the degree of subset containment between two observations, making it particularly valuable for analyzing relationships in sparse binary or presence-absence data matrices. Unlike the Jaccard coefficient, which normalizes by the union of both sets and thus penalizes differences in set sizes, the Overlap Coefficient normalizes by the smaller of the two sets, effectively measuring what proportion of the smaller set is contained within the larger set. This mathematical property makes it especially suitable for asymmetric relationships where one observation may represent a subset of another's characteristics.

For two observations i and j represented as binary vectors, the *Overlap Coefficient* is formally defined as:

$$Overlap(i,j) = \frac{\min\left(|A_i|, |A_j|\right)}{|A_i \cap A_j|}$$

where:

- $A_i = \{k : x_{ik} > 0\}$ represents the set of active features for observation i
- $A_j = \{k : x_{jk} > 0\}$ represents the set of active features for observation j
- $|A_i \cap A_j|$ is the cardinality of the intersection (shared active features)
- $\min\left(|A_i|, |A_j|\right)$ is the size of the smaller set

Stating that

$$Overlap(i,j) \in [0,1]$$

The interpretation of the results is:

$$\text{Interpretation} = \begin{cases} 1 & \text{Complete subset relationship (smaller set} \subseteq \text{larger set)} \\ 0 & \text{No shared features between observations} \\ (0,1) & \text{Partial overlap proportional to subset containment} \end{cases}$$

It is less sensitive to size differences between sets than Jaccard and it is effective for hierarchical or nested data structures. However, it is a asymmetric measure (depends on

which set is smaller), it can produce high similarity even when sets differ substantially in size, being less discriminative than Jaccard for sets of similar sizes and it may overestimate similarity in imbalanced comparisons[36].

2.4.9. Gram matrix

The Gram matrix is a matrix of inner products between vectors in a dataset. For vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, the Gram matrix Ghas elements

$$G_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. It captures pairwise similarities in the original or transformed feature space.

In a kernel setting, where data are mapped to a (possibly high-dimensional) feature space via $\phi: \mathbb{R}^d \to \mathcal{H}$, the Gram (or kernel) matrix becomes

$$K_{ij} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = k(\mathbf{x}_i, \mathbf{x}_j)$$

where k is a positive semi-definite kernel function.

Among the main advantages of using a Gram matrix, there is its role as the foundation of kernel methods: many algorithms, such as kernel PCA or support vector machines, can be expressed purely in terms of inner products, so replacing $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$with $k(\mathbf{x}_i, \mathbf{x}_j)$allows non-linear transformations without explicitly working in high-dimensional feature spaces. On the other hand, a Gram matrix scales is computational expensive for large sample sizes. Its entries are less directly interpretable than distances such as Euclidean distance, especially when a non-linear kernel is used, because K_{ij}measures an inner product in an implicit feature space rather than a straightforward geometric distance in the original space. Furthermore, without appropriate standardization or normalization[37].

Given the similarities matrix estimated, to identify potential communities among the 29 industrial sites considered, eight community detection approaches were applied: Girvan-Newman, Fast-Greedy (Clauset-Newman-Moore), Louvain, Leiden, Walktrap, and Infomap, Spin-glass and Interacting fluids. A hierarchical clustering model was also implemented to compare the results with a more traditional model.

Community detection analyses were conducted using the *igraph* package in R (version 4.4.1). The following algorithms were applied via specific *igraph* functions: Girvan-ewman (*cluster_edge_betweenness*), Fast-Greedy (*cluster_fast_greedy*), Louvain (*cluster_louvain*), Leiden (*cluster_leiden*), Walktrap (*cluster_walktrap*), Infomap (*cluster_infomap*), Spin-glass (*cluster_spinglass*) and Interacting fluids (*cluster_fluid_communities*). For the latter, since number of communities should be set *a priori*, four different number of communities were indicated (from 2 to 5).

Below is a brief explanation of the community clustering methods adopted.

2.4.10. Girvan and Newman

Girvan and Newman introduced the optimization of modularity and the associated edge-betweenness approach. Rather than assembling communities from their most strongly connected cores, these methods aim to identify and remove inter-community bridges, thereby revealing cohesive groups from the "outside in"[38]. Modularity is a quality function that quantifies how well a given division of a network reflects a community structure. Its core idea is to assess whether the number of edges within proposed communities is significantly higher than would be expected by chance, under a null model that preserves the degree sequence but randomizes connections. Formally, for an undirected network with adjacency matrix A, k_i as the (weighted) degree of i, total edge count m, γ as resolution parameter and community labels c_i, modularity Q of a partition is

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \gamma\frac{k_i k_j}{2m}\right)\partial(c_i, c_j)$$

where $\delta(ci, cj)$ equals 1 if nodes i and j are placed in the same community and 0 otherwise[13]. High values of Q indicate partitions with denser-than-expected intra-community edges and sparser inter-community connections. Optimizing modularity entails searching over partitions to maximize Q. Because the space of partitions grows super-exponentially with network size, exact maximization is computationally infeasible for large graphs. Consequently, a variety of efficient heuristics and approximations have been developed, including greedy agglomerative strategies, spectral methods that relax the problem into an eigenvector formulation, and multi-level schemes such as the Louvain

algorithm. These methods have enabled practical community detection across large real-world networks.

Despite its widespread success, modularity optimization has well-known limitations. Chief among these is the resolution limit: small yet well-formed communities can be merged into larger ones when considered within a large network, leading to under-segmentation. Nevertheless, modularity remains a foundational concept that continues to shape contemporary approaches, inspiring refinements and alternatives that address its shortcomings. Related algorithms and extensions frequently discussed alongside modularity include the Louvain method (multi-level greedy optimization), spectral partitioning approaches, and Kernighan-Lin-type local refinement procedures.

As introduced before, to address the shortcomings of hierarchical clustering, Girvan and Newman propose an alternative approach that focuses on identifying edges most likely to lie between communities rather than those most central within them. The method builds on the concept of betweenness centrality, originally defined for vertices by Freeman[39] as the number of shortest paths between all pairs of other vertices that pass through a given vertex. This measure reflects a node's influence over the flow of information in networks where such flow tends to follow shortest paths. The authors generalise this concept to edges, defining the "edge betweenness" of an edge as the total number of shortest paths between all vertex pairs that traverse it. When multiple shortest paths exist between a pair, each is assigned equal fractional weight so that their contributions sum to one. In networks with loosely connected communities, most shortest paths between different communities must pass through a small set of inter-community edges; these edges therefore exhibit high edge betweenness.

The algorithm proceeds iteratively:

1. Compute the edge betweenness for all edges in the network.
2. Remove the edge with the highest betweenness.
3. Recalculate betweenness values for all edges affected by the removal.
4. Repeat steps 2–3 until no edges remain.

Recalculation at each step is essential: if two communities are connected by multiple edges, not all will initially have high betweenness, but after one is removed, another may become the unique high-betweenness bridge. This dynamic updating ensures that

inter-community links are progressively identified and removed, revealing the underlying community divisions.

From a computational perspective, betweenness for all m edges in a graph of n vertices can be computed in $O(mn)$ time using an efficient algorithm by Newman. Since this computation is repeated after each removal, the worst-case complexity is $O(m2n)$, though performance is often better in practice for networks with strong community structure, as they fragment early in the process[38,40,41].

The output of the algorithm can be represented as a dendrogram, analogous to that produced by hierarchical clustering, but derived from the successive removal of inter-community edges rather than the addition of intra-community ones. This inversion of perspective, focusing on boundaries rather than core, is the key innovation that allows the method to overcome the misclassification issues of traditional approaches.

Subsequently, several improvements to the initial algorithm followed and new approaches were also explored. Some examples will be described below.

2.4.11. Clauset–Newman–Moore (CNM)

Starting from an undirected network with n vertices and m edges, represented by the adjacency matrix A_{ij}, the CNM algorithm approaches community detection as an optimization problem on the modularity function presented earlier. This method is also known as the fast-greedy algorithm due to its use of a greedy agglomerative strategy to maximize modularity efficiently[42].

The algorithm begins with the finest possible partition (each vertex in its own module) and modularity (Q) via the community-community edge fractions

$$e_{rs} = \frac{1}{2m}\sum_{i,j} A_{ij}\partial(c_i, r)\partial(c_j, s)$$

denoting the proportion of edges between communities r and s, and

$$a_r = \frac{1}{2m}\sum_{i} k_i\,\partial(c_i, r)$$

so that

$$Q = \sum_r (e_{rr} - \gamma a_r^2)$$

making explicit the balance between the actual internal edge fraction e_{rr} and its null-model expectation a_r^2. At each agglomerative step, the algorithm selects the pair (r, s) whose fusion yields the largest increment ΔQ, which for initially connected singletons is

$$\Delta Q_{rs} = \frac{1}{2m} - \gamma \frac{k_r k_s}{(2m)^2}$$

and which in subsequent steps is updated according to the merger topology: if a third community t is adjacent to both *i* and *j*, then

$$\Delta Q'_{st} = \Delta Q_{rt} + \Delta Q_{st}$$

if *k* is adjacent only to *i*, then

$$\Delta Q'_{st} = \Delta Q_{rt} - 2\gamma a_s a_t$$

if only to *j*,

$$\Delta Q'_{st} = \Delta Q_{st} - 2\gamma a_r a_t$$

The attachment fraction is simultaneously updated via $a'_s = a_s + a_r$. To ensure scalability, only ΔQ_{rs} values for connected pairs are stored in a sparse structure, each row organized as a balanced tree with an associated row-local max-heap, and the largest entry from each row kept in a global max-heap to permit $O(1)$ access to the best merge candidate and $O(\log n)$ update cost. Iteration continues until a single community remains, producing a dendrogram from which the cut yielding the maximum recorded Q (for the chosen γ) is taken as the optimal partition. For sparse graphs with dendrogram depth $d \sim \log n$ the total complexity $O(m\, d\, \log n)$ reduces to $O(n \log^2 n)$, making the method applicable to very large-scale networks.

Building on the pivot to direct modularity optimization, the CNM greedy agglomeration method grows communities from the bottom up. Starting with each node as its own community, it repeatedly merges the neighbouring pair that yields the largest gain in modularity, producing a dendrogram and identifying the partition with the best observed score. The guiding intuition is simple: link together those pairs whose internal connectivity is most statistically surprising relative to chance. Using sparse bookkeeping

and a priority queue over merge candidates, CNM scales efficiently on large, sparse graphs, delivering a complete community hierarchy and a clear optimal cut. Its main strengths lie in this scalability, conceptual simplicity, reproducibility, and the hierarchical insight it provides. Its limitations stem from its greedy nature (early choices can lock in suboptimal partitions) and from modularity's resolution limit, which can cause small communities to be absorbed into larger ones; performance can also degrade on very dense networks. In practice, these drawbacks are often mitigated by multilevel coarsening, sparse ΔQ updates, and a final local node-move refinement to escape "merge myopia" while preserving speed.

2.4.12. Louvain

In 2008, Blondel et al. presented a multilevel heuristic method that optimizes modularity through local node moves and graph aggregation, trading global searches for fast, iterative improvements[43]. The so-called Louvain method seeks a high-modularity partition by alternating two phases until no further gain is possible.

The approach begins with each node in its own community and performs greedy local moves: for each node i, it evaluates the modularity gain from moving i into each neighbouring community C. With weighted degree $k_i = \sum_j A_{ij}$, community strength $\sum_{tot}(C) = \sum_{u \in C} k_u$ and $k_{i,\text{in}}(C) = \sum_{j \in C} A_{ij}$, the change in modularity for moving i from its current community into C can be computed from local aggregates as

$$\Delta Q(i \to C) = \frac{1}{2m}\left[\left(\sum\nolimits_{\text{in}}(C) + 2k_{i,\text{in}}(C)\right) - \gamma\frac{(\sum_{\text{tot}}(C) + k_i)^2}{2m}\right] - \frac{1}{2m}\left[\sum\nolimits_{\text{in}}(C) - \gamma\frac{(\sum_{\text{tot}}(C))^2}{2m}\right]$$

where

$$\sum\nolimits_{\text{in}}(C) = \sum\nolimits_{u,v \in C} A_{uv}$$

is twice the internal edge weight of C. For selection purposes, this simplifies to maximizing over neighbouring communities the local score

$$\arg \max_{C \ni nmbr\ of\ i} \left\{k_{i,\text{in}}(C) - \gamma\frac{k_i \sum_{tot}(C)}{2m}\right\}$$

since terms independent of C cancel when comparing moves. Nodes are swept (in random or fixed order) and moved whenever $\Delta Q > 0$ until no single move improves Q (end of phase one). Phase two coarsens the graph by contracting each community into a super-node; edges between super-nodes carry the sum of inter-community weights, and self-loops preserve intra-community weight. The two phases repeat on this reduced graph, building a hierarchy, and the best observed partition across levels is retained. *Louvain* is fast in practice (near-linear on sparse graphs) because each node move uses only local quantities, and the number of levels stays modest. Its main strengths are speed and scalability, a clear multilevel hierarchy, and good empirical modularity with minimal bookkeeping; its limitations are those of modularity (resolution limit, many near-degenerate optima) and the heuristic itself (results can vary across runs, and communities may be poorly connected internally). Common refinements include multiple restarts with consensus clustering, modest post-processing (e.g., Kernighan–Lin-style local swaps), tuning γ for multiscale structure, or adopting Leiden as a drop-in replacement to guarantee well-connected communities while retaining Louvain's efficiency.

2.4.13. Leiden

Originally proposed by Traag et al. (2019)[44], the Leiden algorithm builds on the multilevel strategy of Louvain while directly addressing one of its key shortcomings: the possibility of obtaining communities that are internally disconnected. By introducing an intermediate refinement phase, Leiden guarantees that each community is well-connected before it is aggregated to a higher level, thus improving both the robustness and interpretability of the detected structure.

The algorithm proceeds iteratively in three phases. First of all, a Local Moving Phase. Each node i is evaluated for potential movement to a neighbouring community. A move is accepted if it increases the value of a chosen quality function Q, most often *modularity*. The gain in modularity when moving i from its current community C to a target community C' is computed as:

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

where m is the total edge weight in the network, $\sum_{in}$ is the sum of weights of edges inside C', $\sum_{tot}$ is the sum of weights of edges incident to C', k_i is the degree of node i, $k_{i,in}$ is the sum of weights of edges from i to nodes in C'.

Secondly, it follows a Refinement Phase. Communities obtained in the first phase are internally partitioned again with the same local moving process. This step splits off disconnected or weakly connected subgroups, ensuring that all remaining nodes in a community are part of a single connected component.

Finally, the Aggregation Phase. The refined communities are collapsed into super-nodes. Edges between super-nodes are weighted sums of the original edges. The algorithm then repeats from Phase 1 on this reduced network until no modularity gain is possible.

Leiden can also optimise alternative quality functions such as the Constant Potts Model (CPM) for resolution-parameter control. By design, Leiden's refinement step ensures γ-connectedness, preventing the "disconnected community" artefacts sometimes produced by Louvain. Despite the extra phase, the algorithm remains fast, often matching or exceeding Louvain's runtime on large sparse graphs and, finally, it could be applicable to weighted, unweighted, directed, and undirected networks.

While modularity-based methods have proven highly influential and remain among the most widely used techniques for community detection, they are not without limitations. Chief among these is the resolution limit, which can cause smaller communities to be merged into larger ones, even when they are structurally distinct. Additionally, the optimization of modularity is an NP-hard problem, often requiring heuristic or approximate solutions that may converge to local optima. In response to these challenges, a range of alternative approaches has been developed that do not rely on modularity as an objective function. These include methods based on information theory, random walks, and local label dynamics, offering complementary perspectives on what constitutes a meaningful community in a network[45].

2.4.14. Infomap

Originally proposed by Rosvall and Bergstrom (2008)[46], the Infomap algorithm approaches community detection from an information-theoretic perspective, modelling the movement of a random walker on the network and seeking a partition that minimises

the average description length of its trajectory. The underlying idea is that a walker tends to spend long stretches within dense groups of nodes before crossing to another part of the graph; by assigning short, reusable codewords for movements within the same module and separate codes for inter-module jumps, one can compress the description of the walk. The optimal community structure is found by minimising the map equation

$$L(M) = q_{\curvearrowright} H(Q) + \sum_{i=1}^{m} p_{\circlearrowright}^{i} H\left(\mathcal{P}^{i}\right)$$

where M is a given partition into mm modules, $q_{\curvearrowright}$ is the probability of the random walk exiting a module, $H(Q)$ is the Shannon entropy of the codebook used for inter-module moves, $p_{\circlearrowright}^{i}$is the probability of being in module i (including the exit step), and $H\left(\mathcal{P}^{i}\right)$ is the entropy of the codebook for moves within module i. By construction, this formulation rewards partitions where the walker is "trapped" longer within modules, as repeated use of short codewords reduces the total description length. Infomap can be applied to weighted, unweighted, directed, or undirected graphs, and its optimisation is performed through a greedy search with possible multilevel refinements, yielding communities that are often free from the resolution limit that affects modularity-based methods.

2.4.15. Walktrap

Walktrap is a community detection algorithm based on the principle that short random walks are more likely to stay within the same community[47].

Let $G = (V, E)$ be the graph, possibly weighted, with $n = \mid V \mid$ nodes. For each node i, we consider a short random walk of length t, described by the t-step transition probability vector $p_i^{(t)}$. The distance between two nodes i and j after t-steps is defined as

$$r_{ij} = \sqrt{\sum_{k=1}^{n} \frac{\left(p_{ik}^{(t)} - p_{jk}^{(t)}\right)^2}{d(k)}}$$

where $d(k)$ is the degree of node k (or the sum of its edge weights).

Initially, each node is its own community. Communities are merged iteratively in a bottom-up (agglomerative) manner, at each step choosing the merge that results in the smallest increase in average distance. This process builds a dendrogram from which an

optimal partition is chosen, typically by maximizing modularity Q. Walktrap outputs a hierarchical community structure, allowing the resolution to be adjusted by "cutting" the dendrogram at different heights. Its strengths lie in the intuitive nature of its underlying principle, the flexibility to handle weighted or unweighted graphs, and the hierarchical structure it returns, allowing to examine communities at multiple resolutions without rerunning the method. This makes it particularly effective for medium-sized networks where subtle boundaries matter. However, its iterative merging process and reliance on distance computations render it more computationally demanding than faster heuristics such as Louvain or Leiden, especially for very large datasets. Its performance can also depend on the choice of walk length, with values too short missing structure and those too long blurring community boundaries. Like many modularity-based approaches, it may struggle with very small communities embedded in large graphs, and memory usage can become an issue for large-scale applications.

2.4.16. Spin-Glass

In this physical analogy, every vertex i in the graph possesses a spin state $\sigma_i \in \{1, \dots, q\}$, which represents the community to which it belongs. Just as magnetic spins in a material interact with their neighbors to minimize the system's overall energy, nodes in a network "interact" through their connections to find the most stable community structure[48].

The theoretical foundation rests on the q-state Potts model, a generalization of the Ising model from condensed matter physics[49]. The algorithm's central quest is to discover the spin configuration $\{\sigma_i\}$ that minimizes the system's Hamiltonian (energy function):

$$H = -\sum_i <_j J_{ij}\, \delta(\sigma_i, \sigma_j) + \gamma \sum_i <_j P_{ij}\, \delta(\sigma_i, \sigma_j)$$

The interaction strength J_{ij} represents the observed connection between nodes i and j often simply the adjacency matrix entry A_{ij}, though it can accommodate weighted networks. The expected interaction P_{ij}, under a "null model" provides a baseline against which to measure the significance of observed connections. This baseline can take different forms: the simple uniform model

$$P_{ij} = \frac{2m}{[n(n-1)]}$$

assumes all node pairs are equally likely to connect, while the more sophisticated "configuration model"

$$P_{ij} = \frac{k_i k_j}{2m}$$

preserves each node's degree while randomizing connections.

The Kronecker delta function $\delta(\sigma_i, \sigma_j)$ acts as a community indicator, equaling 1 when nodes i and j share the same spin state (belong to the same community) and 0 otherwise. The resolution parameter γ serves as a tuning knob that controls the granularity of community detection, much like adjusting the magnification on a microscope to reveal different scales of structure[49,50].

The energy function tells a compelling physical story. The first term, with its negative sign, rewards internal community connections (when two connected nodes share the same spin state, they contribute negatively to the total energy, making the configuration more stable). Conversely, the second term acts as a penalty mechanism, discouraging communities that appear denser than what the null model would predict. The resolution parameter γ modulates this tension: values greater than 1 make the penalty more severe, favouring smaller, tighter communities, while $\gamma < 1$ relaxes the constraint, allowing larger communities to form.

Finding the optimal spin configuration — the ground state that minimizes H — presents a formidable computational challenge. The algorithm employs simulated annealing, a probabilistic optimization technique that mimics the physical process of slowly cooling a material to achieve its most stable crystalline structure. This process unfolds through several stages:

The journey begins with initialization, where spins are assigned randomly across the network, creating a high-energy, disordered state analogous to a molten material. The algorithm then enters an iterative phase of spin updates, where it randomly proposes changes to individual node assignments. These proposals are accepted or rejected based on their impact on the total energy, following the Boltzmann probability distribution:

$$p = \begin{cases} 1 & if\ \Delta H \leq 0 \\ e^{\left(-\frac{\Delta H}{T}\right)} & if\ \Delta H > 0 \end{cases}$$

Here, T represents the current "temperature" of the system. Energy-reducing moves ($\Delta H \leq 0$) are always accepted, as they lead the system toward greater stability. Energy-increasing moves face probabilistic acceptance that decreases exponentially with both the energy penalty ΔH and the inverse temperature. This mechanism allows the algorithm to escape local minima early in the process when temperatures are high, while gradually becoming more selective as the system cools.

The cooling schedule forms the algorithm's backbone, progressively reducing T according to a predetermined cooling factor. This gradual temperature reduction mirrors the controlled cooling of metals in metallurgy — too fast, and the system becomes trapped in suboptimal local configurations; too slow, and computational resources are wasted on marginal improvements.

As the algorithm converges toward its final state, a natural community structure emerges from the energy landscape. The method produces a partition into at most q communities, where q is specified by the spins parameter, though the actual number typically falls below this maximum as the optimization process naturally eliminates unnecessary community distinctions. This organic emergence of community count represents a key advantage over methods that require pre-specification of the number of clusters.

When the configuration model serves as the null model and $\gamma = 1$, a mathematical relationship emerges: the Hamiltonian becomes directly related to generalized modularity, bridging the gap between physics-inspired and network-theoretic approaches. The optimization framework remains fundamentally physics-driven, leveraging the rich theoretical apparatus of statistical mechanics to navigate the complex landscape of possible community structures.

The algorithm's strength lies in its ability to naturally balance local connectivity patterns with global network properties, guided by the physical intuition that stable configurations minimize overall system energy while respecting the constraints imposed by the network's structure.

2.4.17. Fluid Community Detection

Developed by Parés et al. (2017)[51], this approach represents an innovative paradigm in network analysis that leverages the intuitive analogy of fluid dynamics to discover

community structures through iterative propagation processes**.** Unlike traditional optimization-based approaches, the algorithm operates by treating each community as a distinct fluid that flows through the network topology, with nodes adopting the community identity of their most influential neighbors at each iteration. The mathematical framework encompasses several key components: initialization strategies that can accommodate both random and seed-based approaches, propagation dynamics that govern how community assignments evolve over time, convergence criteria that determine algorithmic termination, and quality measures that assess the resulting community structure.

The algorithm's strength lies in its simplicity and computational efficiency, requiring only $O(|E| + |V|k)$ operations per iteration, making it particularly suitable for large-scale networks where traditional modularity optimization methods become computationally prohibitive. The tie-breaking mechanisms ensure deterministic behaviour even in ambiguous scenarios, while the various algorithm variants (synchronous *vs* asynchronous updates, multi-resolution approaches) provide flexibility for different network characteristics and computational constraints.

The mathematical foundations also include comprehensive quality assessment metrics, parameter selection guidelines, and comparative measures that enable systematic evaluation against other community detection methods. This makes Fluid Community.

The algorithm begins by choosing one random seed for each of the k target communities and then “coloring” every node according to its closest seed. Concretely, let

$$S_i \subset V, \quad |S_i| = 1,\ i = 1, \dots, k$$

be the k seed nodes (one per community). Them the algorithm define the initial labelling

$$F_0 : V \to \{1, \dots, k\}$$

by the piecewise rule

$$F_0(v) = \begin{cases} i & \text{if } v \in S_i \text{ (seed set for community } i) \\ argmaxj \sum u \in S_i \dfrac{1}{d(v,u)+1} & \text{if } v \notin \bigcup_{i=1}^{k} S_i \end{cases}$$

Here $d(v,u)$ is the graph-distance between v and u, so each non-seed node picks the community whose seed is closest in the inverse-distance sense.

Equivalently, one can view this as assigning every node a community label uniformly at random:

$$F_0(v) = \text{Uniform}(\{1,2,\dots,k\}), \quad \forall v \in V$$

which has the same effect of seeding each community once and then roughly propagating labels by proximity.

After the random initialization, the algorithm proceeds through a iterative label propagation. Within each iteration, it compute for each community label $i \in \{1,\dots,k\}$ the "support"

$$\text{support}_i(\text{v}) \;=\; \sum_{\text{u}\in \text{N(v)}} w(v,u)\,\mathbf{1}\{F_t(u) = i\}$$

for every node $v \in V$, where $N(v)$ is the neighbourhood of v, $w(v,u)$ is the edge-weight (or 1 if unweighted), and $1\{\cdot\}$ is the indicator.

It then update the label of v by picking the community with maximum support, breaking ties uniformly at random:

$$F_{t+1}(v) \;=\; \arg\max_{1\le i\le k} \text{support}_\text{i}(v)$$

It then repeats this step until labels stabilize (i.e.$F_{t+1}(v) = F_t(v)$ for all v) or until reaching a preset maximum iteration T.

Each synchronous pass over all nodes costs $O(\sum_v \deg(v)) = O(m)$ time, where m is the number of edges. In practice the method converges in a small number of rounds (often < 20). Convergence is not guaranteed to the global optimum of any particular objective, but it typically yields high-quality, locally coherent communities.

Its main strengths include scalability to massive graphs, minimal parameter tuning, rapid convergence, flexible extensions, robustness of solutions.

As a methodological benchmark against the network-based community detection approaches reviewed above, we implemented hierarchical agglomerative clustering using R's *hclust* function from the *stats* package, systematically evaluating four linkage criteria across the nine similarity matrices (Jaccard, cosine, Kendall's τ, normalized Euclidean, overlap coefficient, normalized mutual information, weighted Jaccard, Gram matrix, and Dice coefficient.

While community detection leverages network topology to identify densely connected modules via modularity optimization or random walks, hierarchical clustering operates directly on feature similarity space, constructing dendrograms by iteratively merging clusters according to linkage-specific distance criteria[52,53]. The similarity matrices were converted to dissimilarity matrices via inversion (1 - similarity), providing a non-topological baseline that reveals phenotypic hierarchies without structural assumptions and enabling evaluation of whether network representation enhances grouping beyond raw similarity[54].

We employed four linkage methods:

- Complete linkage: maximum inter-cluster distance, yields compact clusters;

- Average linkage (UPGMA): mean pairwise distances;

- Single linkage: minimum inter-cluster distance ("friends of friends");

- Ward.D2: minimum within-cluster variance via sum-of-squares.

For each of the 36 similarity-linkage combinations, the optimal number of clusters $k \in [2,10]$ was determined via maximum average silhouette width[55], yielding silhouette-optimized partitions. Newman-Girvan modularity Q was then computed for all hierarchical solutions using the original similarity-derived *igraph* objects, enabling unified quantitative comparison against the 10 community detection methods.

For all model combinations, the corresponding modularity values were reported, with 0.3 considered the optimal reference threshold[56]. Graphical representations were provided only for the best-performing community detection model and for the hierarchical clustering model.

2.5. Results

For 2022, the dataset was highly sparse, with 80.9% of participation values for individual diagnostic tests across sites equal to zero (missing values are treated as zero participation). Additionally, around 1.4% of the cells contained participation rates exceeding 100%.

After removing diagnostic tests with zero participation across all sites, the resulting dataset included 77 diagnostic tests. Among all combinations, 31 models returns a

modularity value higher or ≈ 0 (Figure 1). Figure 2 shows the models with a modularity value lower than 0. The first six best models are presented in Figures 3-8.

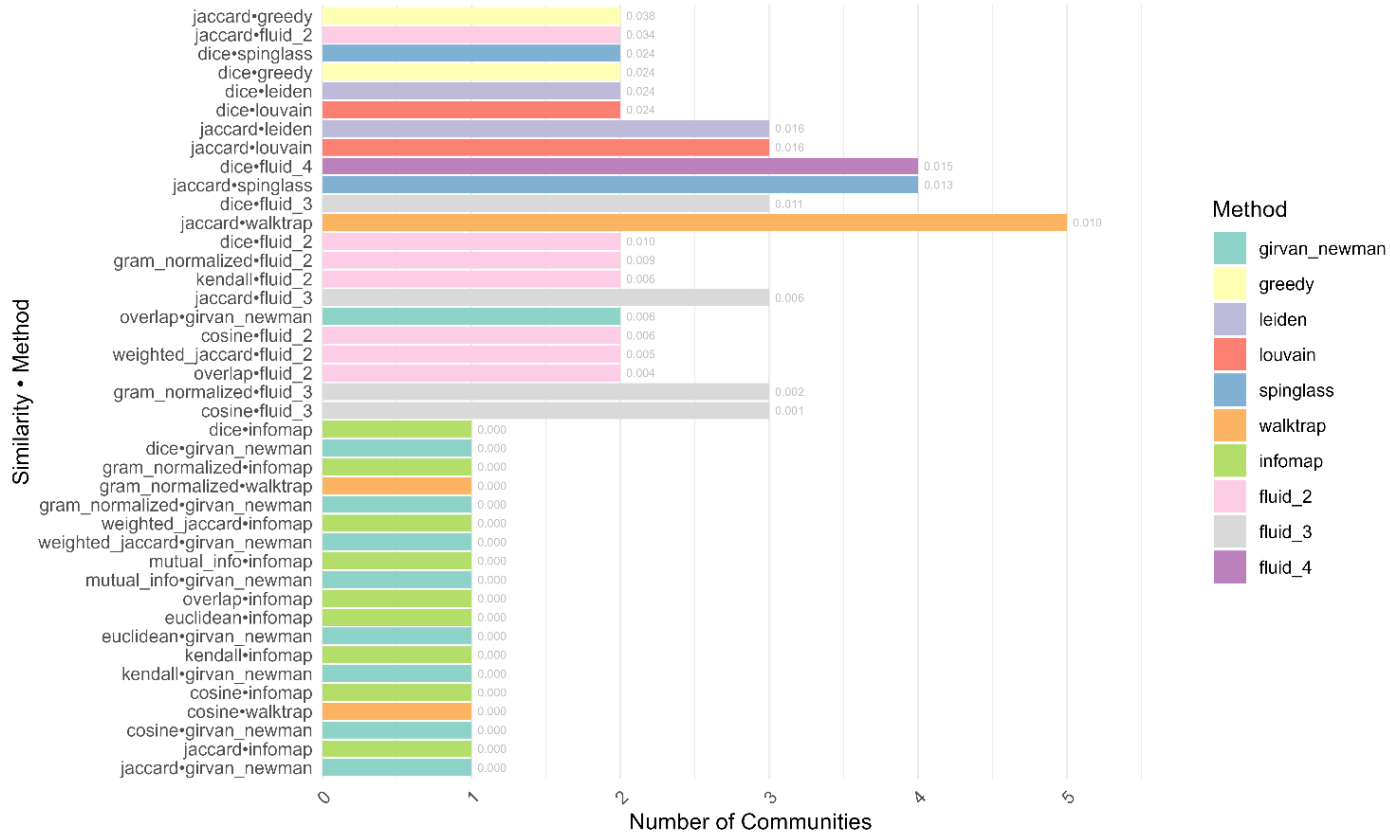


Figure 1. Number of communities detected by each community-detection model and similarity-matrix combination for the year 2022. Results is relative to models with a modularity value ≥ 0. Bars are sorted top-to-bottom in descending order of modularity. The x-axis denotes the number of communities identified.

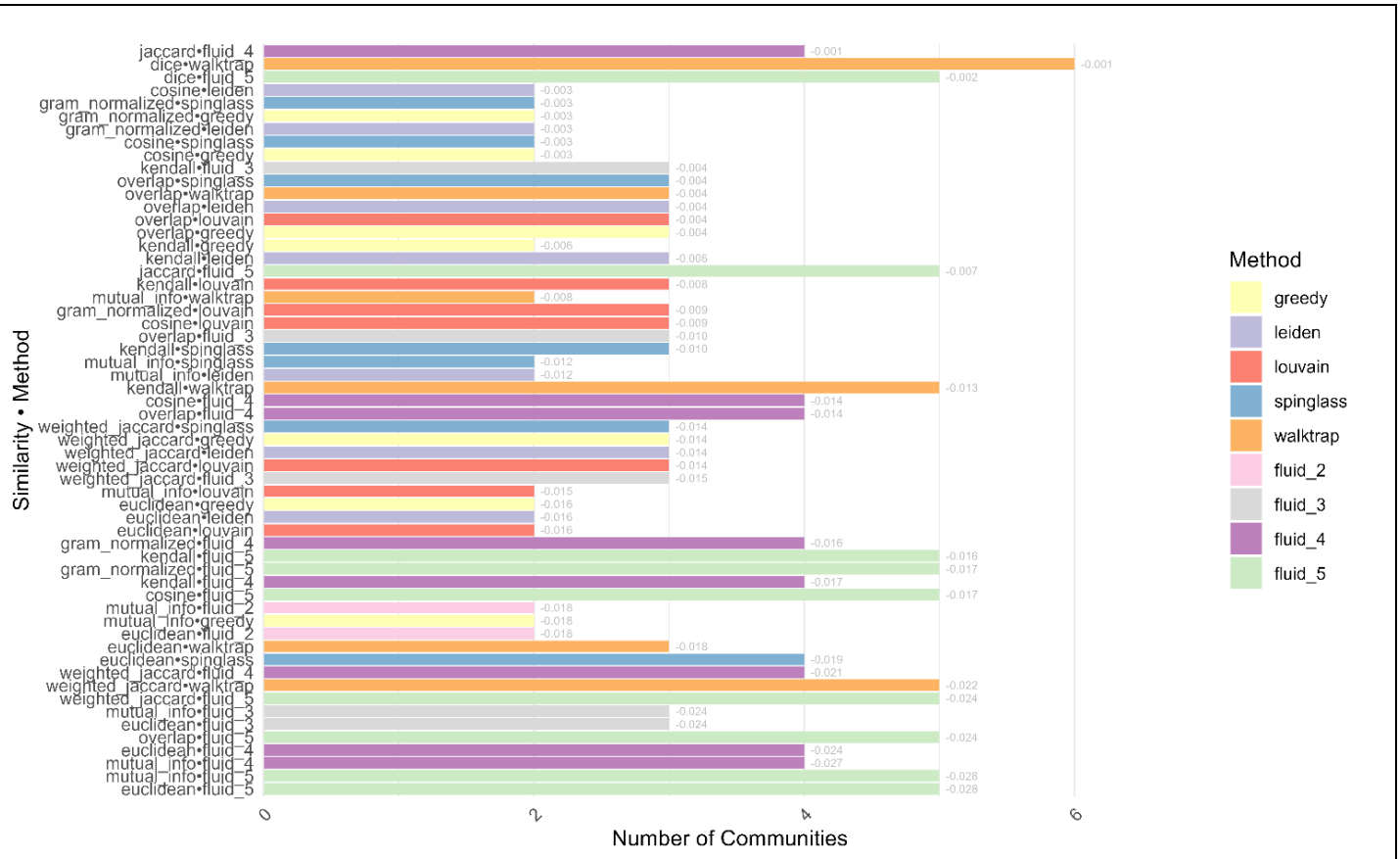
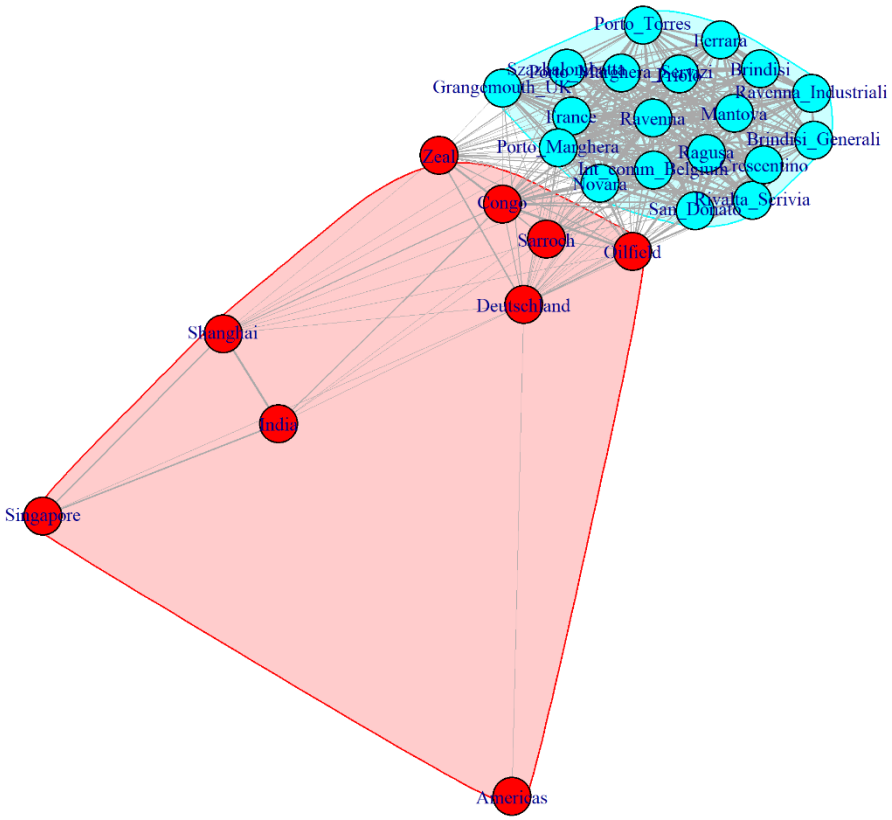


Figure 2. Number of communities detected by each community-detection model and similarity-matrix combination for the year 2022. Results is relative to models with a modularity value < 0. Bars are sorted top-to-bottom in descending order of modularity. The x-axis denotes the number of communities identified.

Two of the best models (fig. 3-4) excluded the "México" site, as it had no connections with other nodes and was too isolated to belong to any of the identified communities.

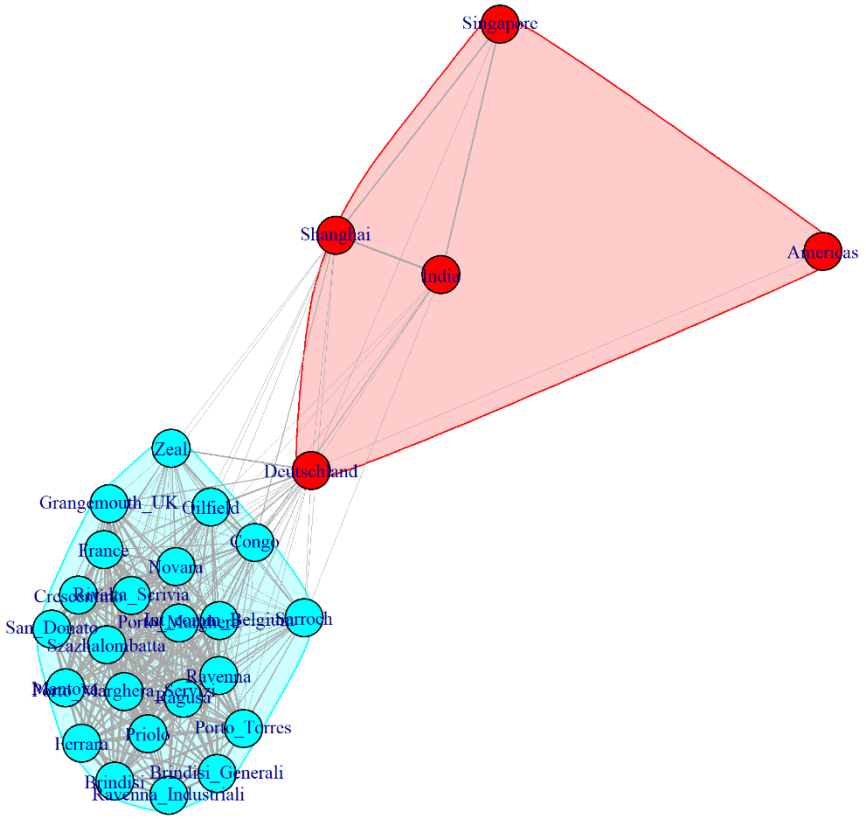
Jaccard matrix – Greedy community detection model



[1]
"Americas"; "Congo"; "Deutschland"; "India"; "Shanghai"; "Singapore"; "Oilfield"; "Sarroch"; "Zeal"))
[2]
"Brindisi_Generali"; "Szazhalombatta"; "Ravenna_Industriali"; "Porto_Marghera_Servizi"; "France"; "Int_comm_Belgium"; "Novara"; "Rivalta_Scrivia"; "San_Donato"; "Brindisi"; "Crescentino"; "Ferrara"; "Mantova"; "Porto_Marghera"; "Porto_Torres"; "Priolo"; "Ragusa"; "Ravenna"; "Grangemouth_UK"))

Figure 3. Network of sites clustered into two communities by the *greedy* algorithm on the *Jaccard* similarity matrix. Node colors reflect community membership; edge widths are proportional to *Jaccard* similarity.

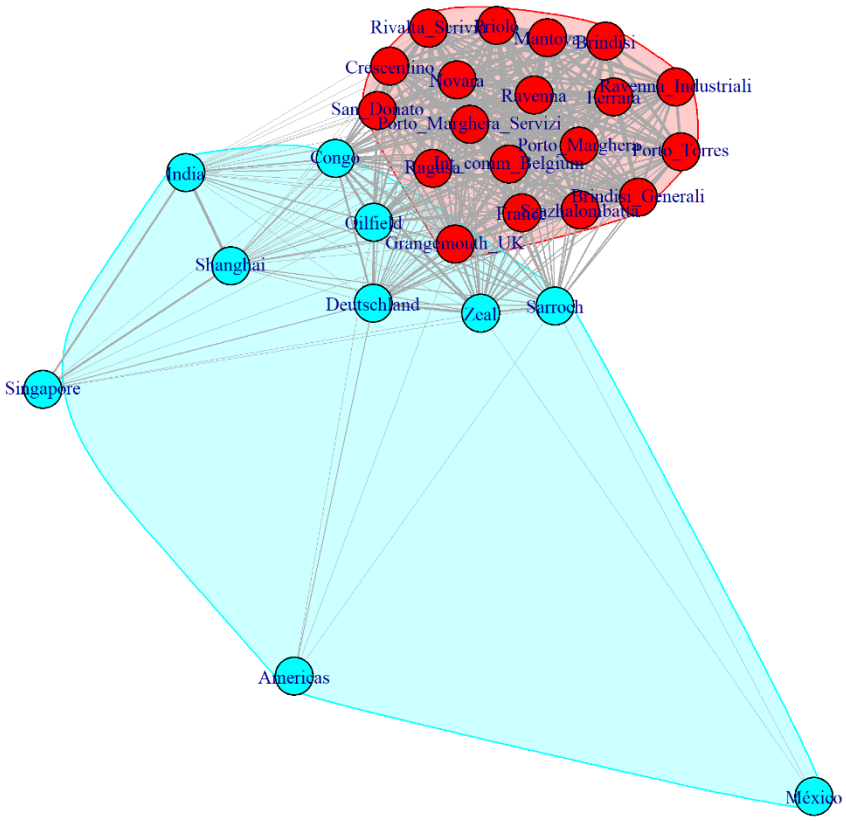
Jaccard matrix – Fluid community detection model



[1]
"Americas"; "Deutschland"; "India"; "Shanghai"; "Singapore"
[2]
"Brindisi_Generali"; "Szazhalombatta"; "Ravenna_Industriali"; "Porto_Marghera_Servizi"; "Congo"; "France"; "Int_comm_Belgium"; "Oilfield"; "Novara"; "Rivalta_Scrivia"; "San_Donato"; "Brindisi"; "Crescentino"; "Ferrara"; "Mantova"; "Porto_Marghera"; "Porto_Torres"; "Priolo"; "Ragusa"; "Ravenna"; "Sarroch"; "Grangemouth_UK"; "Zeal"

Figure 4. Network of sites clustered into two communities by the *fluid* algorithm on the *Jaccard* similarity matrix. Node colors reflect community membership; edge widths are proportional to *Jaccard* similarity.

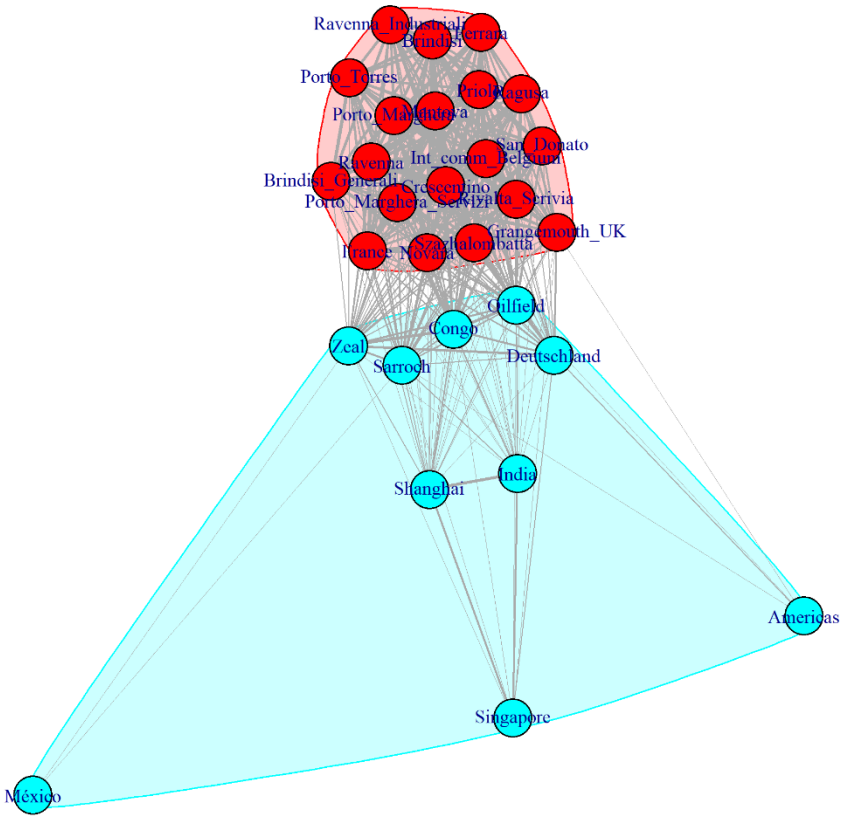
Dice matrix – Louvain community detection model



[1]
"Brindisi_Generali", "Szazhalombatta", "Ravenna_Industriali", "Porto_Marghera_Servizi", "France", "Int_comm_Belgium", "Novara", "Rivalta_Scrivia", "San_Donato", "Brindisi", "Crescentino", "Ferrara", "Mantova", "Porto_Marghera", "Porto_Torres", "Priolo", "Ragusa", "Ravenna", "Grangemouth_UK"
[2]
"Americas", "Congo", "Deutschland", "México", "India", "Shanghai", "Singapore", "Oilfield", "Sarroch", "Zeal"

Figure 5. Network of sites clustered into two communities by the *Louvain* algorithm on the *Dice* similarity matrix. Node colors reflect community membership; edge widths are proportional to *Dice* similarity.

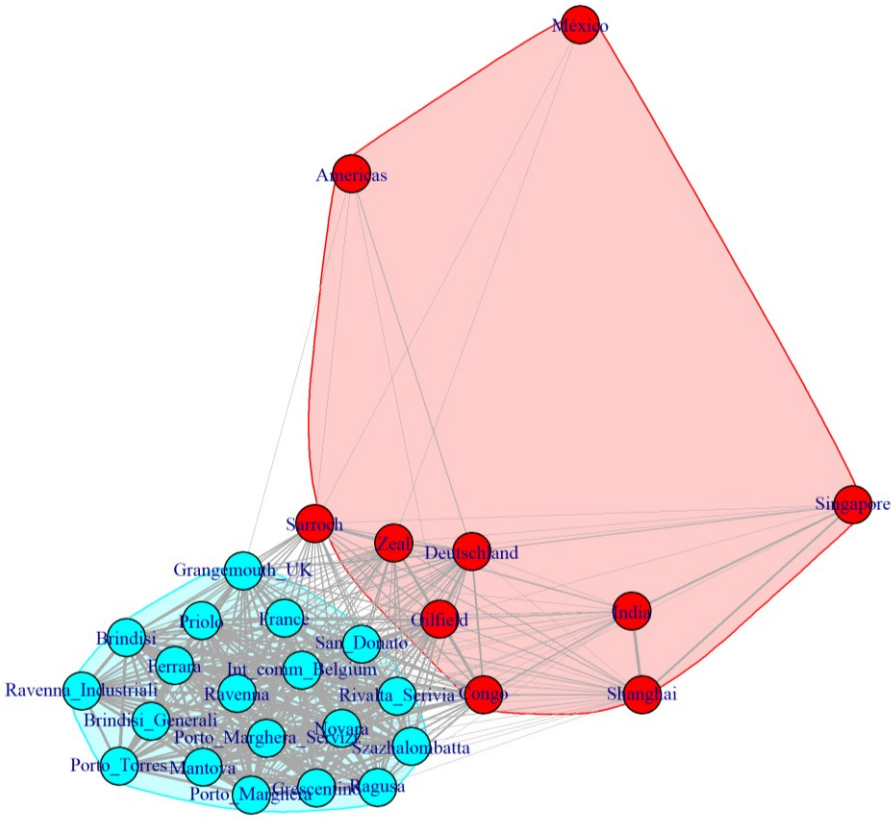
Dice matrix – Leiden community detection model



[1]
"Brindisi_Generali", "Szazhalombatta", "Ravenna_Industriali", "Porto_Marghera_Servizi", "France", "Int_comm_Belgium", "Novara", "Rivalta_Scrivia", "San_Donato", "Brindisi", "Crescentino", "Ferrara", "Mantova", "Porto_Marghera", "Porto_Torres", "Priolo", "Ragusa", "Ravenna", "Grangemouth_UK"
[2]
"Americas", "Congo", "Deutschland", "México", "India", "Shanghai", "Singapore", "Oilfield", "Sarroch", "Zeal"

Figure 6. Network of sites clustered into two communities by the *Leiden* algorithm on the *Dice* similarity matrix. Node colors reflect community membership; edge widths are proportional to *Dice* similarity.

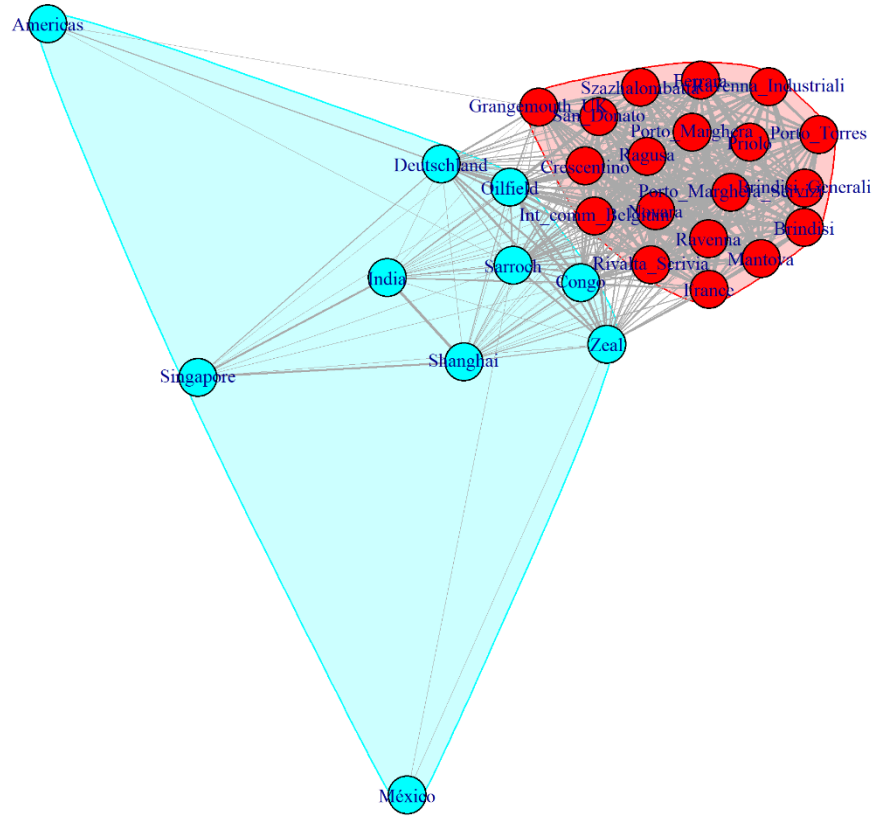
Dice matrix – Greedy community detection model



[1]
"Americas", "Congo", "Deutschland", "México", "India", "Shanghai", "Singapore", "Oilfield", "Sarroch", "Zeal
[2]
Brindisi_Generali", "Szazhalombatta", "Ravenna_Industriali", "Porto_Marghera_Servizi", "France", "Int_comm_Belgium", "Novara", "Rivalta_Scrivia", "San_Donato", "Brindisi", "Crescentino", "Ferrara", "Mantova", "Porto_Marghera", "Porto_Torres", "Priolo", "Ragusa", "Ravenna", "Grangemouth_UK"

Figure 7. Network of sites clustered into two communities by the *greedy* algorithm on the *Dice* similarity matrix. Node colors reflect community membership; edge widths are proportional to *Dice* similarity.

Dice matrix – Spinglass community detection model



[1]
"Brindisi_Generali", "Szazhalombatta", "Ravenna_Industriali", "Porto_Marghera_Servizi", "France", Int_comm_Belgium, "Novara", "Rivalta_Scrivia", "San_Donato", "Brindisi", "Crescentino", "Ferrara", "Mantova", "Porto_Marghera", "Porto_Torres", "Priolo", "Ragusa", "Ravenna", "Grangemouth_UK"
[2]
"Americas", "Congo", "Deutschland", "México", "India", "Shanghai", "Singapore", "Oilfield", "Sarroch", "Zeal"

Figure 8. Network of sites clustered into two communities by the *spinglass* algorithm on the *Dice* similarity matrix. Node colors reflect community membership; edge widths are proportional to *Dice* similarity.

The performance obtained using the dissimilarity matrices in combination with the various arbitrarily selected hierarchical clustering algorithms is reported below (Fig.9-10). Overall, these methods performed worse than the community-detection algorithms. However, the lowest performance scores of the hierarchical clustering models were not as poor as the lowest scores observed among the community-detection models. Since the best hierarchical clustering model exhibited a lower modularity value than the top six

community-detection models, we present the dendrogram corresponding to the best-performing hierarchical clustering model (Fig. 11).

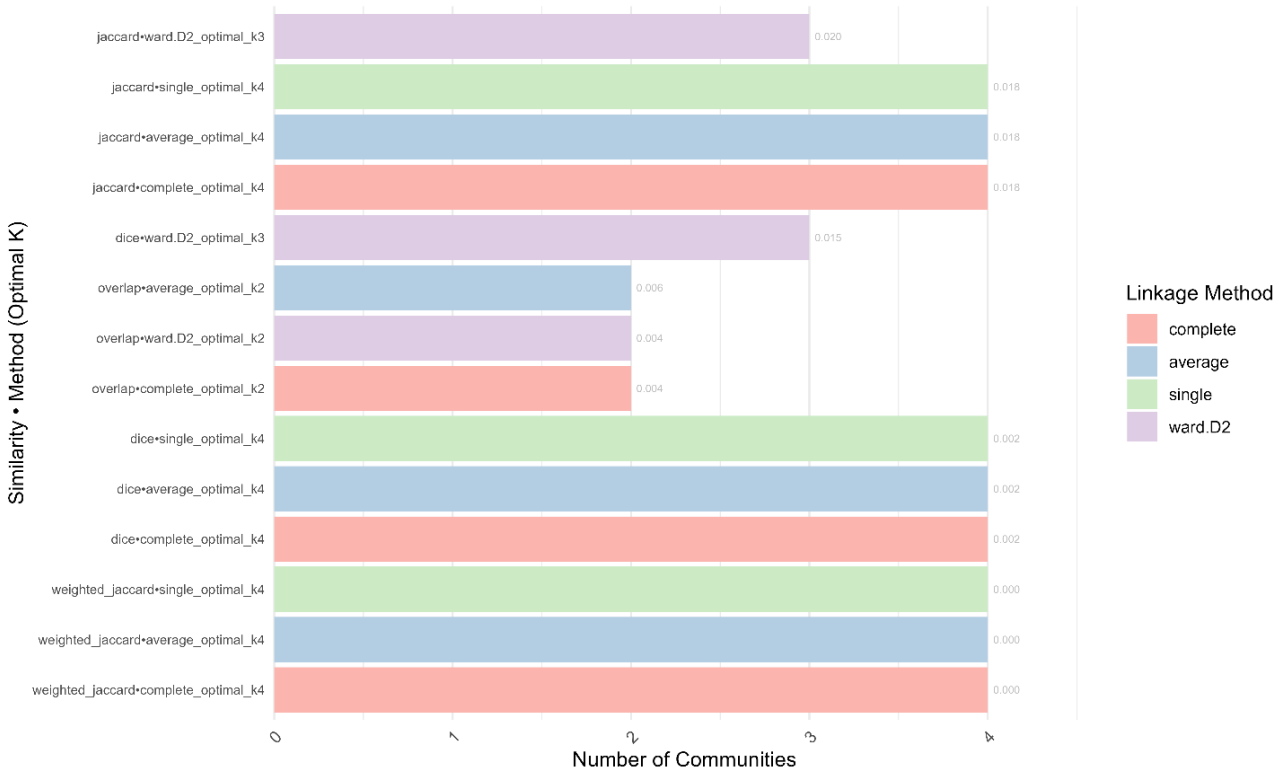


Figure 9. Number of communities detected by each hierarchical model and similarity-matrix combination for the year 2022. Results is relative to models with a modularity value ≥ 0. Bars are sorted top-to-bottom in descending order of modularity. The x-axis denotes the number of communities identified.

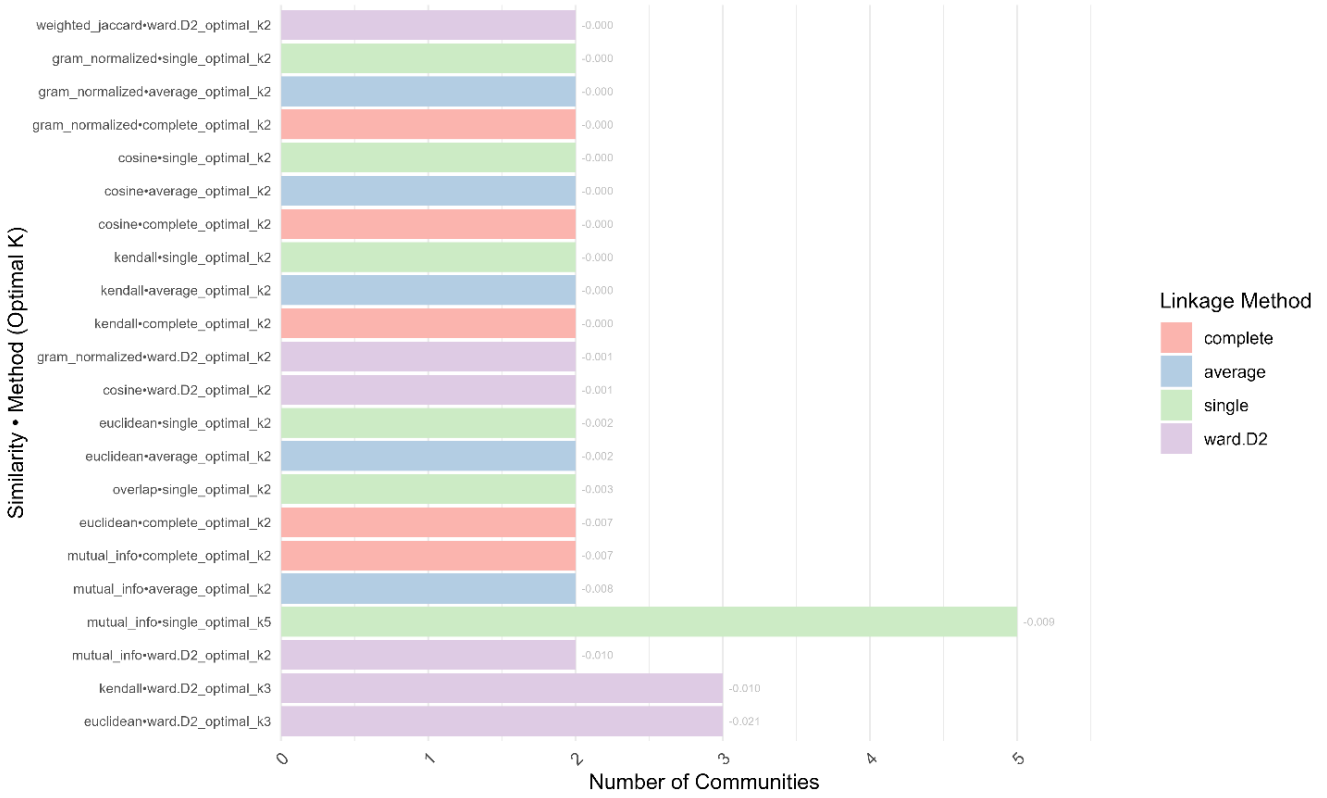


Figure 10. Number of communities detected by each hierarchical model and similarity-matrix combination for the year 2022. Results is relative to models with a modularity value < 0. Bars are sorted top-to-bottom in descending order of modularity. The x-axis denotes the number of communities identified.

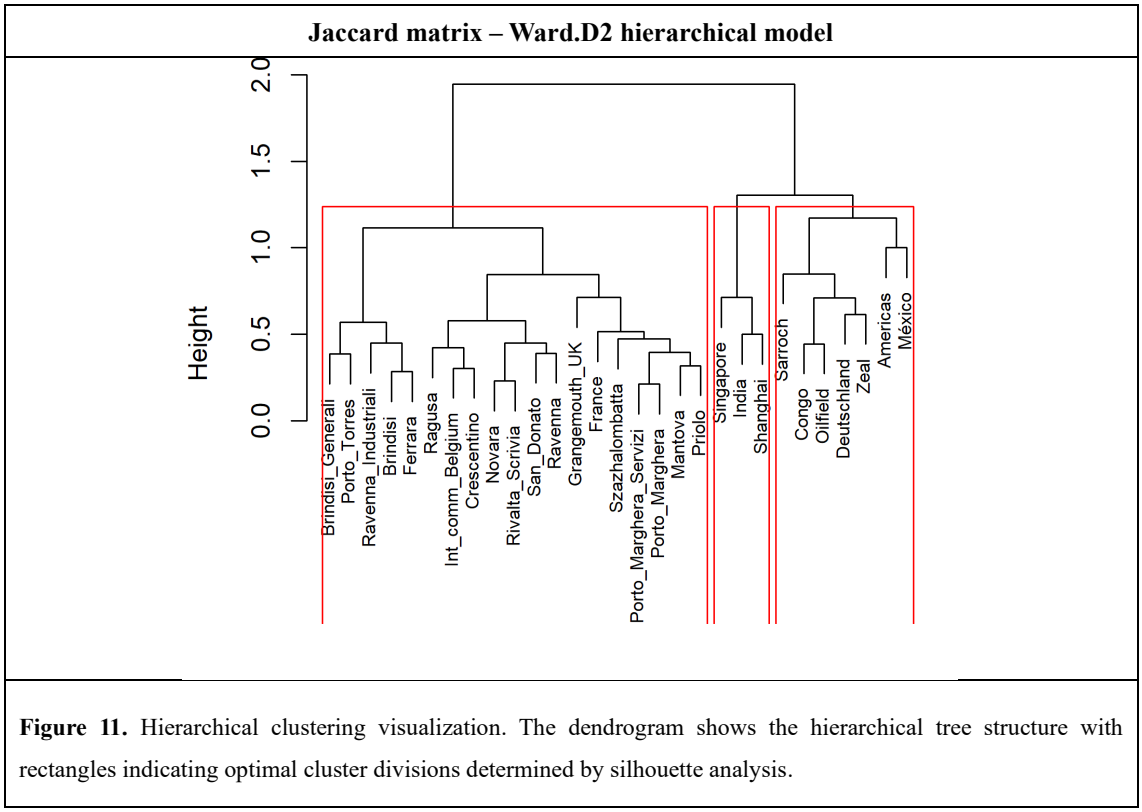


Figure 11. Hierarchical clustering visualization. The dendrogram shows the hierarchical tree structure with rectangles indicating optimal cluster divisions determined by silhouette analysis.

For 2023, null participation accounted for 82.1% of the dataset, and approximately 1.5% of cells had participation values exceeding 100%. The number of diagnostic tests with at least one participation across all company sites was 75 ─ (two fewer than the previous year). Among all combinations, 20 models returns a modularity value higher or ≈ 0 (Figure 12). The first six best models are presented in Figures 13-17.

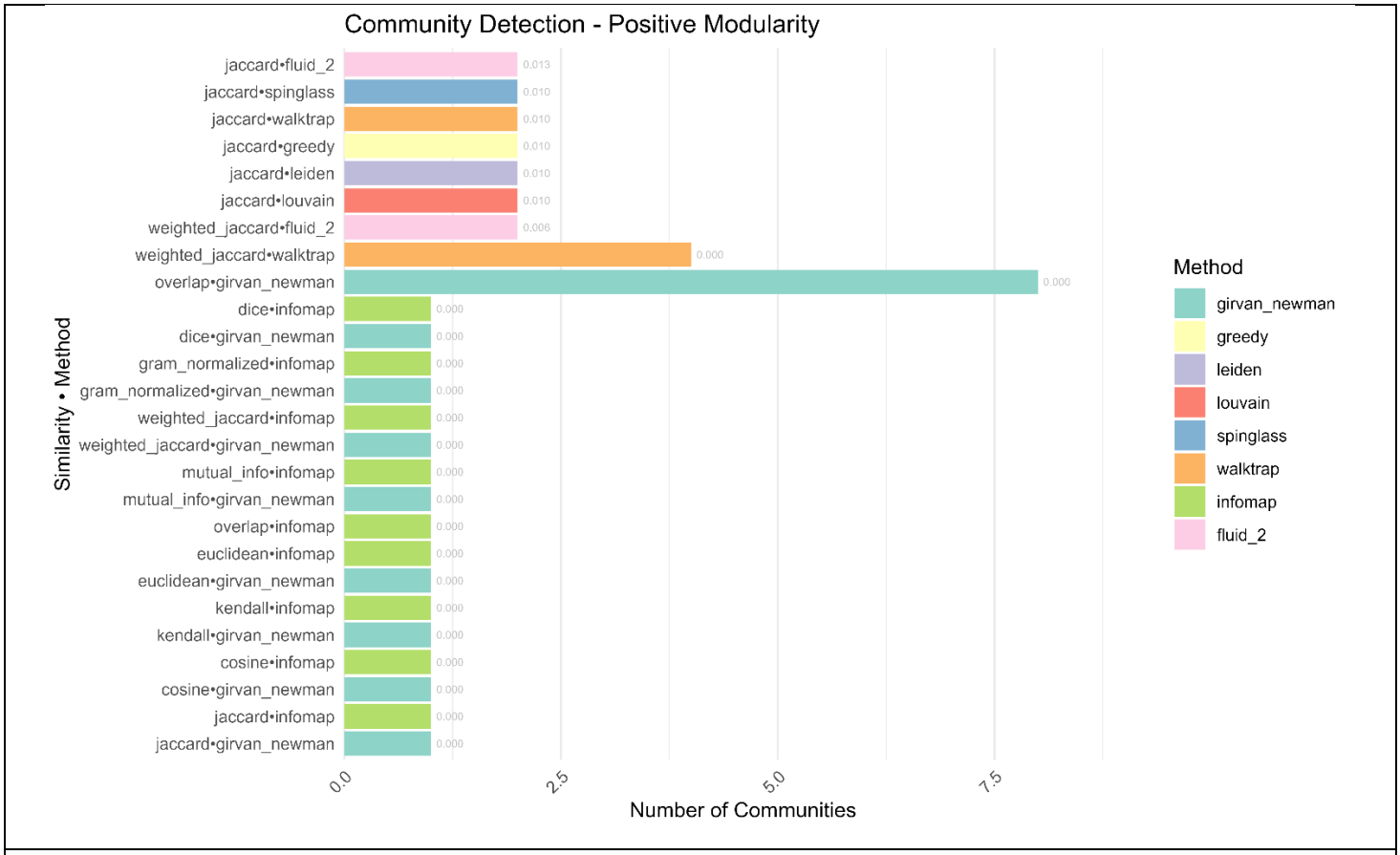


Figure 12. Number of communities detected by each community-detection model and similarity-matrix combination for the year 2023. Results is relative to models with a modularity value ≥ 0. Bars are sorted top-to-bottom in descending order of modularity. The x-axis denotes the number of communities identified.

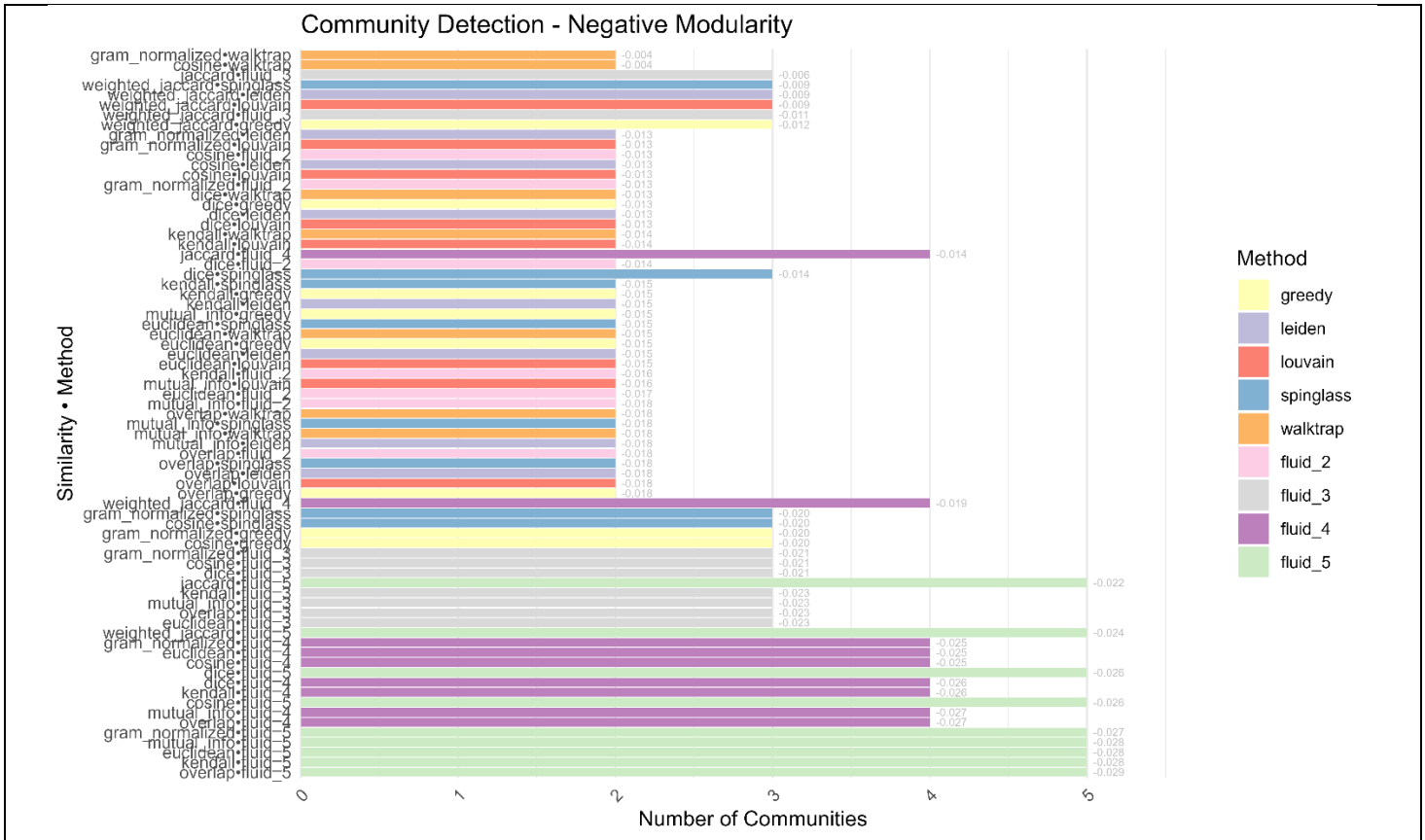
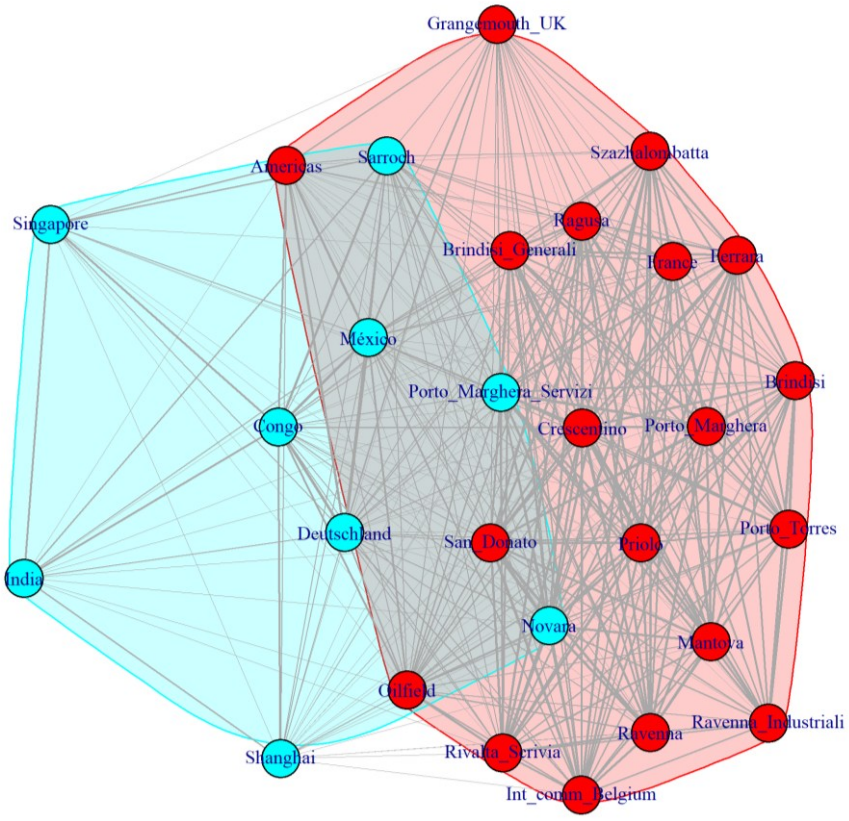


Figure 13. Number of communities detected by each community-detection model and similarity-matrix combination for the year 2023. Results is relative to models with a modularity value < 0. Bars are sorted top-to-bottom in descending order of modularity. The x-axis denotes the number of communities identified.

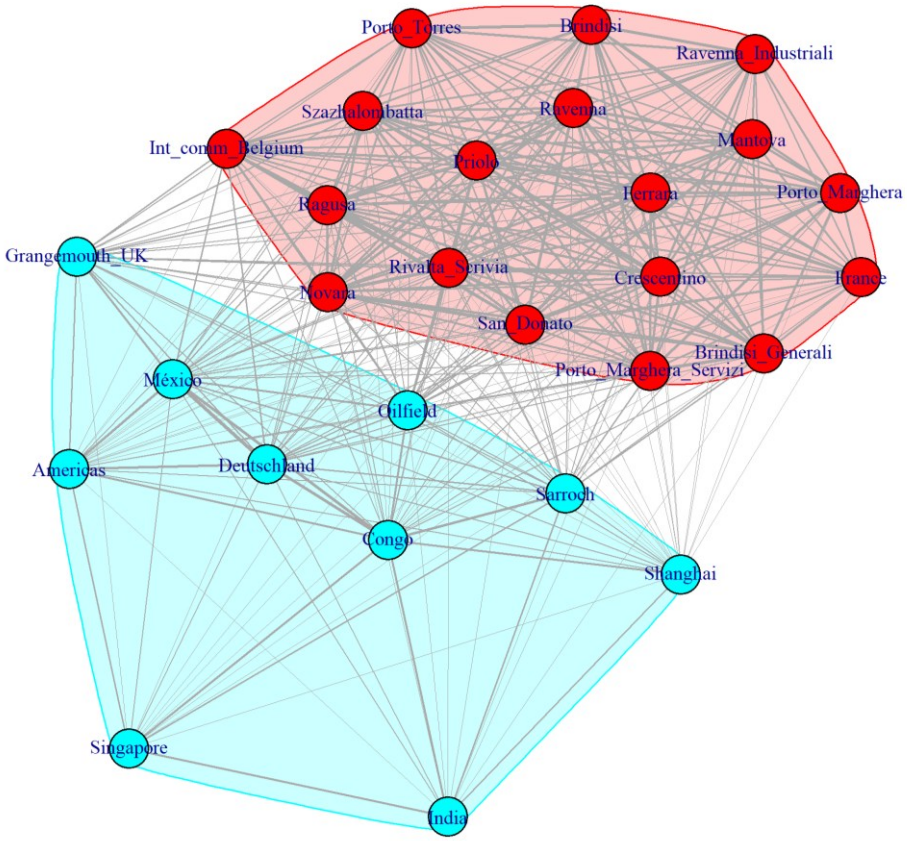
Jaccard matrix – Fluid community detection model



[1]
"Brindisi_Generali"; "Szazhalombatta"; "Ravenna_Industriali"; "Americas"; "France"; "Int_comm_Belgium"; "Oilfield"; "Rivalta_Scrivia"; "San_Donato"; "Brindisi"; "Crescentino"; "Ferrara"; "Mantova"; "Porto_Marghera"; "Porto_Torres"; "Priolo"; "Ragusa"; "Ravenna"; "Grangemouth_UK"
[2]
Porto_Marghera_Servizi"; "Congo"; "Deutschland"; "México"; "India"; "Shanghai"; "Singapore"; "Novara"; "Sarroch"

Figure 14. Network of sites clustered into two communities by the *fluid propagation* algorithm on the *Jaccard* similarity matrix. Node colors reflect community membership; edge widths are proportional to Jaccard similarity.

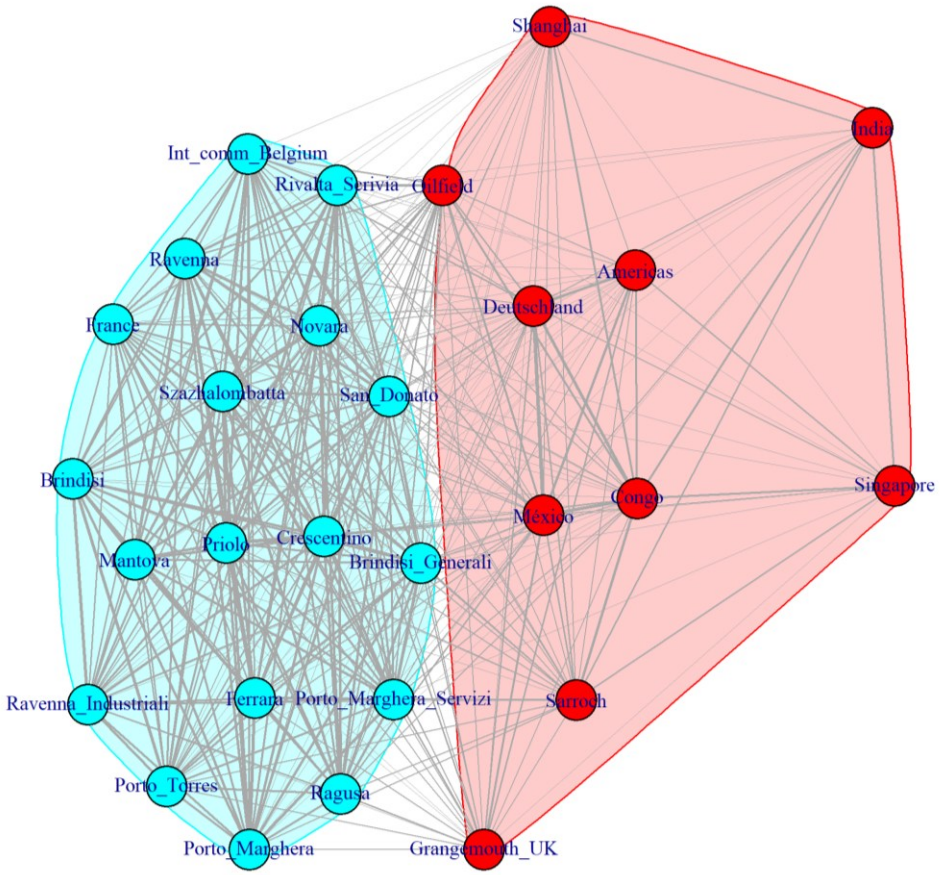
Jaccard matrix – Spinglass community detection model



[1]
"Brindisi_Generali"; "Szazhalombatta"; "Ravenna_Industriali"; "Porto_Marghera_Servizi"; "France"; "Int_comm_Belgium"; "Novara"; "Rivalta_Scrivia"; "San_Donato"; "Brindisi"; "Crescentino"; "Ferrara"; "Mantova"; "Porto_Marghera"; "Porto_Torres"; "Priolo"; "Ragusa"; "Ravenna"
[2]
"Americas"; "Congo"; "Deutschland"; "México"; "India"; "Shanghai"; "Singapore"; "Oilfield"; "Sarroch"; "Grangemouth_UK"

Figure 15. Network of sites clustered into two communities by the *spinglass* algorithm on the *Jaccard* similarity matrix. Node colours reflect community membership; edge widths are proportional to Jaccard similarity.

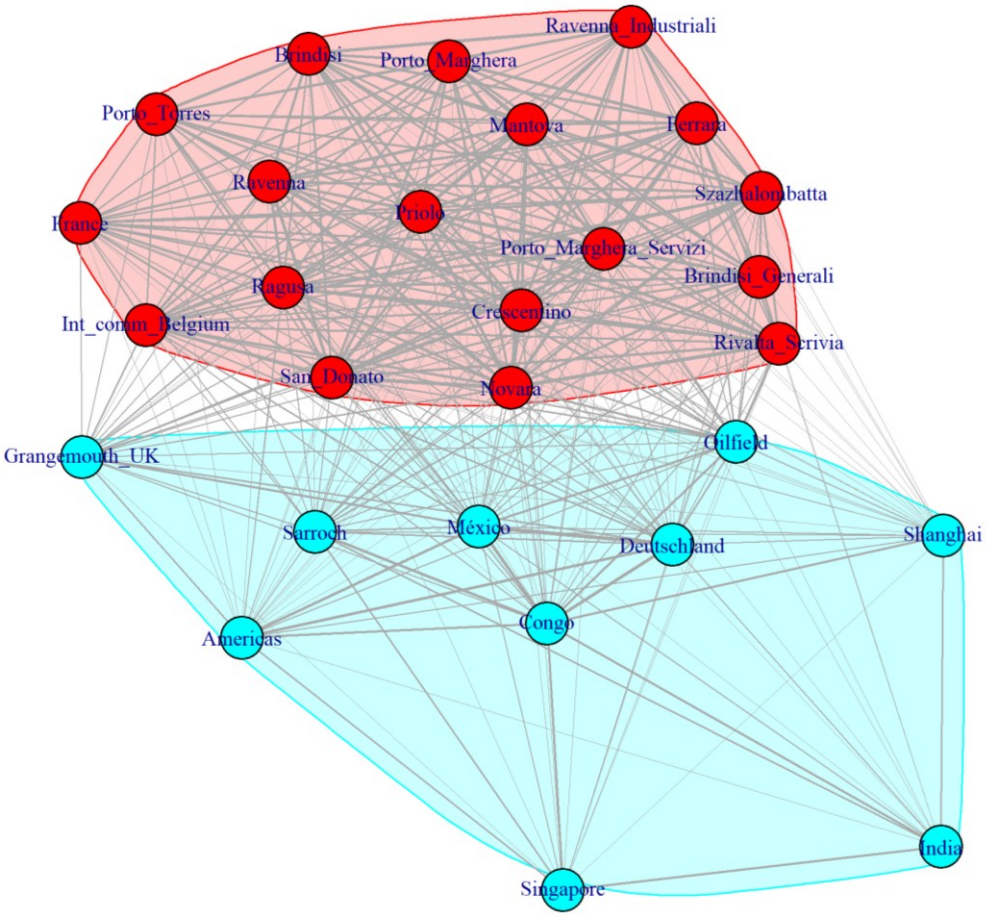
Jaccard matrix – Walktrap community detection model



[1]
"Americas", "Congo"; "Deutschland"; "México"; "India"; "Shanghai"; "Singapore"; "Oilfield"; "Sarroch"; "Grangemouth_UK"
[2]
"Brindisi_Generali"; "Szazhalombatta"; "Ravenna_Industriali"; "Porto_Marghera_Servizi"; "France"; "Int_comm_Belgium"; "Novara"; "Rivalta_Scrivia"; "San_Donato"; "Brindisi"; "Crescentino"; "Ferrara"; "Mantova"; "Porto_Marghera"; "Porto_Torres"; "Priolo"; "Ragusa"; "Ravenna"

Figure 16. Network of sites clustered into two communities by the *walktrap* algorithm on the *Jaccard* similarity matrix. Node colours reflect community membership; edge widths are proportional to Jaccard similarity.

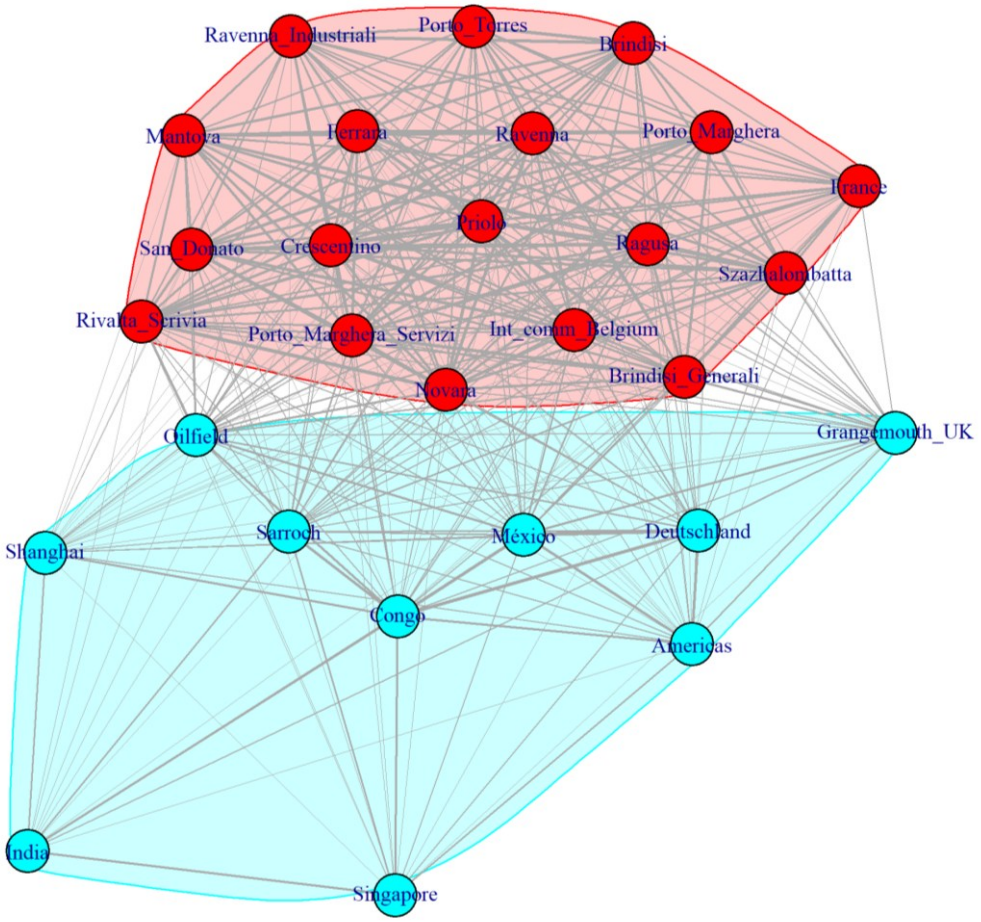
Jaccard matrix – Greedy community detection model



[1]
"Brindisi_Generali"; "Szazhalombatta"; "Ravenna_Industriali"; "Porto_Marghera_Servizi"; "France"; "Int_comm_Belgium"; "Novara"; "Rivalta_Scrivia"; "San_Donato"; "Brindisi"; "Crescentino"; "Ferrara"; "Mantova"; "Porto_Marghera"; "Porto_Torres"; "Priolo"; "Ragusa"; "Ravenna"
[2]
"Americas"; "Congo"; "Deutschland"; "México"; "India"; "Shanghai"; "Singapore"; "Oilfield"; "Sarroch"; "Grangemouth_UK"

Figure 17. Network of sites clustered into two communities by the *greedy* algorithm on the *Jaccard* similarity matrix. Node colours reflect community membership; edge widths are proportional to Jaccard similarity.

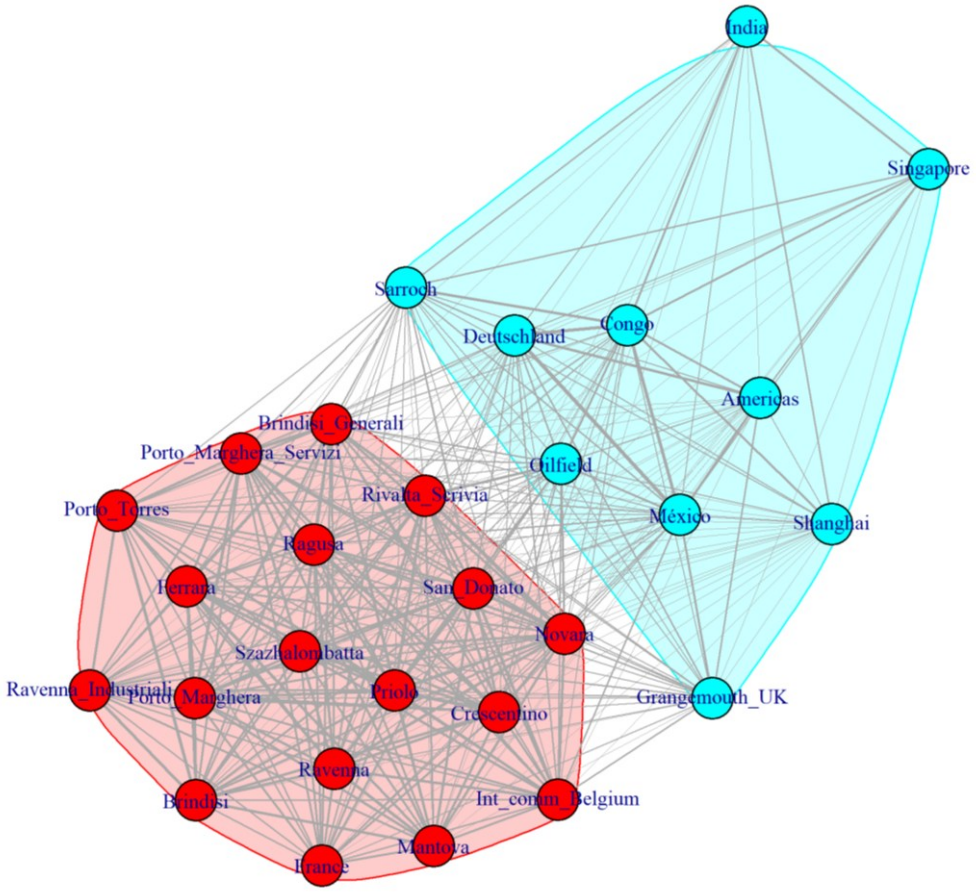
Jaccard matrix – Leiden community detection model



[1]
"Brindisi_Generali"; "Szazhalombatta"; "Ravenna_Industriali"; "Porto_Marghera_Servizi"; "France"; "Int_comm_Belgium"; "Novara"; "Rivalta_Scrivia"; "San_Donato"; "Brindisi"; "Crescentino"; "Ferrara"; "Mantova"; "Porto_Marghera"; "Porto_Torres"; "Priolo"; "Ragusa"; "Ravenna"
[2]
"Americas"; "Congo"; "Deutschland"; "México"; "India"; "Shanghai"; "Singapore"; "Oilfield"; "Sarroch"; "Grangemouth_UK"

Figure 18. Network of sites clustered into two communities by the *Leiden* algorithm on the *Jaccard* similarity matrix. Node colours reflect community membership; edge widths are proportional to Jaccard similarity.

Jaccard matrix – Louvain community detection model



[1]
"Brindisi_Generali"; "Szazhalombatta"; "Ravenna_Industriali"; "Porto_Marghera_Servizi"; "France"; "Int_comm_Belgium"; "Novara"; "Rivalta_Scrivia"; "San_Donato"; "Brindisi"; "Crescentino"; "Ferrara"; "Mantova"; "Porto_Marghera"; "Porto_Torres"; "Priolo"; "Ragusa"; "Ravenna"
[2]
"Americas"; "Congo"; "Deutschland"; "México"; "India"; "Shanghai"; "Singapore"; "Oilfield"; "Sarroch"; "Grangemouth_UK"

Figure 19. Network of sites clustered into two communities by the *Louvain* algorithm on the *Jaccard* similarity matrix. Node colours reflect community membership; edge widths are proportional to Jaccard similarity.

As with the 2022 analysis, dissimilarity matrices were also computed for 2023 and incorporated into a hierarchical clustering approach. The performance of these models is reported in Figures 20-21. “Again, these methods performed worse than the community-detection algorithms, and their performance in 2023 was even poorer than in

2022. The best model ─ also the only one with a modularity value greater than 0 ─ is shown in Figure 22.

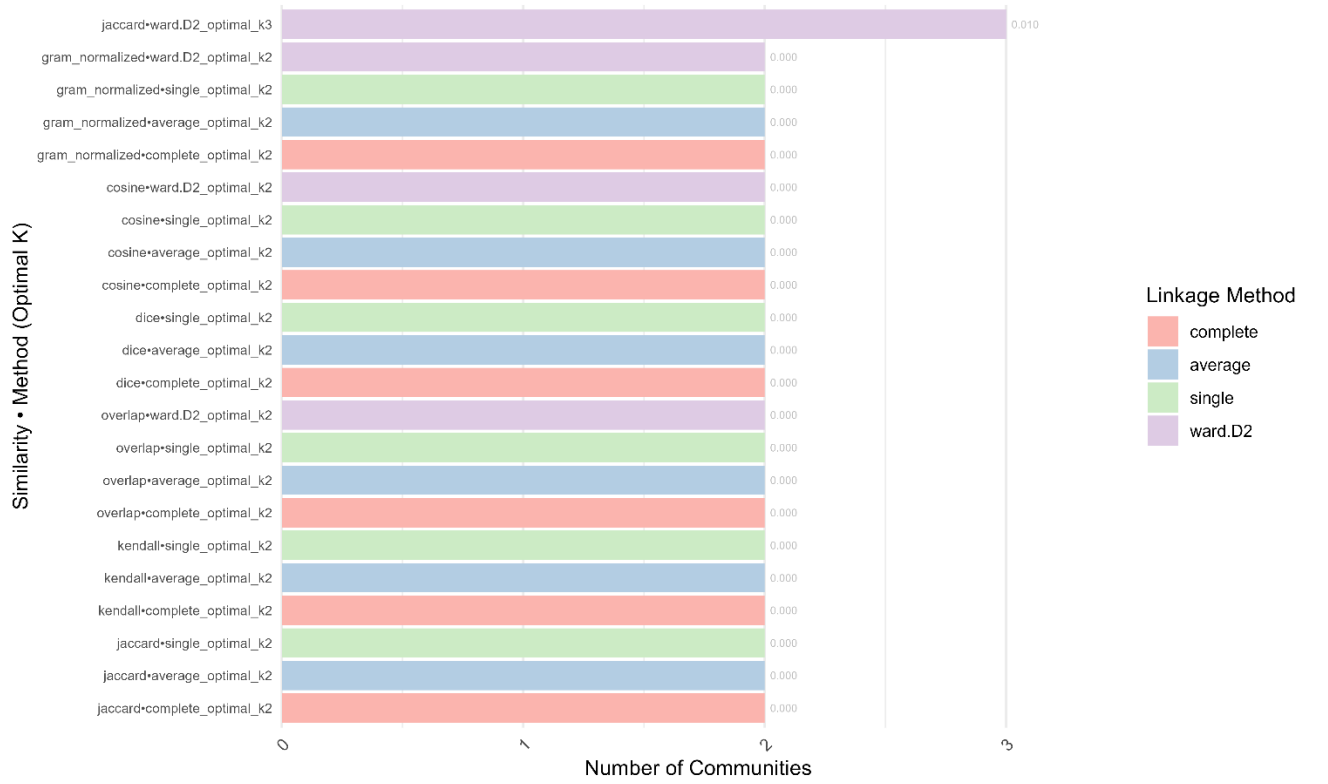


Figure 20. Number of communities detected by each hierarchical model and similarity-matrix combination for the year 2023. Results is relative to models with a modularity value ≥ 0. Bars are sorted top-to-bottom in descending order of modularity. The x-axis denotes the number of communities identified.

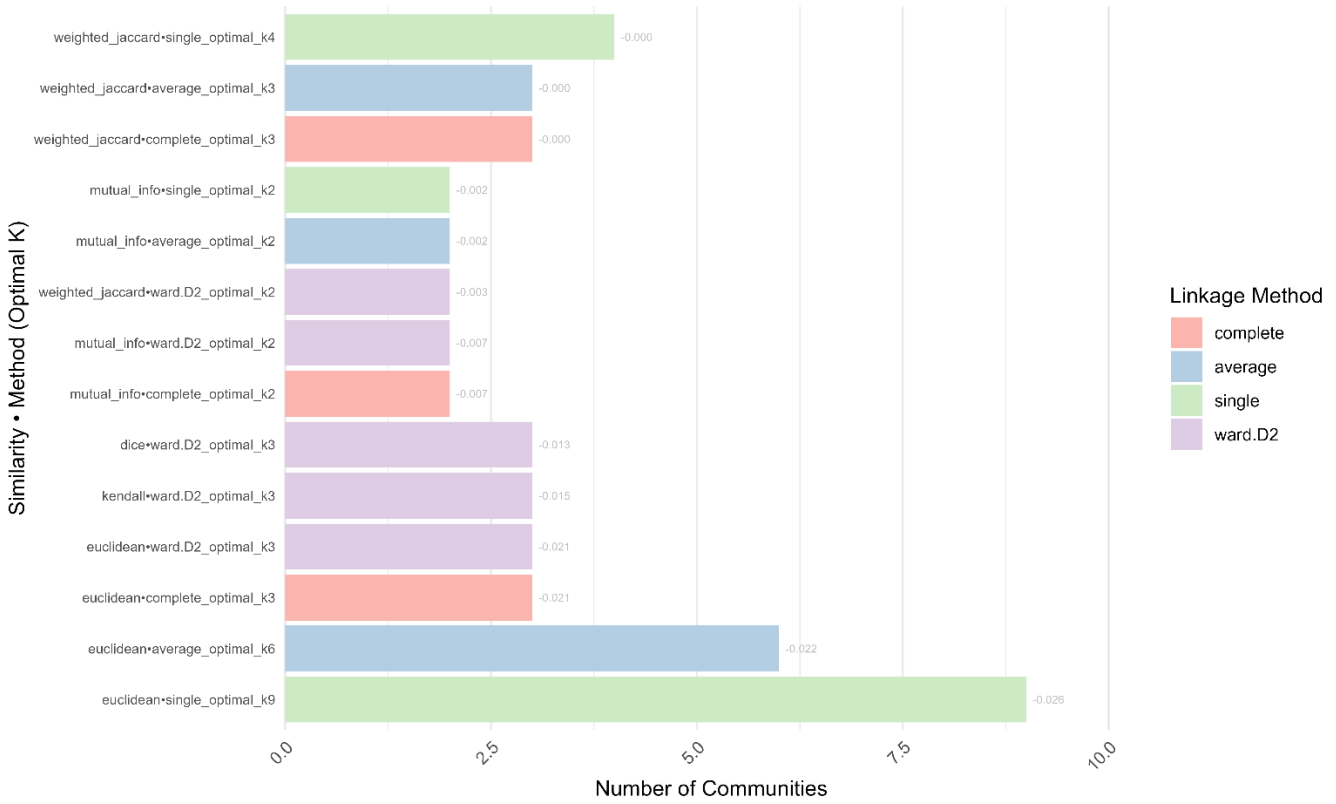


Figure 21. Number of communities detected by each hierarchical model and similarity-matrix combination for the year 2023. Results is relative to models with a modularity value < 0. Bars are sorted top-to-bottom in descending order of modularity. The x-axis denotes the number of communities identified.

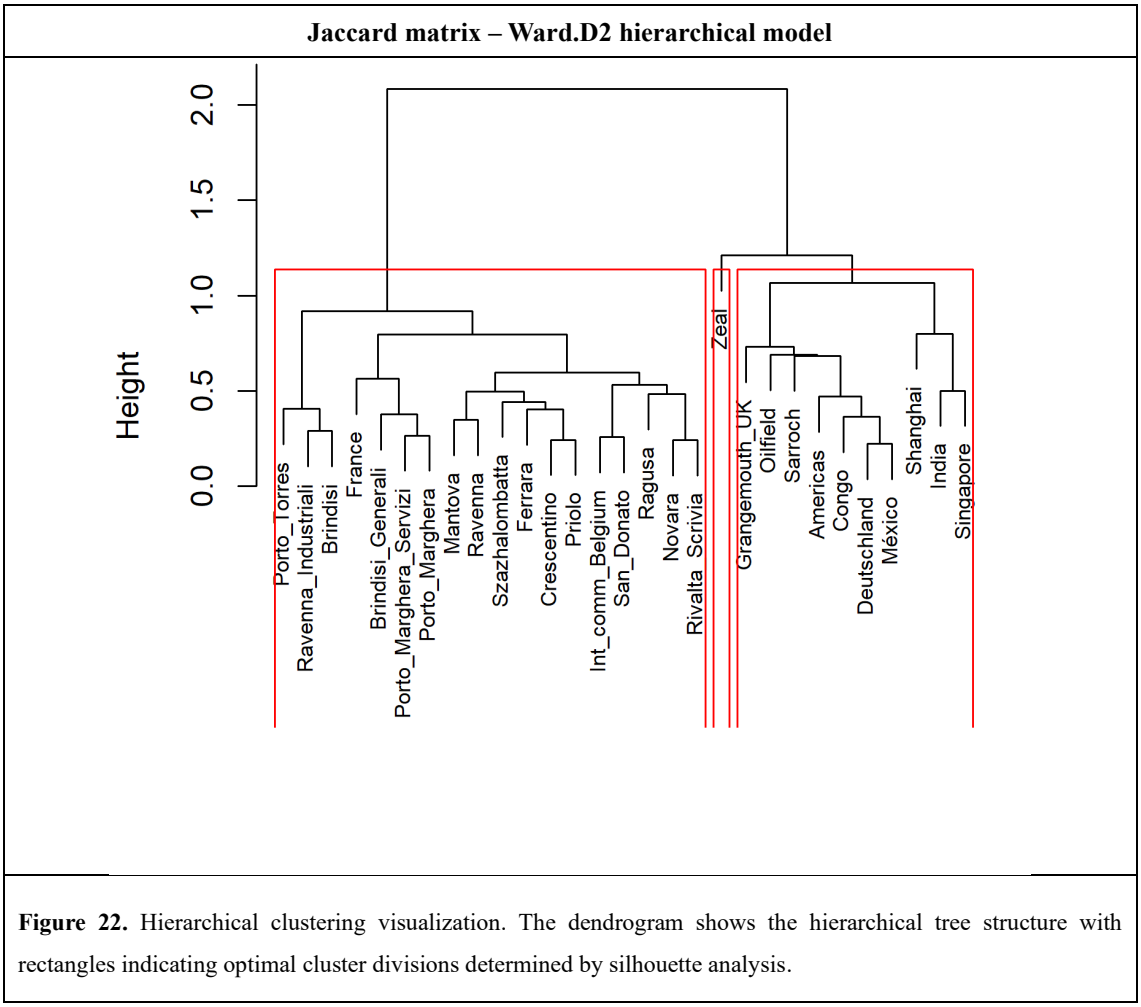


Figure 22. Hierarchical clustering visualization. The dendrogram shows the hierarchical tree structure with rectangles indicating optimal cluster divisions determined by silhouette analysis.

While there are some differences between the two years, the *Jaccard* similarity matrix generally appears to be associated with the best performance across the various community detection algorithms. The *Dice* coefficient also performed well in establishing connections for 2022.

In contrast, the *Fluid* algorithm showed the poorest performance when the number of communities was set to more than two. Most models identified an optimal number of communities equal to 2. The *Girvan–Newman* algorithm, tended to return a single community, but always with poor value of modularity.

Overall, all the models exhibit suboptimal performance, with modularity values close to zero or even negative, indicating a largely random assignment of sites to communities.

In general, the division between communities appears to distinguish Italian from foreign locations. The most frequent exceptions were Sarroch (Sardinia), which was often

grouped with foreign sites, Belgium (commercial site), France and Százhalombatta (Hungary), which were commonly associated with Italian sites. Grangemouth (UK) showed this behaviour only in 2022.

3. Anomaly and novelty detection

3.1. Background

Machine learning in clinical research had to face challenges associated with unbalanced datasets, particularly when modelling rare yet clinically critical outcomes[57]. Classical supervised learning algorithms are especially prone to generating biased decision boundaries and exhibiting poor sensitivity toward minority classes[58,59]. Over the past decade, research on imbalanced medical data has converged on the necessity of domain-aware solutions that integrate data-level techniques (e.g., tailored over- and under-sampling or feature-level preprocessing) with algorithmic adaptations and hybrid strategies[60]. These combined approaches have proven essential for developing reliable models in high-stakes diagnostic contexts[61].

Even well-established machine learning techniques ─ such as decision trees, support vector machines, and neural networks ─ often struggle with highly skewed class distributions, leading to biased predictions and limited generalization ability. To address these challenges, advanced methods such as synthetic over-sampling (e.g., SMOTE and its derivatives like ADASYN), cost-sensitive learning, and ensemble modelling have been developed. These techniques enhance the robustness of predictive models against imbalance and are particularly relevant when combined with anomaly detection frameworks[60].

Recent advances in anomaly detection have demonstrated that unsupervised and semi-supervised methods, such as Isolation Forest, Local Outlier Factor, and hybrid SVM-based pipelines, can effectively identify contextual anomalies in electronic health records and continuous clinical data streams, even when labelled anomalies are scarce[62]. However, the performance of these models often remains modest if there is no strong characterization of the majority class[60,63].

Anomaly and novelty detection represent fundamental and closely related problems in machine learning, both concerned with identifying observations that deviate from expected behaviour. Unlike supervised learning, these methods typically operate in unsupervised or semi-supervised settings, where the model must infer the concept of "normality" based on predominantly normal data. The subtle distinction lies in the treatment of the training data: anomaly detection assumes some anomalies are already present, whereas novelty detection presumes that only normal observations are available and aims to flag previously unseen deviations during deployment.

A wide range of methodologies underpin anomaly detection. Classical statistical models, such as Gaussian mixture models and kernel density estimators, define normal behaviour probabilistically and label low-likelihood observations as anomalies. Distance-based methods, including k-nearest neighbours and the Local Outlier Factor (LOF), rely on the principle that normal points cluster densely while outliers are relatively isolated; LOF enhances this approach by considering local density variations. Machine learning techniques, such as One-Class Support Vector Machines (OCSVM) and Isolation Forests, have expanded the toolkit further. OCSVM defines a boundary around normal data in a transformed feature space, whereas Isolation Forest exploits the principle that anomalies are easier to isolate due to their distinctiveness, making it computationally efficient and robust in high-dimensional data. More recently, deep learning methods such as autoencoders and generative adversarial networks (GANs) have proven highly effective at learning representations of normal patterns and identifying deviations via reconstruction or discrimination errors[60,62,64].

Class imbalance remains one of the most pressing methodological issues in clinical data analysis, particularly in studies of rare diseases or low-incidence clinical events. These datasets often require extensive longitudinal collection or multicentre collaborations to achieve sufficient statistical power for traditional inferential approaches. Improve statistical methods to predict the occurrence of these events is relevant not only for rare genetic diseases but also for clinically significant events that occur rarely but have serious consequences.

One of the possible scenarios in which the occurrence of the event is not so frequent in the clinical setting concerns catheter-related bloodstream infections (CRBSI), central line-associated bloodstream infection (CLABSI) or more generally catheter-associated

bloodstream infections (CABSI). Potential infection episodes can be treated as deviations from established baselines of catheter use and patient trajectories. This modelling perspective has the potential to enable earlier identification of infection, mitigate complications, and optimize resource allocation within hospital settings. Although several studies have attempted to determine the most suitable models for predicting infectious events related to vascular catheterization, significant methodological limitations and potential biases remain[65,66].

Catheter-associated infections exemplify this scenario, affecting only a minority of patients with central venous catheters, with reported incidence ranging from less than one to several episodes per 1,000 catheter-days. Despite their rarity, catheter infections remain a major cause of hospital-acquired bloodstream infection and are associated with increased morbidity, prolonged hospitalization, considerable mortality, and substantial healthcare costs[67–69]. The economic burden is substantial at the health-system level, with per-episode costs for CRBSI/CABSI estimated to range from approximately $6,000 to over $70,000, and hospital-level analyses reporting tens of thousands of euros or dollars in incremental costs per 1,000 catheter-days[70–73]. In addition to these clinical and economic consequences, patients frequently report pain, discomfort, functional limitations, and restrictions in social and daily activities following catheter-related complications, indicating a meaningful deterioration in health-related quality of life[74–76].

No validated pre-insertion risk score exists for catheter-associated infections. Prolonged dwell time, type of catheters lumen, insertion sites, and patient frailty consistently emerge as key risk factors[77]. However, available evidence remains limited to selected cohorts (paediatrics, specific pathologies)[78–80], necessitating broader studies to strengthen generalizability and inform universal clinical protocols.

Moreover, several studies have shown a strong association between these infections and multidrug-resistant organisms, underscoring the complexity of their management and their contribution to the broader challenge of antimicrobial resistance[73,81].

It therefore follows that the ability to automatically identify patient groups whose trajectories deviate from expected patterns can enable timely interventions, inform personalized treatment planning, and drive continuous improvement in the quality of clinical care.

3.2. Objective

The aim of this second section is to introduce a new novelty-detection model that may outperform existing approaches described in the literature. To this end, we use retrospective data from patients admitted to the “Luigi Sacco Hospital” in Milan, a referral center for infectious disease management. The dataset includes clinical and biometric characteristics of the patients, as well as information on the type of catheter used.

The proposed model (OC-Cat) is a three-stage pipeline designed to identify observations that deviate from the majority class (described in detail in the Methods section). Its components ─ feature reduction, majority-class characterization with anomaly detection (the core objective of this work), and feature-importance classification ─ can operate independently or be integrated with other analytical frameworks. To evaluate the predictive performance of OC-Cat, we selected two established and methodologically complementary anomaly-detection algorithms, Isolation Forest and One-Class SVM, as benchmarks due to their robustness and scalability in high-dimensional clinical settings.

The overarching goal is to determine whether the proposed model can reliably identify patients at increased risk of CRBSI/CABSI (or more broadly, catheter-associated infections) using only pre-insertion patient biometrics, clinical variables, and catheter characteristics. By highlighting patient- and catheter-specific risk signals at insertion, the model enables clinicians to implement heightened monitoring and intensive catheter care protocols for high-risk cases, potentially preventing infections before clinical manifestation.

3.3. Materials

Data are referred to a cohort of patients admitted at “Luigi Sacco Hospital” in Milan (Italy) from January 2021 to January 2025. All included patients underwent central or peripheral catheterization. Parameters reported are referring to the moment of catheter insertion. Patients admitted to an intensive care unit, or who had a CRBSI/CABSI diagnosis or a vascular access device (VAD) placement either in the ICU or within 48 hours of transfer to a non-ICU department, patients with catheter insertion during Day-

Hospital admission or patients who received short peripheral catheters were excluded from the analysis.

The study protocol received approval from the Institutional Review Board of "Luigi Sacco Hospital" (Research Ethics Committee approval number 2021/ST/180). Information collected included age, biological sex, and several comorbidities, such as neurological diseases, cardiovascular diseases, lung diseases and liver diseases. Hypertension was defined as blood pressure ≥140mmHg or ≥90 mmHg, obesity as BMI ≥30 kg/m2, while diabetes as suggested by WHO[82]. Chronic kidney disease (CKD) was defined as either kidney damage or a decreased glomerular filtration rate (GFR) of less than 60 mL/min/1.73m2 for at least 3 months[83]. Liver disease was defined as presence cirrhosis. Active haematological or solid neoplasm are considered together. We collected also active chemotherapy, use of parenteral nutrition, drug addiction, condition of potential immunosuppression ─ Hematopoietic Steam Cells Transplantation (HSCT) and Solid organ transplantation, Human Immunodeficiency Virus (HIV) positivity, immunosuppressant therapy ─ autoimmune disorders, presence of SARS-CoV-2 positivity and eventually associated pneumonia, tracheostomy and age-adjusted Charlson Comorbidity Index (CCI)[84]. Regarding hospitalization data, we considered the type of ward (clinical or surgical), a possible coming from ICU or a previous infection within 30 days. In this analysis, we excluded day hospital catheter placements. Information about catheter included number of lumens (1 vs >1), and presence of a tunnel, number of attempts (1 vs >1), type of VAD (power injectable or not) and site of insertion (upper vs lower limb vs head/neck).

We considered CABSI to be any bloodstream infection occurring in a patient with a catheter, both central or peripheric. We deemed CRBSI as an infection where the same organism is isolated from both a blood culture and a catheter tip culture, with the tip culture yielding more than 15 colony-forming units (CFUs), or, alternatively, when the same organism was isolated from both a peripheral vein and a VAD cultures, with the catheter sample turning positive at least 2 hours earlier than the peripheral sample, according to differential time to positivity (DTP) criterion[85,86].

3.4. Methods

Three anomaly detection methods were benchmarked to identify potential outliers among all catheter insertion records: the proposed OC-Cat model, Isolation Forest, and OCSVM). These methods are detailed below.

3.4.1. Isolation Forest

Isolation Forest represents one of the anomaly detection approaches. It exploits the fundamental principle that anomalies are inherently easier to isolate than normal observations through recursive binary partitioning. Developed by Liu et al. in 2008[87], this algorithm operates on the intuitive premise that outliers, being few and different, require fewer random partitions to be separated from the bulk of the data, thus exhibiting shorter path lengths in randomly constructed binary trees[87,88]. The method constructs an ensemble of isolation trees, where each tree T is built by recursively selecting a random feature $q\ in\ \{1, 2, \ldots, d\}$ from the d-dimensional feature space and a random split value p uniformly distributed between the minimum and maximum values of that feature: $p \sim \mathcal{U}(\min(X_q), \max(X_q))$, where X_q represents all values of feature q in the current node. This recursive partitioning continues until each observation is isolated in its own leaf node or a predetermined maximum depth $\ell = \lceil \log_2 n \rceil$ is reached, where n is the number of observations in the training set.

The mathematical foundation of Isolation Forest rests upon the analysis of path lengths in binary search trees and their relationship to data density. For a dataset $X = \{x_1, x_2, \ldots, x_n\}$ with $x_i \in \mathbb{R}^d$, the path length $h(x, T)$ of an observation x in tree T represents the number of edges traversed from the root to the terminating node. The expected path length of an observation is computed as:

$$E[h(x)] = \frac{1}{t}\sum_{i=1}^{t} h(x, T_i)$$

where t is the number of trees in the ensemble. The algorithm's theoretical foundation leverages the fact that the average path length of unsuccessful searches in a Binary Search Tree (BST) with n nodes follows the relationship:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} = 2(\ln(n-1) + \gamma) - \frac{2(n-1)}{n}$$

where $H(k) = \sum_{i=1}^{k} \frac{1}{i}$ is the k-th harmonic number and $\gamma \approx 0.5772$ is the Euler-Mascheroni constant. This expression provides the normalization constant that accounts for the expected path length in a randomly constructed binary tree[87,89].

The anomaly score computation forms the core of the Isolation Forest's decision mechanism, transforming raw path lengths into interpretable anomaly measures. The anomaly score for an observation x given a dataset of size n is defined as:

$$s(x,n) = 2^{-\frac{E[h(x)]}{c(n)}}$$

This exponential transformation ensures several desirable properties: when $E[h(x)] \to c(n)$, the score approaches $s(x,n) \to 2^{-1} = 0.5$, indicating normal behaviour; when $E[h(x)] \to 0$ (very short paths), $s(x,n) \to 2^{0} = 1$, signaling clear anomalies; and when $E[h(x)] \to n-1$ (maximum possible path length), $s(x,n) \to 0$, suggesting the observation is deeply embedded within normal regions. The score interpretation follows the rule:

- $s(x,n) \approx 1$: Observation isolated with very few splits → Clear anomaly
- $s(x,n) \approx 0.5$: Average path length, typical behaviour → Normal point
- $s(x,n) \approx 0$: Very long path, deeply embedded → Very Normal

The tree construction algorithm follows a precise recursive procedure that balances randomness with structural constraints. Given a node containing a subset $X' \subseteq X$ of the training data, the splitting procedure is formally defined as:

1. *Feature Selection*: Choose $q \sim \mathcal{U}\{1, 2, \ldots, d\}$ uniformly at random
2. *Split Point Selection*: Choose $p \sim \mathcal{U}(\min(X'_q)(\max(X'_q))$, uniformly between the minimum and maximum values of feature q in the current subset
3. *Node Partitioning*: Create left child $X'_L = \{x \in X' : x_q < p\}$ and right child $X'_R = \{x \in X' : x_q < p\}$
4. *Termination Conditions*: Stop if $|X'| \leq 1$ or current depth equals ℓ

The path length for an observation x is then computed as:

$$h(x,T) = \begin{cases} e + c(|X'|) & \text{if } x \text{ reaches external node with } |X'| > 1 \\ e & \text{if } x \text{ reaches external node with } |X'| = 1 \end{cases}$$

where e is the current depth and $c(|X'|)$ adjusts for unresolved instances when early termination occurs.

The ensemble construction and score aggregation process involves several critical parameters that influence the algorithm's performance and computational complexity. The complete Isolation Forest algorithm operates as follows:

$$IsolationForest(\mathrm{X}, t, \psi) = T_1, T_2, \ldots, T_t$$

where t is the number of trees (typically $t = 100$) and ψ is the subsampling size (typically $\psi = 256$ or $\psi = \min(256, n)$). Each tree T_i is constructed using a random subsample $\mathrm{X}_i \subset \mathrm{X}$ with $|\mathrm{X}_i| = \psi$, selected uniformly without replacement. The subsampling serves dual purposes: reducing computational complexity from $O(n^2)$ to $O(n\psi)$ and improving the algorithm's focus on anomalies by reducing the masking effect of normal instances.

The computational complexity analysis reveals Isolation Forest's significant efficiency advantages over traditional anomaly detection methods. The training phase complexity is $O(t \cdot \psi \cdot \log\psi)$, where the logarithmic factor comes from the expected tree depth. For evaluation, each query requires $O(t \cdot \log\psi)$ operations to traverse all trees. Memory complexity is $O(t \cdot \psi)$ for storing the ensemble. Compared to distance-based methods with $O(n^2)$ complexity for computing pairwise distances, or density-based methods requiring $O(n \log n)$ for nearest neighbor searches, Isolation Forest offers substantial computational savings, particularly for large datasets where $\psi \ll n$.

Under the assumption that anomalies constitute a small fraction α of the data and are distributed differently from the normal instances, the probability that an anomaly is isolated earlier than normal points can be bounded. Specifically, if $P_a(h \leq k)$ denotes the probability that an anomaly has path length at most k, and $P_n(h \leq k)$ denotes the same for normal points, then under mild regularity conditions:

$$\lim_{n\to\infty} \frac{P_a(h \leq k)}{P_n(h \leq k)} > 1$$

for appropriately chosen k. This theoretical result validates the algorithm's core assumption and provides confidence bounds for anomaly detection performance.

Furthermore, the consistency of the anomaly score estimator can be established: as the number of trees $t \to \infty$, the sample average

$$\frac{1}{t}\sum_{i=1}^{t} h(x, T_i)$$

converges to the true expected path length $E[h(x)]$ by the strong law of large numbers, ensuring that the anomaly scores become increasingly reliable with larger ensembles.

3.4.2. One-Class Support Vector Machine

One-Class Support Vector Machine (OCSVM) represents a sophisticated adaptation of the classical SVM framework for the unsupervised anomaly detection problem, fundamentally reformulating the binary classification paradigm to identify a single class boundary that encapsulates normal data while excluding outliers. The algorithm operates by finding a hyperplane in a high-dimensional feature space that separates the training data from the origin with maximum margin, effectively creating a decision boundary that encompasses the majority of normal observations while maintaining sensitivity to anomalous patterns through the careful balance of the ν-parameter and kernel-induced geometric transformations[90,91].

OCSVM starts from the idea that the vast majority of your data is normal. It learns a compact region in feature space that encloses those normal points. Anything that falls outside this “normalcy region” is treated as a potential anomaly. During training, OCSVM positions a boundary around the normal instances and maximizes the margin between that boundary and the data. This margin acts like a safety buffer, making the model robust to noise and small variations. The tighter the boundary, the more confident you can be that points outside it are true outliers.

OCSVM differs from traditional SVMs by using only examples of the normal class. It doesn’t require labelled anomalies, which makes it ideal when anomalies are rare or hard to obtain. By focusing exclusively on normal behaviour, the model crafts a decision boundary that truly reflects what “normal” looks like.

The parameter v serves as an upper bound on the fraction of margin errors and support vectors. Lower v values produce a stricter boundary (fewer false positives, more missed

anomalies). Higher v values loosen the boundary (more tolerance for anomalies, at the cost of potentially flagging some normal points)[92,93].

Once the normalcy region is learned, new data points are tested against that boundary. Points inside the region pass as normal, while those outside trigger an anomaly alert. In this way, OCSVM functions as a robust boundary detector, continuously monitoring incoming data for unusual behaviour.

To formalize this idea, we imagine mapping the input data into a feature space via $\phi(\cdot)$ In that space, the OCSVM aims to push the hyperplane away from the origin so that the majority of points satisfy $w^T\phi(x_i) \geq \rho$, allowing some slack for violations. This leads to:

$$\min_{w,\rho,\xi} \frac{1}{2} \|w\|^2 - \rho + \frac{1}{vn} \sum_{i=1}^{l} \xi_i$$

subject to:

$$w^T\phi(x_i) \geq \rho - \xi_i, \qquad \xi_i \geq 0$$

where ξ_i is the slack variable for point i that allows it to lie on the other side of the decision boundary, n is the size of the training dataset and $v \in (0,1]$ is the regularization parameter controlling trade-off, w is the normal vector to the separating hyperplane, ρ is the offset parameter (distance from origin to hyperplane) and $\varphi(x)$ is the feature mapping to high-dimensional space.

In high-dimensional spaces, direct computation is impractical. By introducing Lagrange multipliers α_i and applying the kernel trick, we obtain the dual problem:

$$\min_{\alpha} \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j)$$

subject to

$$0 \leq \alpha_i \leq \frac{1}{vn}, \quad \sum_{i=1}^{n} \alpha_i = 1$$

This formulation depends only on pairwise kernel evaluations $K(x_i, x_j)$, which define the geometry of the normalcy region[89]:

- **Linear kernel**: flat boundary, efficient for sparse/high-dimensional data. It is suitable when the relationship between the features is approximately linear.

- **Radial Basis Function (RBF) kernel**: smooth closed contours around clusters. It is versatile for handling complex, non-linear relationships.
- **Polynomial kernel**: curved boundaries with tunable complexity. The polynomial kernel introduces non-linearity by considering not just the dot product but also higher-order interactions between features. A higher degree allows the model to capture more complex relationships but in the same time it may increase the risk of overfitting.
- **Sigmoid kernel**: S-shaped separation, related to neural nets. It is particularly suitable for scenarios where the data distribution is not well defined or exhibits sigmoidal patterns.
- **Precomputed Kernel**: it allows users to provide a precomputed kernel matrix instead of the actual data. Useful when the kernel matrix is computed using a custom kernel function or when using pairwise similarities between instances.

Once trained, the OCSVM classifies a new sample x via the decision function:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{\mathrm{l}} \alpha_i K(x_i, x) - \rho\right)$$

where

$$Classification = \begin{cases} \text{Normal (inlier)} & \text{if } f(x) = +1 \\ \text{Anomaly (outlier)} & \text{if } f(x) = -1 \end{cases}$$

This correspond to a continuous anomaly score:

$$\mathrm{Score(x)} = \sum_{\mathrm{i}=1}^{n} \alpha_{\mathrm{i}} \mathrm{K}(\mathrm{x_i}, \mathrm{x}) - \rho$$

where

$$\mathrm{Classification} = \begin{cases} \text{Clear inlier} & if\ \ \mathrm{Score}(x) \gg 0 \\ \text{Boundary region} & if\ \ \mathrm{Score}(x) \approx 0 \\ \text{Clear outlier} & if\ \ \mathrm{Score}(x) \ll 0 \end{cases}$$

Geometrically, OCSVM finds the smallest-volume region in feature space that encloses most training data. The margin:

$$\text{Margin} = \frac{\rho}{\|w\|}$$

acts as a protective buffer between normal points and potential anomalies. The distance of a point from the boundary:

$$d(x) = \frac{|w^T \phi(x) - \rho|}{\|w\|}$$

quantifies how far it lies from the edge of normality.

3.4.3. OC-Cat

The new one-class classification framework (OC-Cat) proposed by Leoni et al[94,95]. integrated a graph search approach for feature selection, a soft one-class classification approach (based on the occurrence of the majority class) and a feature importance method. The three steps can also be applied independently of each other. Below we proceed to a brief description of the three steps used.

- **Feature selection**

In a clinical dataset, variables often tend to be highly correlated with each other. However, with categorical variables, classical methods such as Pearson's or Spearman's correlation may have limitations of applicability, as mentioned before.

So the Authors proposed a graph-based features selection approach. It consists in reducing the number of feature to include in the model starting with the evaluation of feature redundancy. This calculation is then used to build a weighted feature graph. We then find the shortest path that "best" explains the data distribution to identify a minimal feature subset. The objective is to be parsimonious to gain computational time e lost noise, to overcome the curse of dimensionality issue.

This novel approach, instead, do not require specific assumptions. Since the categorical features recorded in data set $\mathcal{D}$ may depend on each other, it is appropriate to converse only the most representative features within a set of dependent attributes while discarding others. To do this, we first assess the redundancy of the features using the "excess over independence" (EOI) metric between any pair of features[96]. An independence matrix that reports the values of this metric for each possible pair of features was then built.

$$Q(i,j) = \sum_{a \in F_i} \sum_{a \in F_j} \frac{\#(\chi_i^a \cap \chi_j^b)\#(\bar{\chi}_i^a \cap \bar{\chi}_j^b) - \#(\bar{\chi}_i^a \cap \chi_j^b)\#(\chi_i^a \cap \bar{\chi}_j^b)}{\#(\chi_i^a \cap \chi_j^b)\#(\bar{\chi}_i^a \cap \bar{\chi}_j^b) + \#(\bar{\chi}_i^a \cap \chi_j^b)\#(\chi_i^a \cap \bar{\chi}_j^b)}$$

So, given two features i and j, with two levels a and b, the resultant normalized excess over independence among the two features is

	***j*-th b_{present}**	***j*-th b_{absent}**
***i*-th a_{present}**	N=40	N=10
***i*-th a_{absent}**	N=14	N=30

$$Q^{ab} = \frac{40 \cdot 30 - 10 \cdot 14}{40 \cdot 30 + 10 \cdot 14} = \frac{1060}{1340} \approx 0.79$$

So it is measured how "redundant" any two features are. By turning features into nodes and their pairwise redundancy into edge-weights, we get a graph where:

- Nodes = features $x_1, x_2, \ldots, x_M$.
- Edge (v,w) weight $Q(v,w)$ = the "excess over independence" (higher means more redundant).

A path $\delta_{i \to j}$ is any sequence of edges that takes you from node i to node j. Its total cost is simply the sum of the Q-values along that route:

$$\text{cost}(\delta_{i \to j}) = \sum_{(v,w) \in \delta_{i \to j}} Q(v,w)$$

If i is connected to j via several intermediate features, there is an accumulation of redundancy at each hop. A smaller sum means a "less redundant" connection. All attributes will have a connection with at least another one, with the exception for those having a zero excess over independence with respect to all other features. Isolated vertices are attributes that have to be retained, as they are not redundant with respect to any other available information.

To find, for each source feature x_i, the path to every other feature x_j with the smallest summed redundancy, it is adopted the Bellman-Ford algorithm[97]. It solves exactly:

$$\delta_{\min i\to j} = \sum_{(v,w)\in\delta_{i\to j}} Q(v,w) \quad \forall\, i,j = 1 \dots M$$

Bellman-Ford repeatedly "relaxes" edges, guaranteeing we've found the least-cost route from i to all j in $O(M \times E)$ time, even if some weights were negative (irrelevant here since our $Q \geq 0$)

Once the minimal-redundancy cost is then computed from each $x_i$to every other feature, it is possible to quantify how "tightly" x_i is coupled to the rest of the set (e.g., by summing its path-costs), prefer features whose aggregate minimal-cost is low and discard features that lie in "dense" regions of the graph (high redundancy to many neighbours).

Once each pair of attributes was associated with the optimal path between, the Bayesian Information Criterion (BIC) was computed for each of the optimal paths retrieved

$$BIC\left(\delta_{i\to j}^{min}\right) = \log\left(P\left(H\middle|\delta_{i\to j}^{min}\right)\right) - \frac{1}{2}\log(H)\,\ell_{i\to j}^{min}), \qquad \forall i,j = 1,\dots,M$$

where $\ell_{i\to j}^{min}$ is the length of $\delta_{i\to j}^{min}$.

The probability $P\left(H\middle|\delta_{i\to j}^{min}\right)$ of the considered path describing the majority class dataset $\mathcal{H}$ is given by

$$\log\left(P\left(H\middle|\delta_{i\to j}^{min}\right)\right) = \sum_{v\in\delta_{i\to j}^{min}} \sum_{v\neq i} \sum_{b\in F_v} \sum_{a\in F_{\rho(v)}} \left(\mathcal{X}_{p(v)}^{a} \cap \mathcal{X}_{v}^{b}\right) \log\left(\frac{\left(\mathcal{X}_{p(v)}^{a} \cap \mathcal{X}_{v}^{b}\right)}{\sum_{a\in F_{\rho(v)}}\left(\mathcal{X}_{p(v)}^{a} \cap \mathcal{X}_{v}^{b}\right)}\right)$$

By finding the path leading to the maximal BIC

$$\delta * = \arg\max_{\delta_{i\to j}^{min}} BIC(\delta_{i\to j}^{min})$$

it is selected the subset S of the available M features that trades-off between the number of features needed to accurately characterize the distribution of the majority class data and the independence between features. The goal is to optimize information gain while imposing a penalty on the number of features used, thereby favouring models that extract maximal information from minimal input.

- **Soft classifier**

The soft-classifier rely on the assumption that a higher occurrence of a specific feature combination in majority class records implies that each new instance with those values is less likely to be infected. As a result, during the learning phase, the distribution of

majority class records was explored and each of them was given a weight proportional to the frequency with which that record occurred in the majority class dataset. The feature combinations should be mutually independent from each other.

During the prediction phase, instead, we estimate the majority-class probability for a new record (based on its i-th attribute combination) using a weighted inverse *Hamming distance*[98]. Given a new record, its distance from any majority-class is calculated and multiplied by the respective weight. It follows that records with higher weights, i.e. those more frequent among majority-class group, contribute more significantly to the calculation of the risk score. The model returns then a prediction probability value $\hat{\mathcal{R}}\left(X_{\text{new}}^{\text{red}}\right) = 1 - \hat{\mathcal{L}}(X_{\text{new}}^{\text{red}} \in \mathcal{H}) \rightarrow [0, 1]$, interpretable as a risk estimate. However, if necessary, it is possible to convert the model's continuous likelihood into binary predictions by choosing a decision threshold.

The Hamming distance is a measure of the difference between two strings (or vectors) of equal length. It counts the number of positions at which the corresponding elements (bits, characters, or numbers) are different. In other words, it provides a straightforward way to quantify how "far apart" two records are by simply counting mismatches.

It is often used for binary strings or vectors, but it can also be applied to categorical data if the elements are comparable.

$$D_{ij} = \sum_{k=1}^{n} 1(x_{ik} \neq x_{jk})$$

so that

$$1(x_{ik} \neq x_{jk}) \begin{cases} 1 \; if \; (x_{ik} \neq x_{jk}) \\ 0 \; if \; (x_{ik} = x_{jk}) \end{cases}$$

Because it simply sums mismatches, Hamming distance ranges from 0 (when the vectors are identical) to *n*, when they differ at every coordinate.

More in detail, after the feature selection, we are in the condition with $S \leq M$ attributes $X_n^{\text{red}} = \{x_n^{\text{red}}(1), \ldots, x_n^{\text{red}}(S)\}$ for each record $n \in \{1, \ldots, N\}$, where red indicates that is a reduced instance, composed of S features only. These S features have a negligible excess over independence by construction. These features can now be used to characterize the distribution of nominal data $\mathcal{H}$ and, thus, recognize instances of the minority class.

We do so by relying on the idea that the more a specific combination of the features' values appears among the majority class records, the less likely it is for a new instance characterized by the same values' combination to belong to the minority class.

Assuming that the different combinations of attributes $\chi_1, \ldots, \chi_C$ are mutually independent, it is possible to characterize the likelihood of a new record belonging to the majority class set $\mathcal{H}$ according to the following result.

Let $X_n^{\text{red}} = \{x_n^{\text{red}}(1), \ldots, x_n^{\text{red}}(S)\}$ be an unseen (and unlabelled) record from a set of N_{new} records, where $new \in \{1, \ldots, N_{\text{new}}\}$, which we have to assign to the majority class $\mathcal{H}$ or minority class $\mathcal{U}$ set.

$$\mathcal{L}(X_{\text{new}}^{\text{red}} \in \mathcal{H}) = p(_{\text{new}}^{\text{red}} \in \mathcal{H}|\mathcal{H}) = \sum_{c=1}^{C} p(X_{\text{new}}^{\text{red}} \in \mathcal{H}|\chi_c, \mathcal{H})\, p(\chi_c|\mathcal{H})$$

where $p(X_{\text{new}}^{\text{red}} \in \mathcal{H}|\chi_c, \mathcal{H})$ is the probability that the new record belongs to the majority class distribution given the c-th combination of attributes we have observed in it and $p(\chi_c|\mathcal{H})$ is the probability of observing such a combination in the majority class data, with $c = 1, \ldots, C$.

At this point is then possible to infer the likelihood of a new record to belong to the majority class and conversely its risk of being attributed to the minority class as

$$\mathcal{R}\left(X_{\text{new}}^{\text{red}}\right) = 1 - \mathcal{L}(X_{\text{new}}^{\text{red}} \in \mathcal{H}), \qquad with\ \mathcal{R}\left(X_{\text{new}}^{\text{red}}\right) \in [0, 1]$$

Interpreting the former as an indicator of the similarity between the new record and the combinations already observed in H, we propose to approximate $p(X_{\text{new}}^{\text{red}} \in \mathcal{H}|\chi_c, \mathcal{H})$ via the inverse Hamming distance between the new record $X_{\text{new}}^{\text{red}}$ and the observed combination χ_c, namely

$$p(X_{\text{new}}^{\text{red}} \in \mathcal{H}|\chi_c, \mathcal{H}) \approx \hat{p}\,(X_{\text{new}}^{\text{red}} \in \mathcal{H}|\chi_c, \mathcal{H}) = \frac{1}{S}\sum_{i=1}^{S} \mathbb{I}\left(x_{\text{new}}^{\text{red}}(i) = \chi_c(i)\right)$$

where $\mathbb{I}\left(x_{\text{new}}^{\text{red}}(i) = \chi_c(i)\right)$ is the indicator function defined as

$$\mathbb{I}\left(x_{\text{new}}^{\text{red}}(i) = \chi_c(i)\right) = \begin{cases} 1 & if \quad x_{\text{new}}^{\text{red}}(i) = \chi_c(i) \\ 0 & otherwise \end{cases}$$

Instead, $p(\chi_c|\mathcal{H})$ is given by

$$p(\chi_c|\mathcal{H}) = p\big(\chi_c(x^{\text{red}}(1) = \chi_c(1),\ldots,x^{\text{red}}(S) = \chi_c(S)\big), \quad \forall c = 1,\ldots,C,$$

With $\sum_{c=1}^{C} p(\chi_c|\mathcal{H}) = 1$, and it provides an indication of how common χ_c is in the majority class dataset. Accordingly, $p(\chi_c|\mathcal{H})$ can be approximated by computing the relative frequency of each combination χ_c in the majority class dataset, namely

$$p(\chi_c|\mathcal{H}) \approx \hat{p}(\chi_c|\mathcal{H}) = \frac{1}{H}\sum_{h=1}^{H} \mathbb{I}\big(\, x_h^{\text{red}}(1) = \chi_c(1),\ldots,\, x_h^{\text{red}}(S) = \chi_c(S)\big)$$

As mentioned before this one-class classification strategy is straightforwardly generalizable to a general case and, thus, usable independently from the previous feature selection approach.

- **Feature ranking**

To unveil the importance of categorical features in the presence of considerably unbalanced categorical data, Jessica et al. proposed to score (and sort) features based on their relevance to characterize the majority class distribution.

The method ranks features based on a tailored definition of importance, stating that a feature ─ or a features set ─ is more important if it consistently exhibits the same value in majority-class data. The algorithm assume that the "a priori" probability of indicating a considered subset of features as relevant to characterize the majority-class distribution is constant and equal across all possible sets of selected features.

The probability of a subset $\tilde{S} \subseteq S$ of attributes (among the eventually selected features) to characterize the majority class distribution is defined as

$$\text{logit}(\phi(\tilde{S})) = \log\left(\frac{\phi(\tilde{S})}{1-\phi(\tilde{S})}\right) \in (-\infty,\infty)$$

A value of values of $\text{logit}(\phi(\tilde{S}))$ being closer to 1 indicate an higher the probability of $\tilde{S}$ to be relevant to characterize the majority class distribution.

Assuming that the probability $\gamma(\tilde{S})$ ─ i.e. the probability of indicating the considered subset of features as relevant to characterize the majority class distribution ─ is constant and equal across all possible sets of selected features $\tilde{S} \subseteq S$, the feature importance can be evaluated by

$$\Omega(\tilde{S}) = \log\left(\frac{\theta_1(\tilde{S})}{\theta_2(\tilde{S})}\right)$$

where $\theta_1(\tilde{S})$ is the probability of indicating the set of features as relevant, provided that they are actually important to characterize the majority class distribution and $\theta_2(\tilde{S})$ is the probability of indicating the $\tilde{S}$ as relevant even if they are not important to characterize the majority class distribution (i.e. a sort of false positive).

The importance estimation of each feature is obtained by iteratively adding attributes to $\tilde{S}$ and quantifying their impact on the characterization of the variability of entropy $\mathcal{H}$ through $\Omega(\tilde{S})$. This happens through the construction of a tree by incrementally considering an increasing number of features starting from the root level ($k = 0$), where no feature is included and $\Omega(\tilde{S}) = +\infty$, to the leaves, computed with all attributes in S but one. There's no inclusion in the tree leaves computed when $S = \tilde{S}$ to avoid $\Omega(\tilde{S}) = -\infty$. Therefore, at level $k = 1$, each branch from the root corresponds to the inclusion of a single feature in $\tilde{S}$ (with the nodes of the tree embedding the value of $\Omega(\tilde{S})$. We then keep iteratively creating new branches, one for each feature not yet included in the subset, until we arrive at the trees leaves.

Once constructed the tree and, thus, explored all possible feature combinations, the problem

$$\delta *_{r\rightarrow} = arg \min_{\delta_{r\rightarrow}} \sum_{k=1}^{K} \Omega_{r\rightarrow}^{k}$$

where K represents the level of the leaves, $\delta_{r\rightarrow}$ represent any path over the tree starting from the root and $\Omega_{r\rightarrow}^{k}$ represents the indicator for the path $\delta_{r\rightarrow}$ up to the k-th level of the tree with dynamic programming. Following this path, the features encountered first are thus the ones deemed most important, while those encountered last are the least important. Indeed, the attributes closer to the root are the ones that predominantly cause an increase in variability. In essence, the selected path is the one in which the incremental increase in entropy from adding each variable is minimized. Variables that cause greater increases in entropy when introduced earlier are considered less relevant for characterizing the majority class.

Since the proposed model (OC-Cat) does not account for within-subject variance when multiple catheters are placed, we included only the first catheter insertion at "Luigi Sacco" Hospital.

To apply the OC-Cat algorithm, we have to discretize continuous variables. We decided to create a 4 classes feature for age $-\ x < 65\ anni;\ 65 \leq x < 75;\ 75 \leq x < 85\ and\ \geq 85\ -$ and 3 classes for age-adjusted Charlson Comorbidity Index (CCI) $-\ x < 3;\ 3 \leq x < 5;\ 5 \leq x < 7\ and\ \geq 7\ -$ according to data distribution and clinical relevance.

The dataset was split into a training (~ 75%) and testing set (~ 25%) using a temporal cut-off at August 2023. Training data was majority-class imbalanced (uninfected patients dominant). We intentionally applied this stratified approach even to Isolation Forest, despite its ability to handle both classes without separation.

The analysis comprised these main steps:

1. Assess feature redundancy using the excess-over-independence metric, i.e. estimating the likelihood of pairwise correlations.
2. Construct a weighted graph in which vertices represent the original features and edges are weighted by the excess-over-independence scores.
3. Apply the Bellman–Ford algorithm to compute shortest paths from each feature to all others, then select the features on the path yielding the highest Bayesian Information Criterion (BIC). To ensure robust feature selection, we applied a 5-fold cross-validation on the training data (i.e. "uninfected" records until August 2023), selecting features in each fold; only features selected at least three times are retained.
4. Train the OC-Cat model exclusively on uninfected patients through August 2023, thereby defining the "normality" class.
5. For every patient record excluded from that training set, predict the probability of membership in the uninfected class. Predictions are generated separately for (a) patients infected before August 2023 and (b) patients never infected or infected on/after that date.
6. Repeat steps 4 and 5 using the iForest and OCSVM algorithms (implemented via Python's scikit-learn and isolationForest packages with default settings).
7. Compare the OC-Cat's predictions against iForest and OCSVM via Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC), Positive Predictive Value (PPV), Negative Predictive Value (NPV), Precision-Recall curves with corresponding PRAUC curves and F1 Score curves across varying

thresholds. All raw model scores are scaled (z-standardized to mean 0 and standard deviation 1) to ensure comparability.

8. Visualizations of the three scaled model score distributions, comparing them by the Mann-Whitney U test and estimate the Hellinger distance between them.
9. Conduct a sensitivity analysis by repeating steps from 4 to 8, taking into account all of the initial features in the dataset for which data are not missing.
10. Conduct a sensitivity analysis by repeating steps from 4 to 8, restricting the dataset to vascular access devices with indwelling times ≥ 7, 14, or 21 days (excluding any removed beforehand). Define infections as those occurring within each corresponding window.
11. Perform final feature selection on the model of interest using the subset of features identified in the preceding steps and all features for sensitivity analysis.

3.5. Results

Overall, the dataset included 2,923 observations (i.e., catheter placements) corresponding to 2,338 individual patients. For the analysis, only records related to the first VAD placement during the first hospitalization were considered. Additionally, patients treated under a day-hospital regime were excluded. As a result, the number of patients included in the final dataset was reduced from 2,338 to 2,120 (Fig. 23). The group variable of interest was the occurrence of a catheter infection during Vascular Access Device (VAD) placement. Among considered observations, 79 (3.7%) CABSI and 57 (2.7%) CRBSI occurred. The dataset is balanced for biological sex: 1,102 (52.0%) female and 1,018 (48.0%) male patients (Table 3).

Duration of VAD placement was higher for infected patients (median=14; 1^{st}-3^{rd} = 8-22 days) compared to healthy (median = 12; 1^{st}-3^{rd} = 7-20 days; *Mann-Whitney U test* = 0.098) (Fig. 24). There also a difference between CRBSI infected patients (median=15; 1^{st}-3^{rd} = 8-19 days) compared to CABSI ones (median=12; 1^{st}-3^{rd} = 7-22 days).

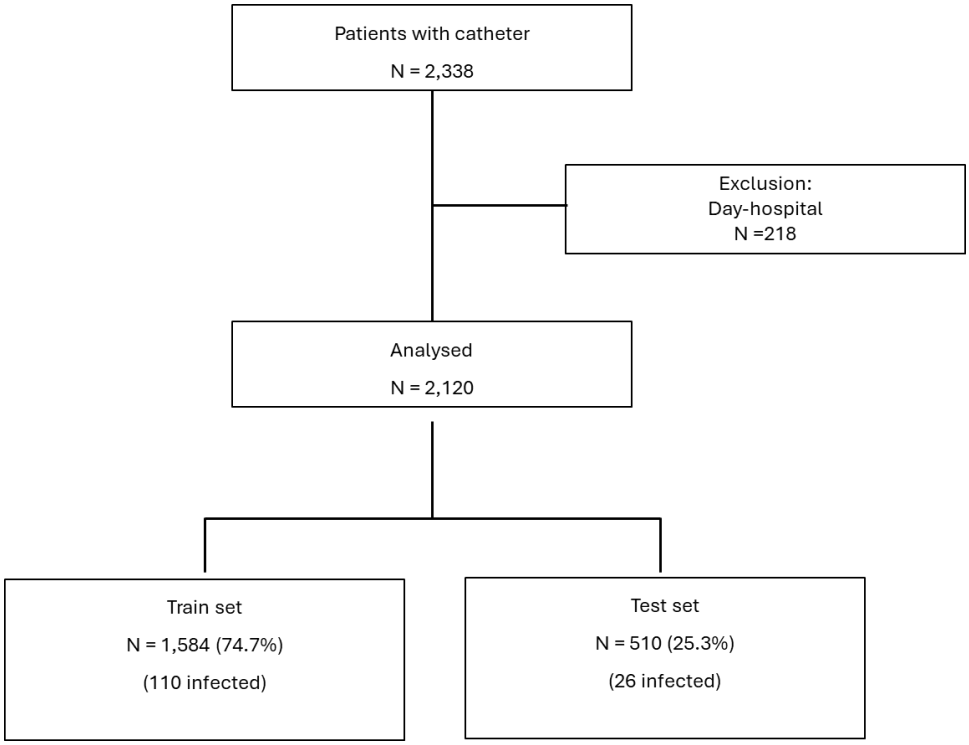


Figure 23. Study flowchart showing cohort selection and train-test split.

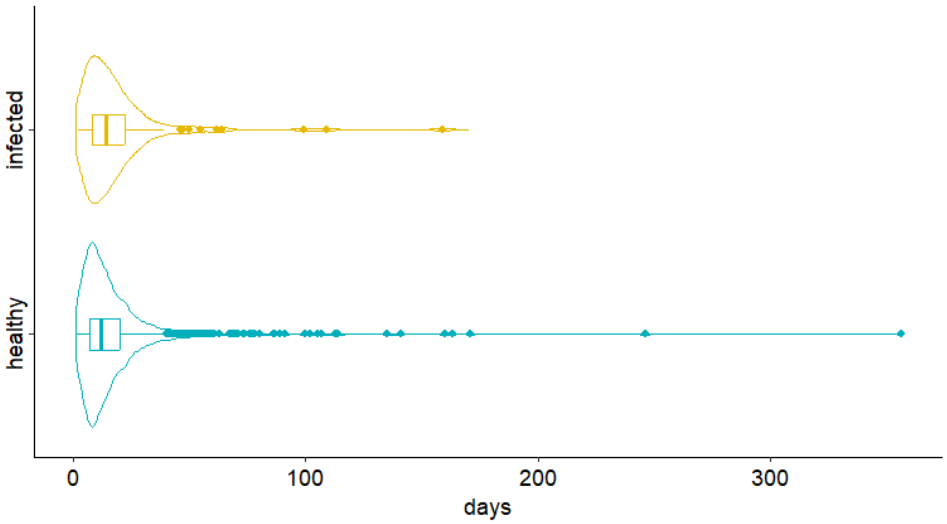


Figure 24. Violin boxplot showing VAD placement duration among healthy and infected patients (CRBSI + CABSI).

	No infection (N=1984)	Infected (N=136)	CABSI (N=79)	CRBSI (N=57)	Overall (N=2120)
Biological sex					
Male	952 (48.0%)	66 (48.5%)	40 (50.6%)	26 (45.6%)	1018 (48.0%)
Female	1032 (52.0%)	70 (51.5%)	39 (49.4%)	31 (54.4%)	1102 (52.0%)
age					
<65	490 (24.7%)	37 (27.2%)	20 (25.3%)	17 (29.8%)	527 (24.9%)
[65; 75)	345 (17.4%)	31 (22.8%)	18 (22.8%)	13 (22.8%)	376 (17.7%)
[75; 85)	595 (30.0%)	36 (26.5%)	23 (29.1%)	13 (22.8%)	631 (29.8%)
≥85	554 (27.9%)	32 (23.5%)	18 (22.8%)	14 (24.6%)	586 (27.6%)
Diabetes [No/Yes]					
Yes	525 (26.5%)	36 (26.5%)	17 (21.5%)	19 (33.3%)	561 (26.5%)
Cardiovascular disease [No/Yes]					
Yes	1004 (50.6%)	51 (37.5%)	25 (31.6%)	26 (45.6%)	1055 (49.8%)
Lung disease [No/Yes]					
Yes	589 (29.7%)	34 (25.0%)	18 (22.8%)	16 (28.1%)	623 (29.4%)
Neurological disease [No/Yes]					
Yes	717 (36.1%)	54 (39.7%)	31 (39.2%)	23 (40.4%)	771 (36.4%)
Liver disease [No/Yes]					
Yes	83 (4.2%)	7 (5.1%)	3 (3.8%)	4 (7.0%)	90 (4.2%)
Chronic kidney disease [No/Yes]					
Yes	375 (18.9%)	14 (10.3%)	6 (7.6%)	8 (14.0%)	389 (18.3%)
Autoimmune disease [No/Yes]					
Yes	215 (10.8%)	18 (13.2%)	11 (13.9%)	7 (12.3%)	233 (11.0%)
Hypertension [No/Yes]					
Yes	1158 (58.4%)	88 (64.7%)	51 (64.6%)	37 (64.9%)	1246 (58.8%)
Psychiatric disease [No/Yes]					
Yes	231 (11.6%)	16 (11.8%)	8 (10.1%)	8 (14.0%)	247 (11.7%)
BMI [<30/≥30]					
≥30	224 (11.3%)	12 (8.8%)	10 (12.7%)	2 (3.5%)	236 (11.1%)
Drug abuser [No/Yes]					
Yes	38 (1.9%)	4 (2.9%)	4 (5.1%)	0 (0%)	42 (2.0%)
aCCI					
<3	288 (14.5%)	18 (13.2%)	9 (11.4%)	9 (15.8%)	306 (14.4%)
[3; 5)	414 (20.9%)	37 (27.2%)	25 (31.6%)	12 (21.1%)	451 (21.3%)
[5; 7)	599 (30.2%)	41 (30.1%)	26 (32.9%)	15 (26.3%)	640 (30.2%)
≥7	683 (34.4%)	40 (29.4%)	19 (24.1%)	21 (36.8%)	723 (34.1%)
Previous infections (30 days) [No/Yes]					
Yes	73 (3.7%)	13 (9.6%)	9 (11.4%)	4 (7.0%)	86 (4.1%)
From ICU [No/Yes]					
Yes	68 (3.4%)	8 (5.9%)	4 (5.1%)	4 (7.0%)	76 (3.6%)
Ward [No/Yes]					
Clinical	1796 (90.5%)	112 (82.4%)	66 (83.5%)	46 (80.7%)	1908 (90.0%)
Surgical	188 (9.5%)	24 (17.6%)	13 (16.5%)	11 (19.3%)	212 (10.0%)

	No infection (N=1984)	Infected (N=136)	CABSI (N=79)	CRBSI (N=57)	Overall (N=2120)
Covid-19 [No/Yes]					
Yes	352 (17.7%)	28 (20.6%)	16 (20.3%)	12 (21.1%)	380 (17.9%)
Covid-associated pneumonia [No/Yes]					
Yes	247 (12.4%)	12 (8.8%)	8 (10.1%)	4 (7.0%)	259 (12.2%)
Tracheostomized [No/Yes]					
Yes	39 (2.0%)	8 (5.9%)	3 (3.8%)	5 (8.8%)	47 (2.2%)
Cachexia [No/Yes]					
Yes	113 (5.7%)	8 (5.9%)	5 (6.3%)	3 (5.3%)	121 (5.7%)
Parenteral nutrition [No/Yes]					
Yes	364 (18.3%)	62 (45.6%)	27 (34.2%)	35 (61.4%)	426 (20.1%)
Chemotherapy [No/Yes]					
Yes	105 (5.3%)	10 (7.4%)	8 (10.1%)	2 (3.5%)	115 (5.4%)
Immunosuppressor [No/Yes]					
Yes	432 (21.8%)	28 (20.6%)	18 (22.8%)	10 (17.5%)	460 (21.7%)
Type of VAD					
Peripheral	1349 (68.0%)	67 (49.3%)	47 (59.5%)	20 (35.1%)	1416 (66.8%)
Central	635 (32.0%)	69 (50.7%)	32 (40.5%)	37 (64.9%)	704 (33.2%)
Number of lumen [1/>1]					
>1	251 (12.7%)	34 (25.0%)	10 (12.7%)	24 (42.1%)	285 (13.4%)
Tunnelized [No/Yes]					
Yes	179 (9.0%)	12 (8.8%)	10 (12.7%)	2 (3.5%)	191 (9.0%)
Position site					
Upper body	57 (2.9%)	3 (2.2%)	2 (2.5%)	1 (1.8%)	60 (2.8%)
Head/neck	853 (43.0%)	65 (47.8%)	41 (51.9%)	24 (42.1%)	918 (43.3%)
Lower body	1074 (54.1%)	68 (50.0%)	36 (45.6%)	32 (56.1%)	1142 (53.9%)
Securement system					
Adhesive	648 (32.7%)	61 (44.9%)	33 (41.8%)	28 (49.1%)	709 (33.4%)
Subcutaneous	1336 (67.3%)	75 (55.1%)	46 (58.2%)	29 (50.9%)	1411 (66.6%)
Number of attempts [1/>1]					
>1	352 (17.7%)	19 (14.0%)	10 (12.7%)	9 (15.8%)	371 (17.5%)
Issue during VAD insertion [No/Yes]					
Yes	25 (1.3%)	0 (0%)	0 (0%)	0 (0%)	25 (1.2%)
Immunocompromised [No/Yes]					
Yes	296 (14.9%)	25 (18.4%)	14 (17.7%)	11 (19.3%)	321 (15.1%)
Neoplasm [No/Yes]					
Yes	495 (24.9%)	33 (24.3%)	18 (22.8%)	15 (26.3%)	528 (24.9%)

Table 3. Distribution of features among healthy and infected patients (CRBSI and CABSI). The 4th and 5th columns detail the distributions for CRBSI and CABSI patients, respectively, while the 6th column shows the distribution across the entire dataset.

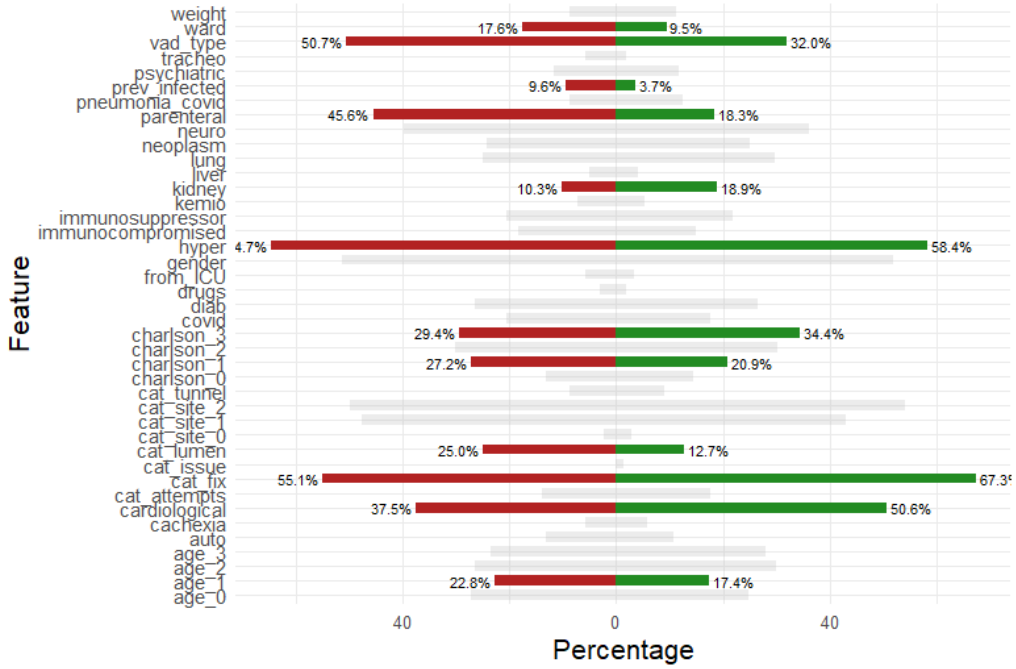


Figure 25. Graphical representation of features distribution among healthy (green) and infected patients (red). Features highlighted in the plot are those with an absolute difference greater than 5 percentage points between the two groups. (multi-class) categorical variables were transformed into binary indicator variables via one-hot encoding.

Among all the variables, those that differed at least five percentage points between healthy and infected were cardiovascular disease (50.6% *vs* 37.5%, χ^2=0.004), hypertension (58.4% *vs* 64.7%, χ^2=0.173), chronic kidney disease (18.9% *vs* 10.3%, χ^2=0.017), age-adjusted Charlson index from 3 to 5 and ≥ 7 (charlson_1: 20.9% *vs* 27.2%, χ^2=0.101; charlson_3: 34.4% *vs* 29.4%, χ^2=0.272), age among 65 and 75 years old (age_1: 17.4% *vs* 22.8%, χ^2=0.139), parenteral nutrition (18.3% *vs* 45.6%, χ^2<0.001), type of ward (9.5% *vs* 17.6%, χ^2=0.003), securement system (cat_fix: 67.3% *vs* 55.1%; χ^2=0.005), number of lumen (cat_lumen: 12.7% *vs* 25.0%, χ^2<0.001), type of VAD (32.0% *vs* 50.7%, χ^2<0.001) and infection in the previous 30 days (3.7% *vs* 9.6%, χ^2=0.002) (Fig.25).

The internal validation was performed applying a splitting of the dataset at the date of August 2023. The train dataset consisted of about 75% of the entire dataset. In the test dataset there are a major proportional quote of CRBSI among all infections compared to the train dataset (Table 4). The catheter insertions that resulted in an infection accounted for 6.9% (110/1584) of the training set and 4.9% (26/536) of the test set.

Full dataset (main analysis)

	No infection (N=1984)	Infected (N=136)	CABSI (N=79)	CRBSI (N=57)	Overall (N=2120)
Train	1474 (74.3%)	110 (80.9%)	66 (83.5%)	44 (77.2%)	1584 (74.7%)
Test	510 (25.7%)	26 (19.1%)	13 (16.5%)	13 (22.8%)	536 (25.3%)

Table 4. Distribution of subjects among training and test sets, stratified by infection status.

Following the dataset split, the distribution of features between healthy and infected subjects differs across the two subsets. In particular, features with a difference of at least five percentage points between categories occur more frequently in the test set than in the training set or the overall dataset (Figs. 26, 27).

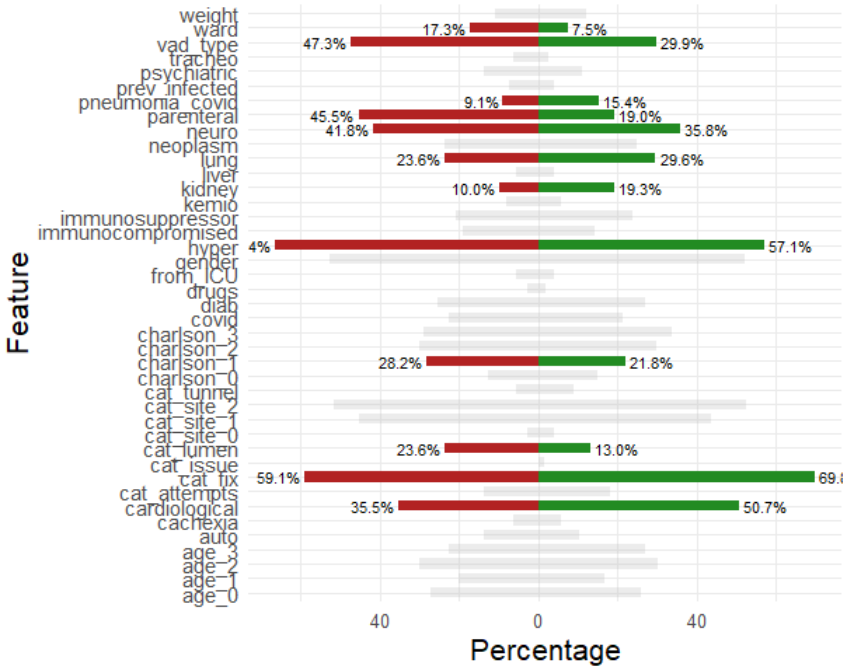


Figure 26. Graphical representation of features distribution among healthy (green) and infected patients (red) in the **train** dataset. Features highlighted in the plot are those with an absolute difference greater than 5 percentage points between the two groups. (multi-class) categorical variables were transformed into binary indicator variables via one-hot encoding.

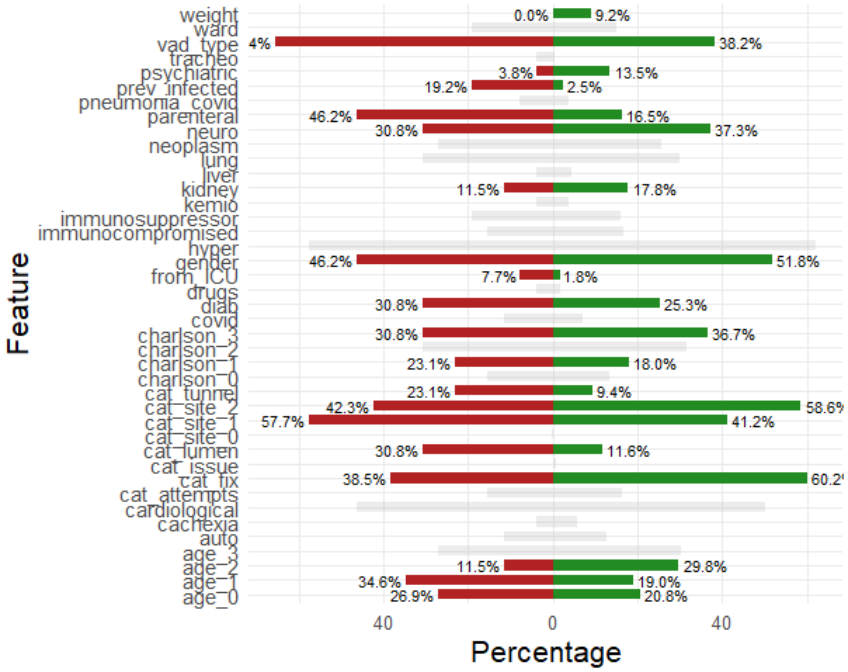


Figure 27. Graphical representation of features distribution among healthy (green) and infected patients (red) in the **test** dataset. Features highlighted in the plot are those with an absolute difference greater than 5 percentage points between the two groups. (multi-class) categorical variables were transformed into binary indicator variables via one-hot encoding.

The first phase of the analysis consisted in the construction of the Q-table using the train dataset (Fig. 28).

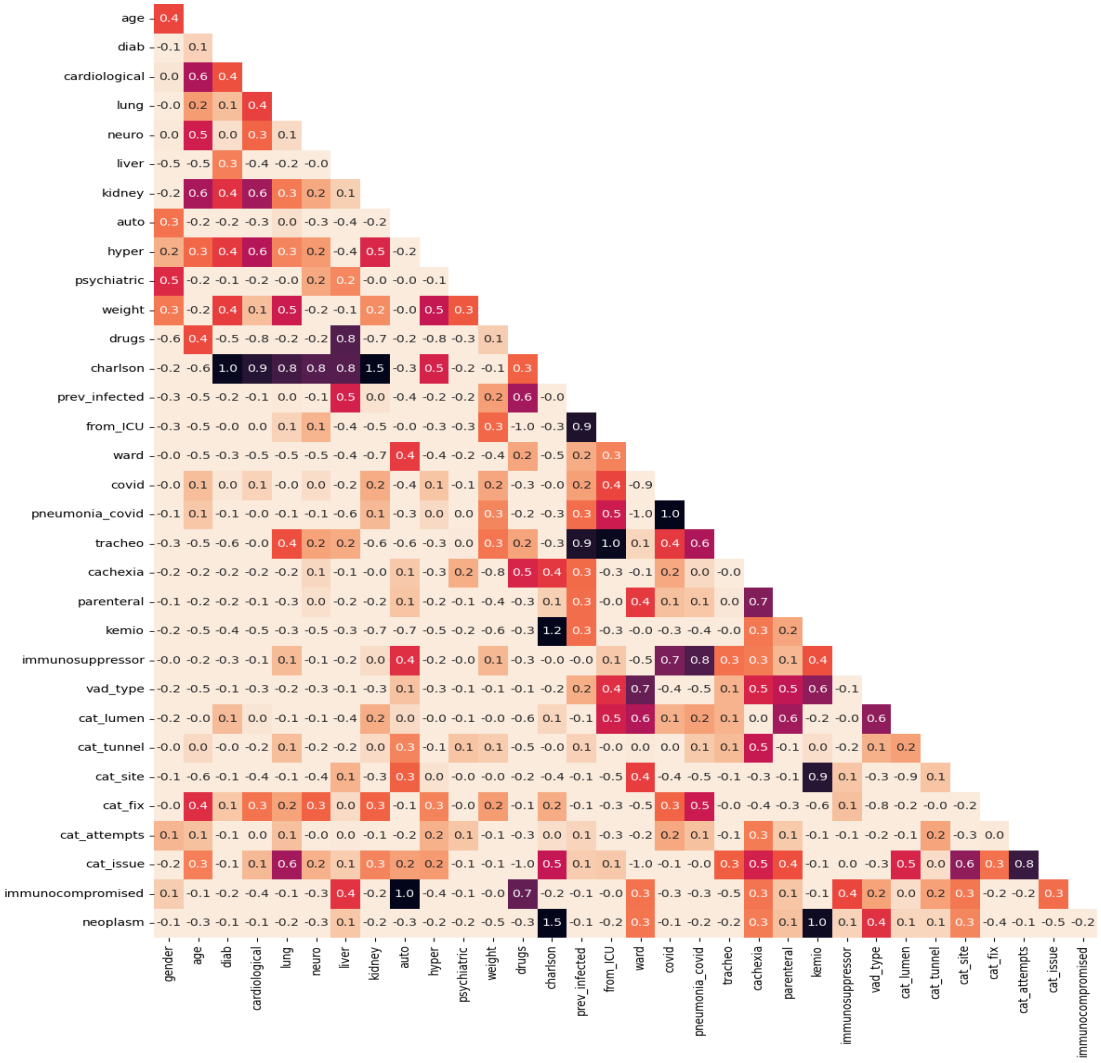


Figure 28. Q-table heatmap among all features of the train dataset. In this representation, value are not normalized between 0 and 1.

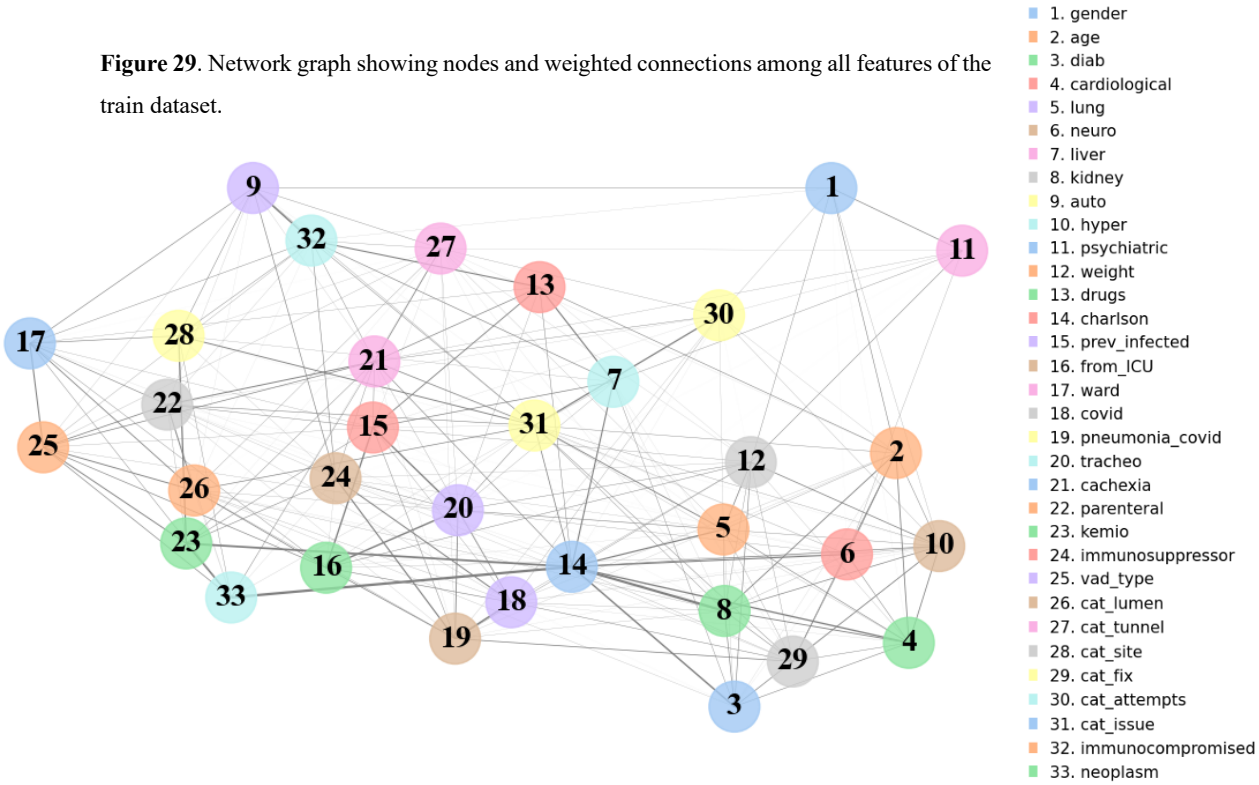
Charlson comorbidity index inherently captures diabetes as well as cardiovascular, pulmonary, neurological, hepatic, and renal diseases (so it's strongly correlated with them). Other variables in the dataset overlap:

- Drug abuse strongly correlates with liver disease.
- Immunocompromised status overlaps with autoimmune diseases and drug use.
- Neoplasm diagnosis is redundant with chemotherapy (other than CCI).
- Catheter complications and insertion attempts are highly collinear.
- VAD type is largely determined by ward assignment.
- Use of immunosuppressants parallels COVID-related coding.

- Tracheostomy status coincides almost perfectly with ICU admission.

Starting from the Q-table, we constructed a graph in which each feature is a node and the strength of their associations is encoded by edge weights (Fig. 29).

Figure 29. Network graph showing nodes and weighted connections among all features of the train dataset.



We then applied the graph-based feature-selection procedure searching for paths maximizing the Bayesian Information Criterion (BIC), with a penalty on the number of features. After this process, 15 features remained for the entire dataset (Fig. 30).

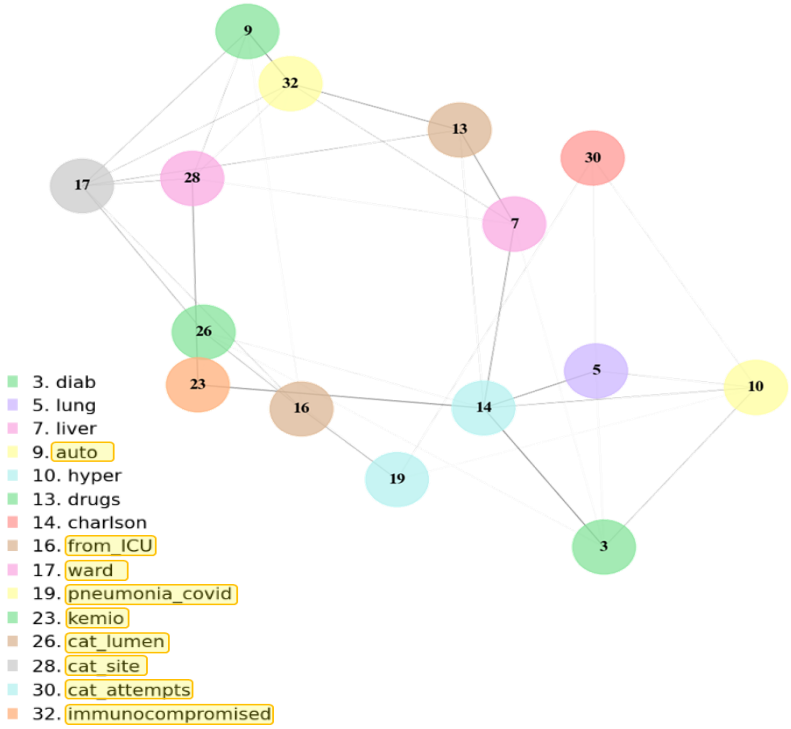


Figure 30. Network graph showing nodes and weighted connections among selected features of the train dataset.

Subsequently, we characterized the distribution of the majority class and applied a soft classifier to compute the probability that each instance belongs to it (Fig. 31).

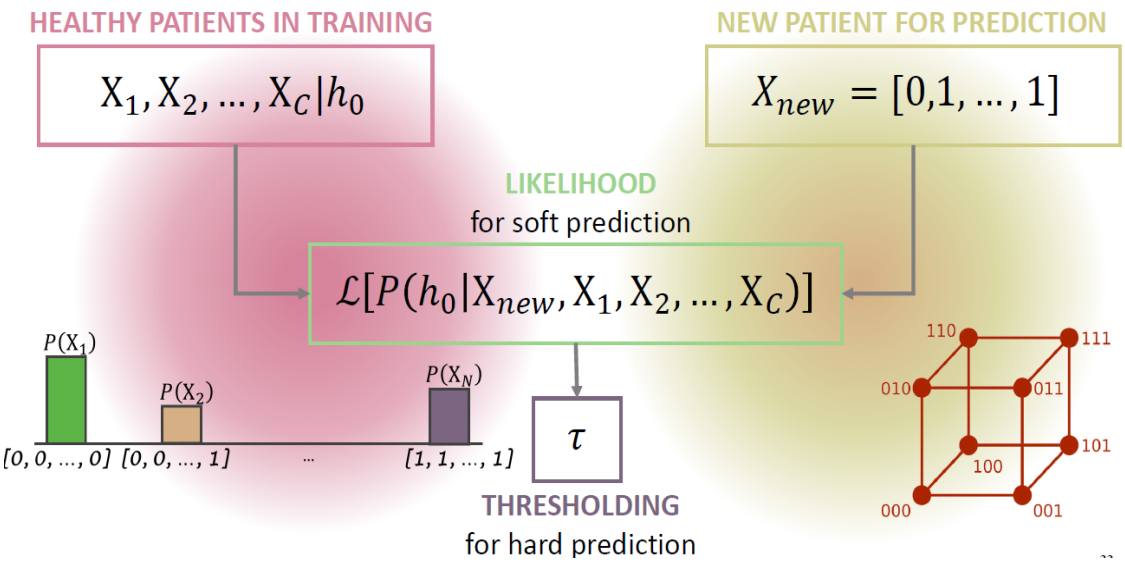


Figure 31. Graphical representation of the soft-classifier algorithm and the possible transformation to an hard-classifier.

We adopted the Isolation Forest and OCSVM models as benchmarks, training them on the subset of uninfected patients from the training dataset. Predictions were then obtained for both the training and test sets.

For the ROC analysis and prediction distribution plots, all infected patients were included, without distinguishing between infection types. In the training set, the prediction estimates indicate superior performance of the OC-Cat model compared to the two benchmark models. In contrast, the ROC curves for the test set show a greater degree of overlap among the models (Fig. 32).

The PPV and NPV curves on train set data (Fig. 33) demonstrate the superior discriminative capacity of the OC-Cat model relative to iForest and OCSVM baselines in identifying truly high-risk patients through threshold optimization. At the intermediate threshold, OC-Cat achieves about 10-20% PPV, exceeding disease prevalence (6.9%), culminating in a peak at high value of threshold. All models maintain robust NPV >90% across thresholds, ensuring safety for low-risk rule-out decisions. The values obtained on the test set are overall comparable, although they show greater overlap with those of the other two models than with the performance of the OC-Cat model (Fig. 34).

The Precision-Recall and F1-score curves further indicate that the OC-Cat model achieved superior performance compared with the other two models on the training dataset, while this superiority was less pronounced on the test dataset. These results suggest that OC-Cat seems more effective at identifying true anomalies while minimizing false positives (Fig. 35). OC-Cat achieves higher F1 scores, particularly in the middle threshold range (Fig. 36)

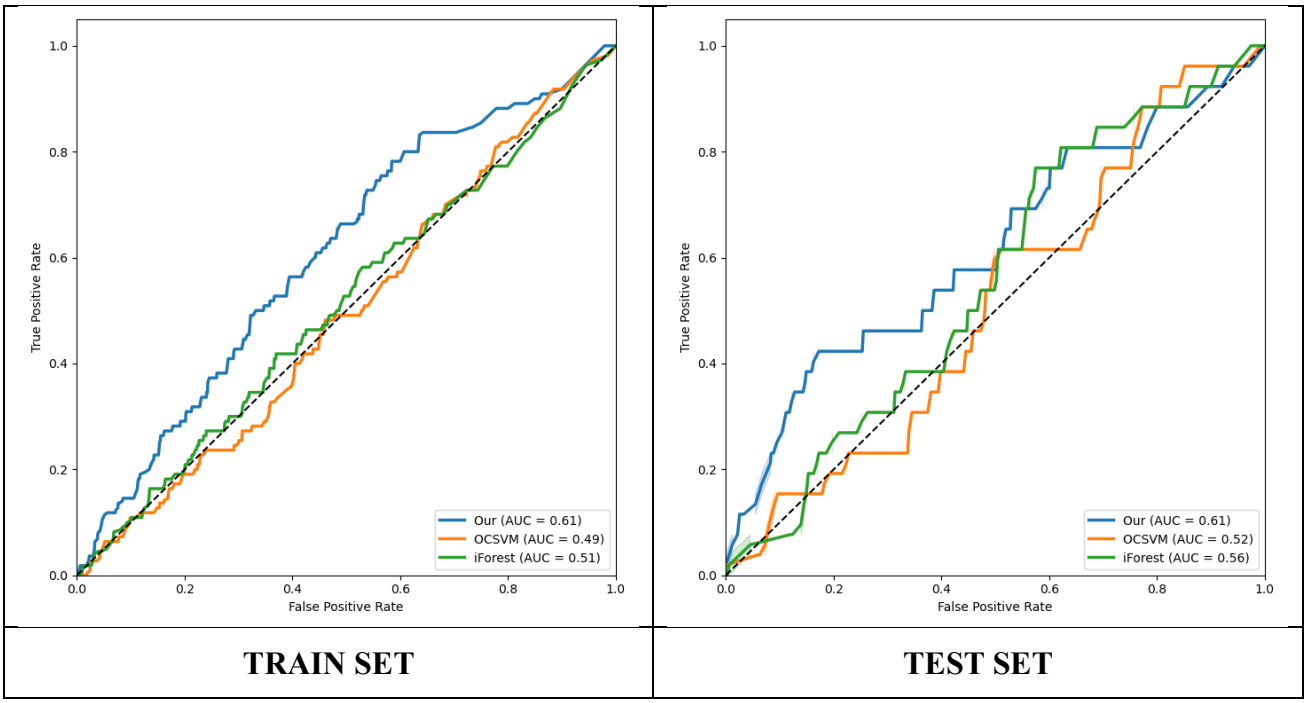


Figure 32. Receiver Operating Characteristic (ROC) curve of Train and Test sets for the three implemented models. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

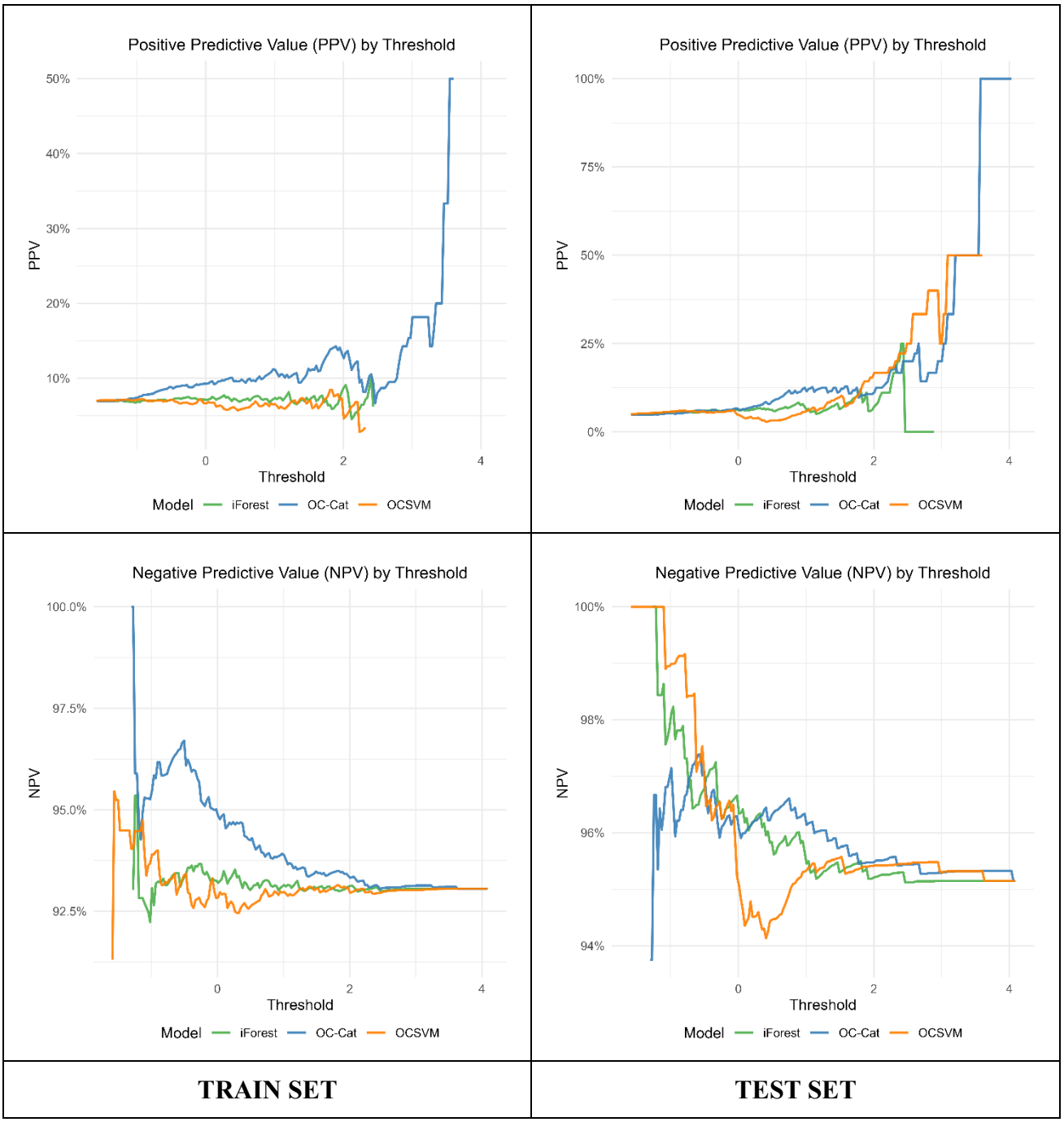


Figure 33. PPV (top) and NPV (bottom) across probability thresholds for infection risk models on Train and Test sets. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one.

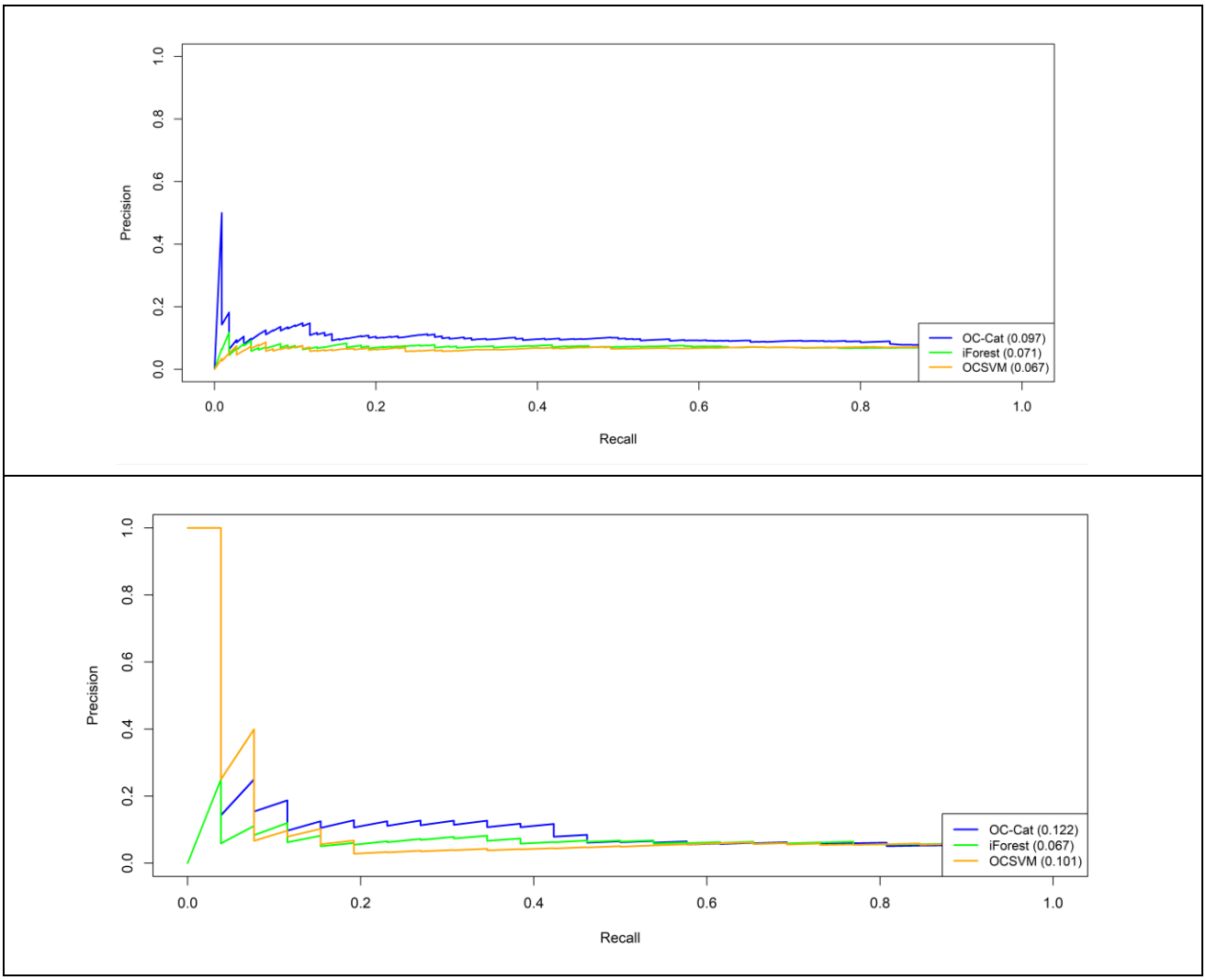


Figure 34. Precision-Recall curves on the Train and Test sets. The OC-Cat model is shown in blue, the iForest model in green, and the OCSVM model in orange. PRAUC values are reported in parentheses.

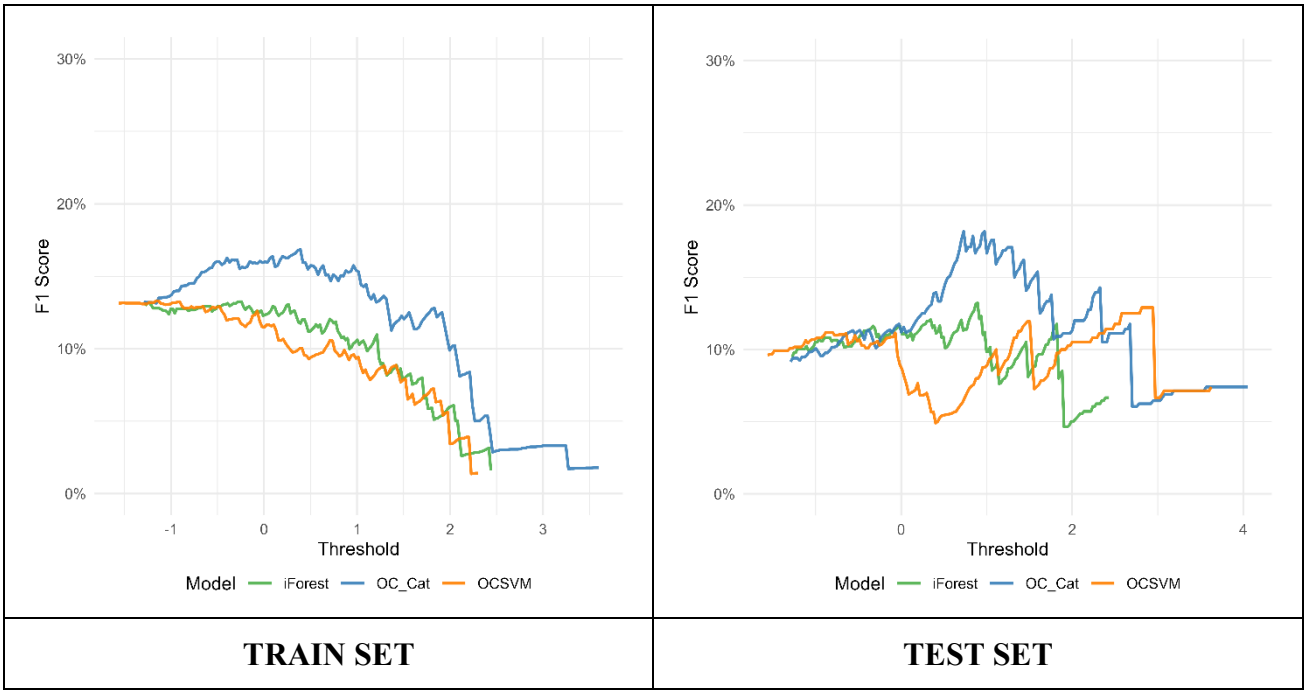


Figure 35. F1 Score curves for the Train and Test sets comparing OC-Cat (blue), iForest (green), and OCSVM (orange) across varying threshold values.

The scaled predictions of the risk scores were represented in Figures 36-38. A Mann-Whitney U test was conducted to compare the standardized predicted scores between infected and uninfected patients, separately for the training and test sets. The results indicated a significantly greater separation distributions for the OC-Cat model compared with the two benchmark models ($p < 0.05$), with higher median predicted raw scores for infected patients. The two benchmark models generally exhibited poorer predictive performance, with the exception of the OCSVM model in the test set, whose Hellinger index indicated a greater distance between distributions (Table 5).

	Mann-Whitney *p-value*		**Hellinger distance**	
	Train	**Test**	**Train**	**Test**
OC-Cat	<0.001	0.049	0.15	0.23
iForest	0.726	0.083	0.09	0.14
OCSVM	0.816	0.447	0.06	0.28

Table 5. Comparison of predicted scores between infected and uninfected patients in training and test sets, assessed using the Mann–Whitney U test *(p-values*) and Hellinger distance.

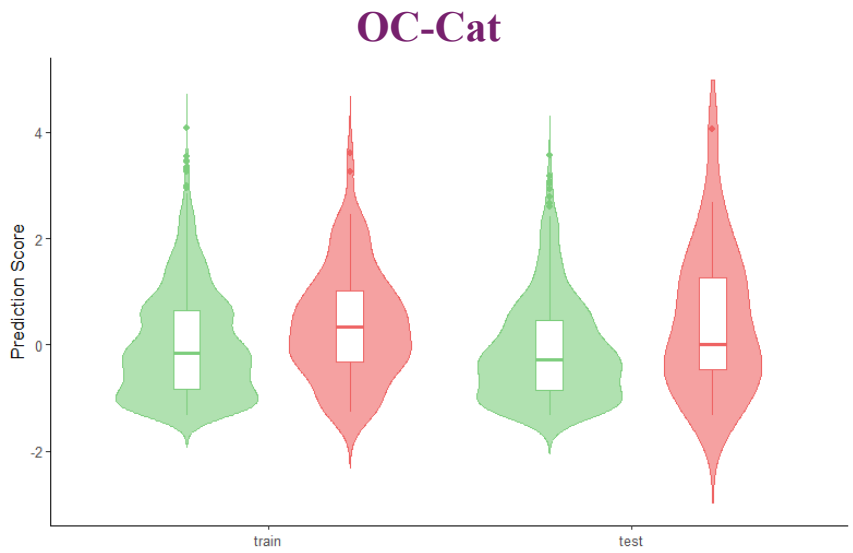


Figure 36. Violin–boxplots showing the OC-Cat model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects.

Figure 37. Violin–boxplots showing the Isolation Forest model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects.

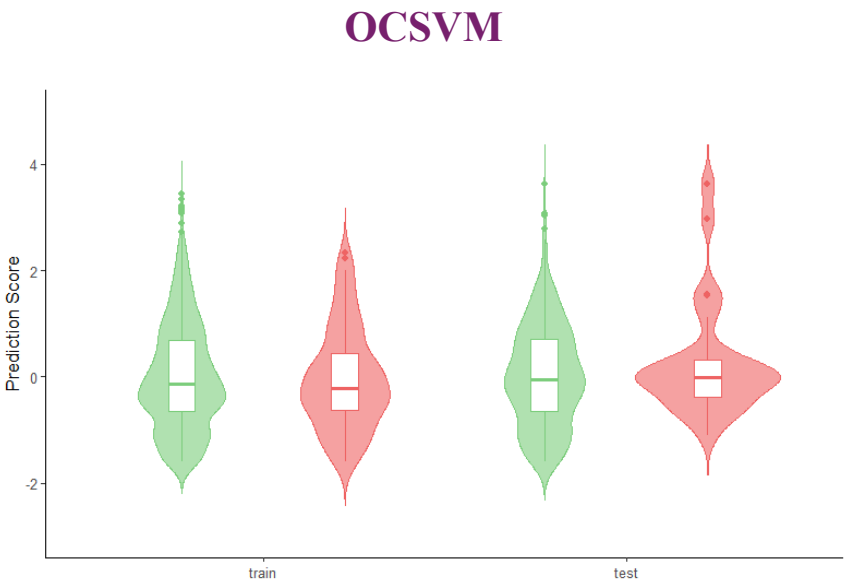


Figure 38. Violin–boxplots showing the OCSVM model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects.

The performance and predictions of the three models were also evaluated separately for CRBSI cases (Figs. 39-45) and CABSI infections (Figs. 46-52) and the resulting score distributions were compared using the Mann–Whitney U test and the Hellinger distance (Tables 6 and 7). When focusing exclusively on CRBSI cases, both the predictions and the comparative curves consistently indicated a superior performance of the OC-Cat model relative to the benchmark methods. The overall performance on the test set was lower, mirroring the trend observed in the broader analysis.

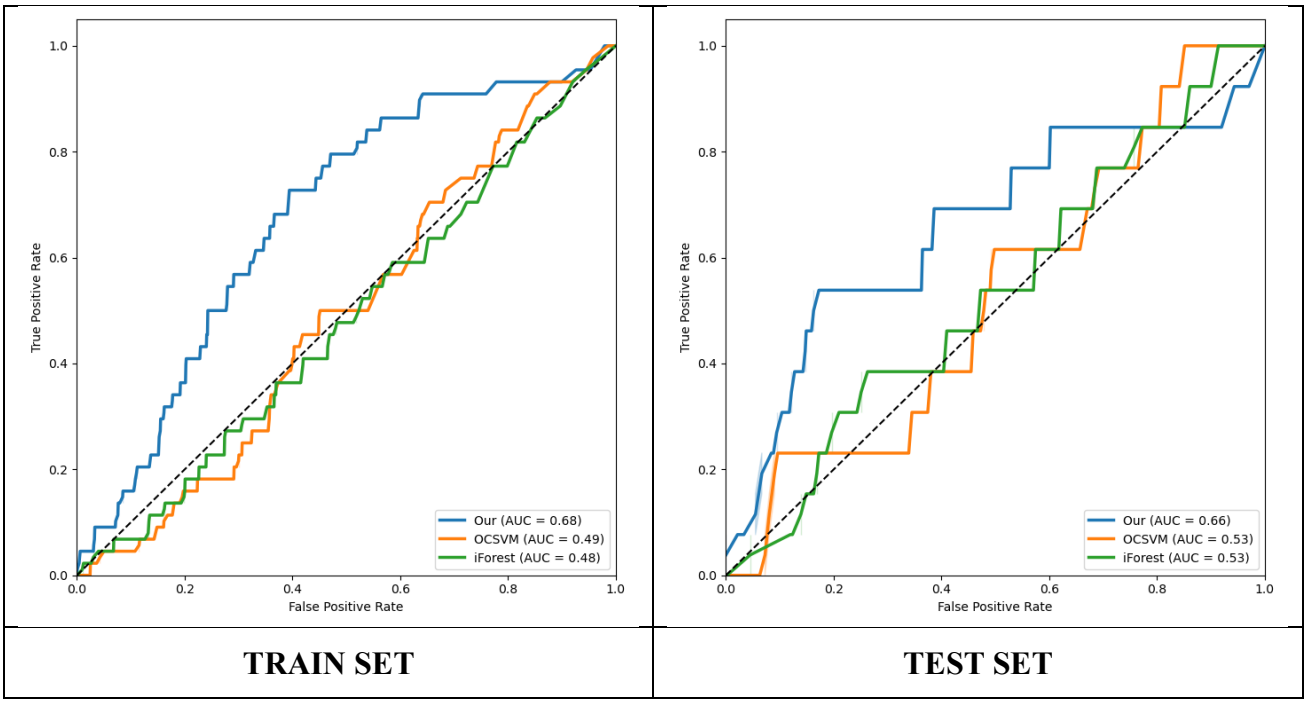


Figure 39. Receiver Operating Characteristic (ROC) curve of Train and Test sets for the three implemented models. Only CRBSI are considered, excluding CABSI events from predictions. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

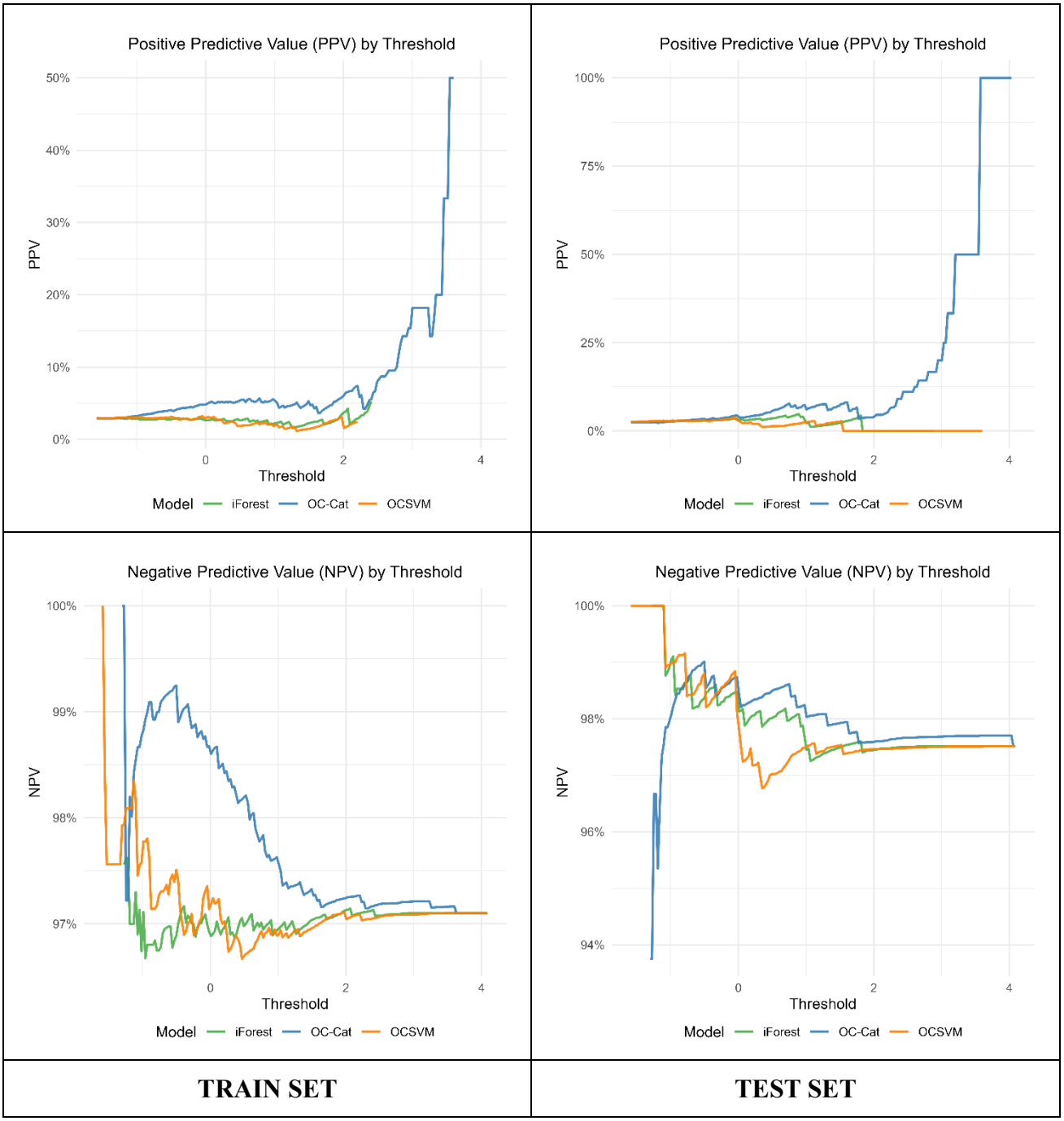


Figure 40. PPV and NPV across probability thresholds for infection risk models on Train and Test sets. Only CRBSI are considered, excluding CABSI events from predictions. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one.

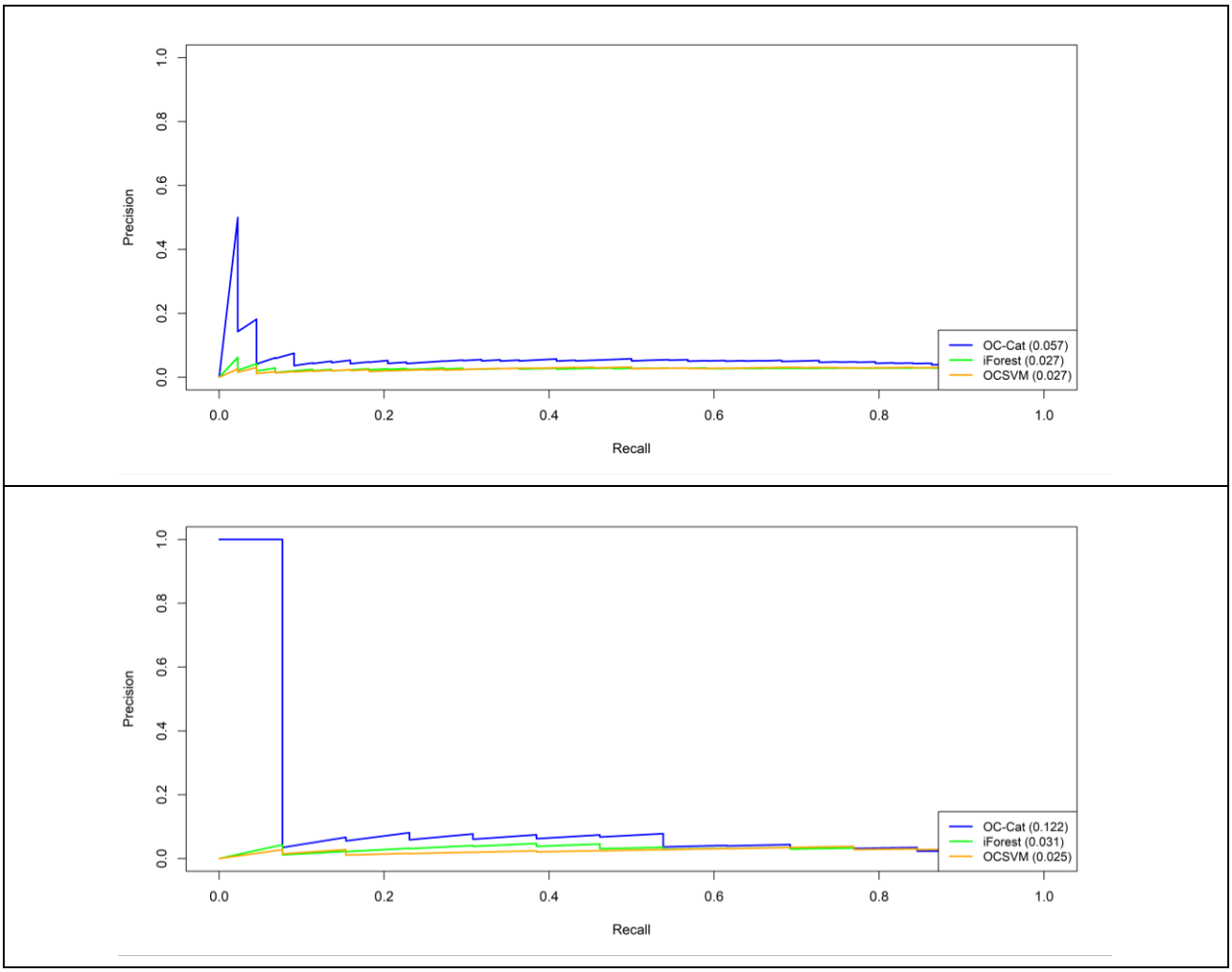


Figure 41. Precision-Recall curves on the Train and Test sets. Only CRBSI are considered, excluding CABSI events from predictions. The OC-Cat model is shown in blue, the iForest model in green, and the OCSVM model in orange. PRAUC values are reported in parentheses.

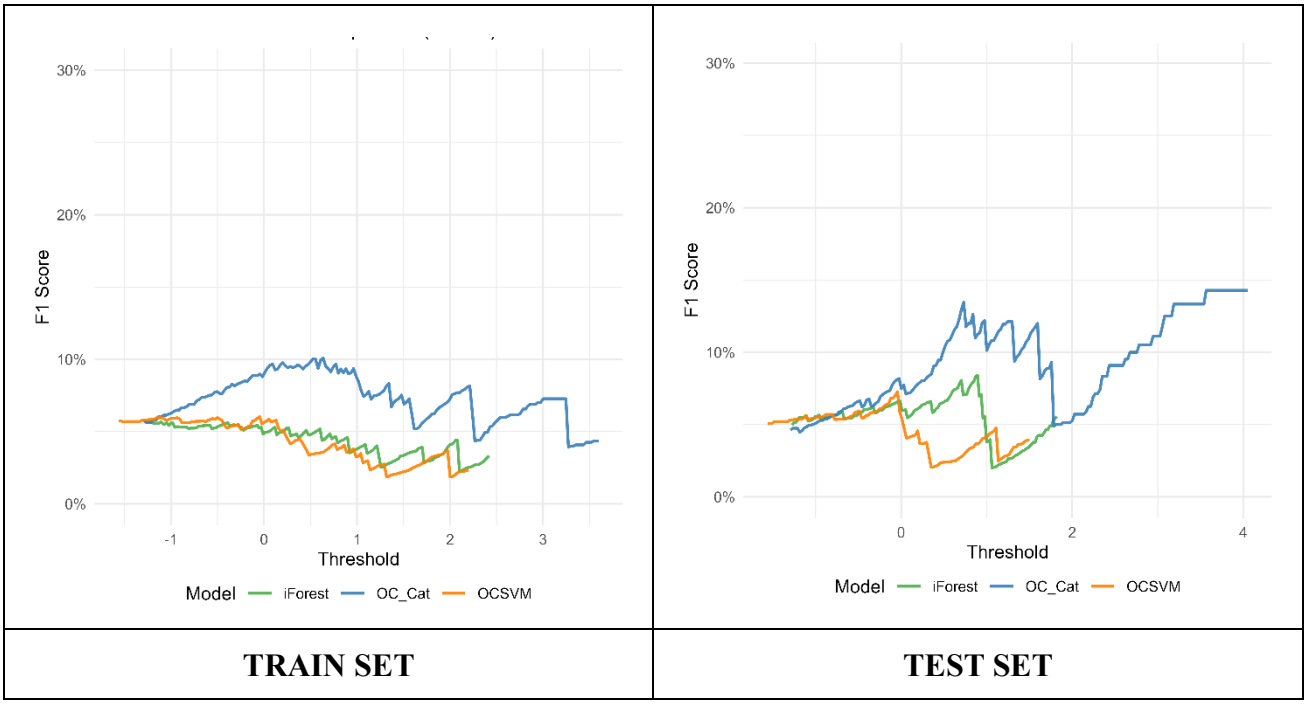


Figure 42. F1 Score curves for the Train and Test sets comparing OC-Cat (blue), iForest (green), and OCSVM (orange) across varying threshold values. Only CRBSI are considered, excluding CABSI events from predictions.

	Mann-Whitney *p-value*		Hellinger distance	
	Train	**Test**	**Train**	**Test**
OC-Cat	<0.001	0.051	0.25	0.30
iForest	0.651	0.223	0.12	0.14
OCSVM	0.853	0.643	0.09	0.34

Table 6. Comparison of predicted scores between infected and uninfected patients in training and test sets, assessed using the Mann–Whitney U test *(p-values*) and Hellinger distance. Only CRBSI are considered, excluding CABSI events from predictions.

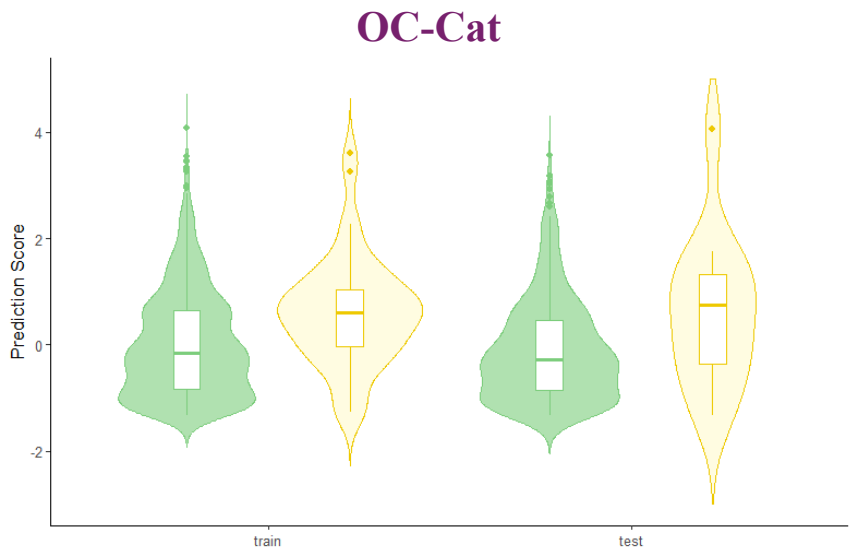


Figure 43. Violin–boxplots of OC-Cat model predicted scores for the training and test sets. Green violins denote predictions for uninfected subjects, while light yellow violins denote predictions for CRBSI infected subjects (CABSI infections were excluded from this analysis).

Figure 44. Violin–boxplots of iForest model predicted scores for the training and test sets. Green violins denote predictions for uninfected subjects, while light yellow violins denote predictions for CRBSI infected subjects (CABSI infections were excluded from this analysis).

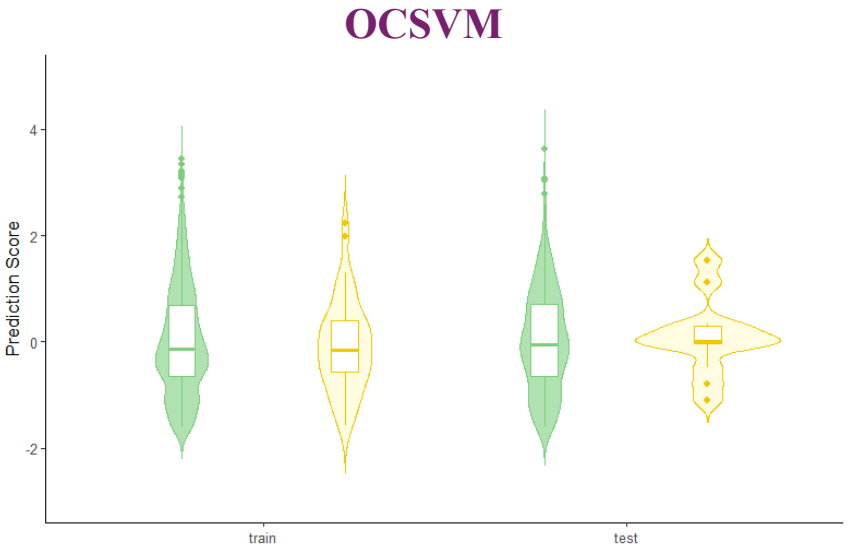


Figure 45. Violin–boxplots of OCSVM model predicted scores for the training and test sets. Green violins denote predictions for uninfected subjects, while light yellow violins denote predictions for CRBSI infected subjects (CABSI infections were excluded from this analysis).

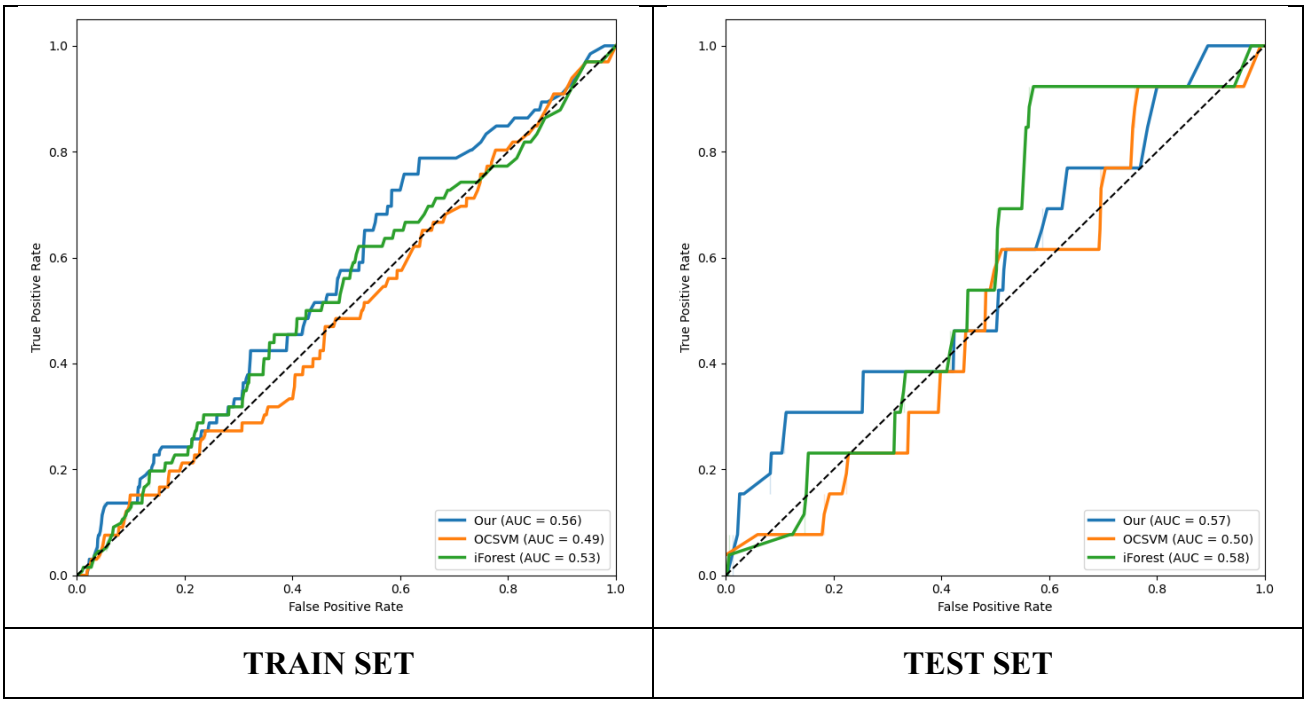


Figure 46. Receiver Operating Characteristic (ROC) curve of Train and Test sets for the three implemented models. Only CABSI are considered, excluding CRBSI events from predictions. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

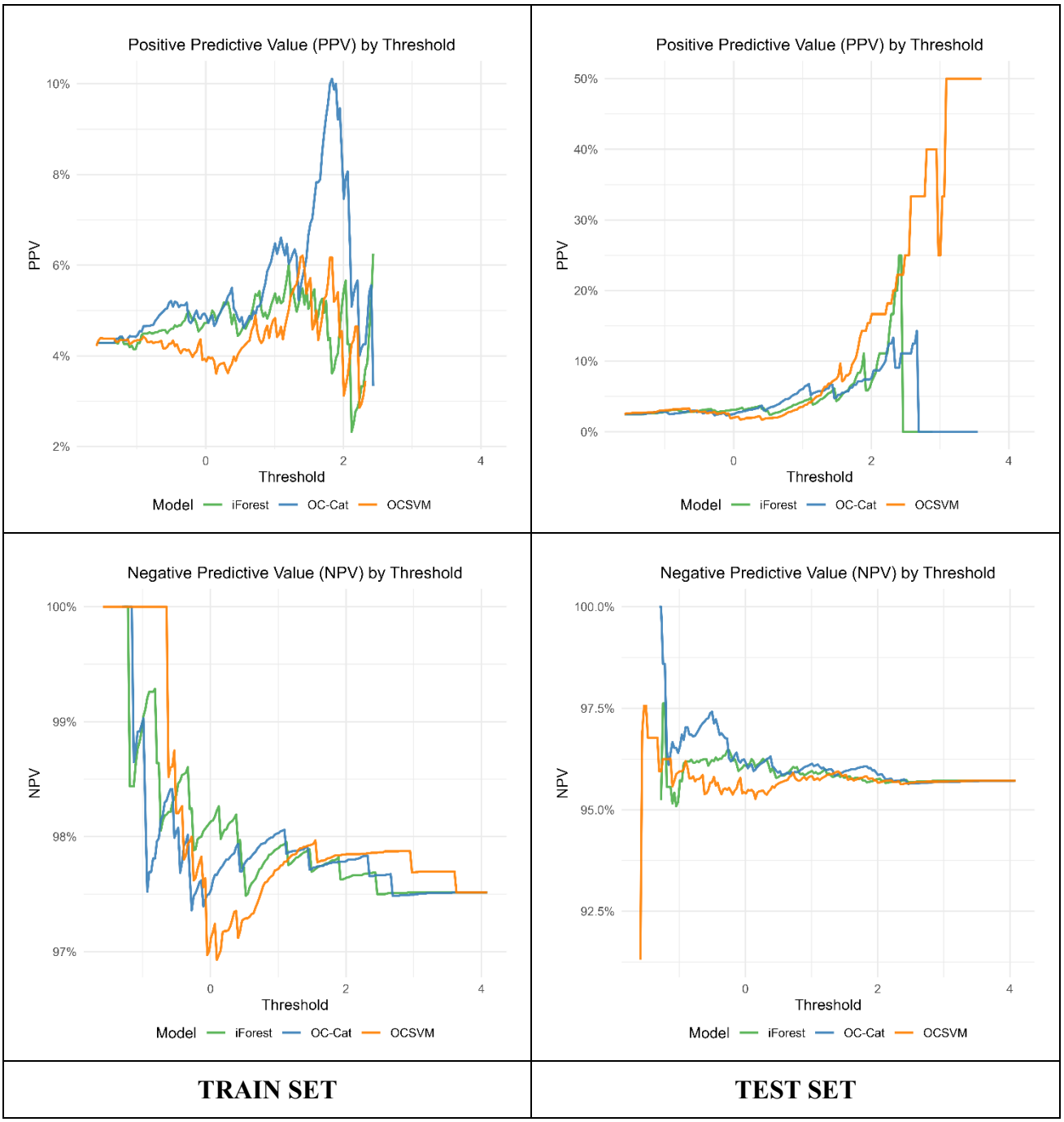


Figure 47. PPV and NPV across probability thresholds for infection risk models on Train and Test sets. Only CABSI are considered, excluding CRBSI events from predictions. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one.

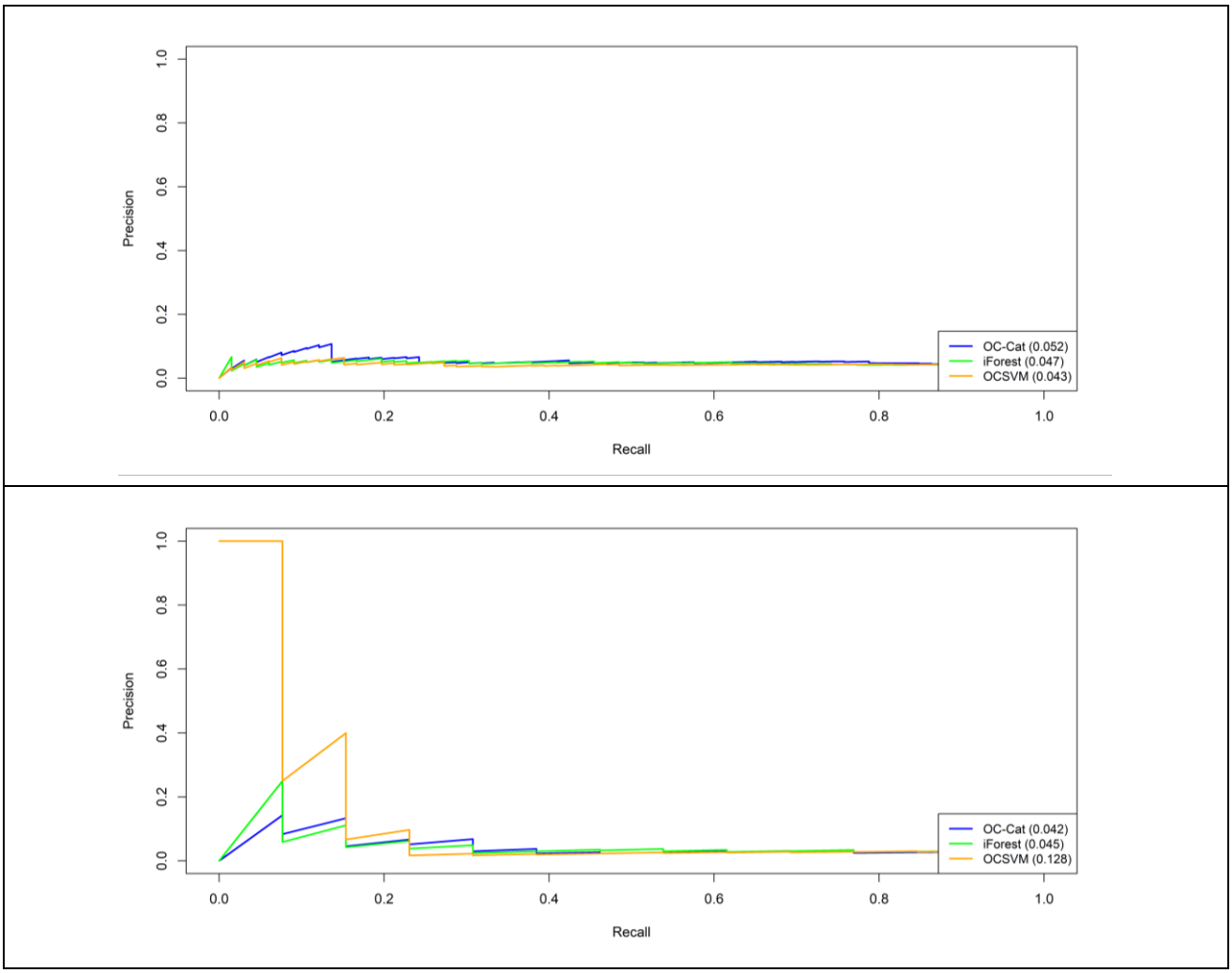


Figure 48. Precision-Recall curves on the Train and Test sets. Only CABSI are considered, excluding CRBSI events from predictions. The OC-Cat model is shown in blue, the iForest model in green, and the OCSVM model in orange. PRAUC values are reported in parentheses.

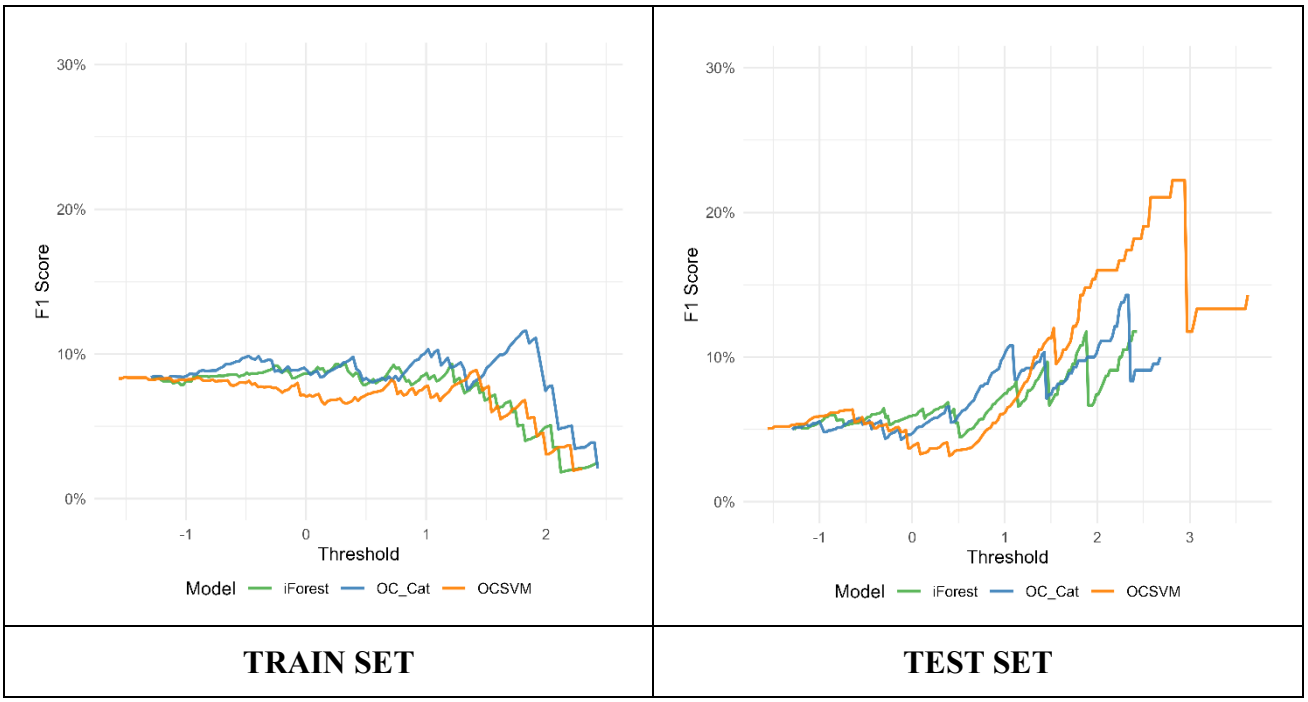


Figure 49. F1 Score curves for the Train and Test sets comparing OC-Cat (blue), iForest (green), and OCSVM (orange) across varying threshold values. Only CABSI are considered, excluding CRBSI events from predictions.

	Mann-Whitney *p-value*		Hellinger distance	
	Train	**Test**	**Train**	**Test**
OC-Cat	0.097	0.383	0.13	0.20
iForest	0.409	0.207	0.12	0.18
OCSVM	0.887	0.532	0.08	0.34

Table 7. Comparison of predicted scores between infected and uninfected patients in training and test sets, assessed using the Mann–Whitney U test *(p-values*) and Hellinger distance. Only CABSI are considered, excluding CRBSI events from predictions.

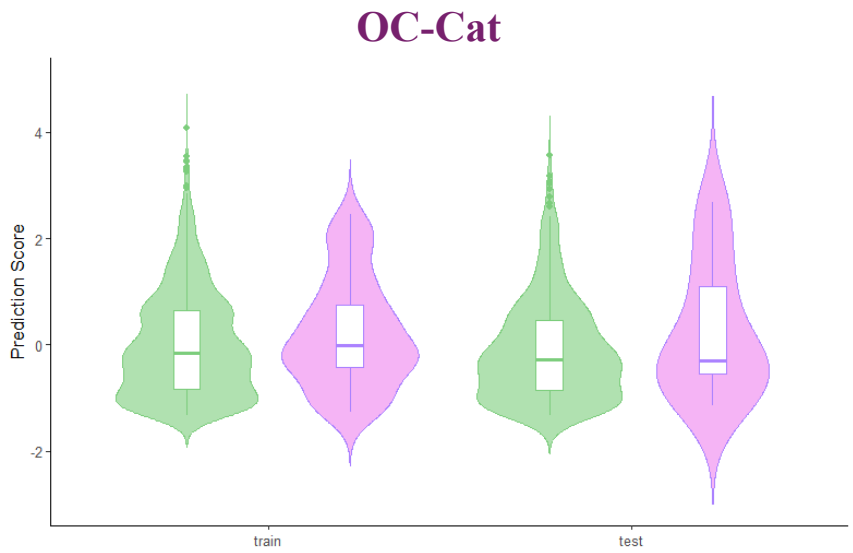


Figure 50. Violin–boxplots of OC-Cat model predicted scores for the training and test sets. Green violins denote predictions for uninfected subjects, while lilac violins denote predictions for CABSI infected subjects (CRBSI infections were excluded from this analysis).

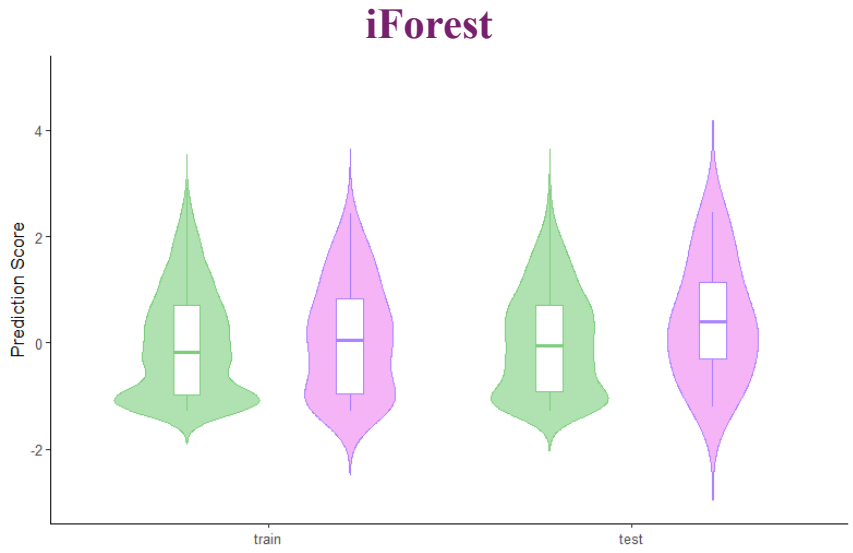


Figure 51. Violin–boxplots of iForest model predicted scores for the training and test sets. Green violins denote predictions for uninfected subjects, lilac violins denote predictions for CABSI infected subjects (CRBSI infections were excluded from this analysis).

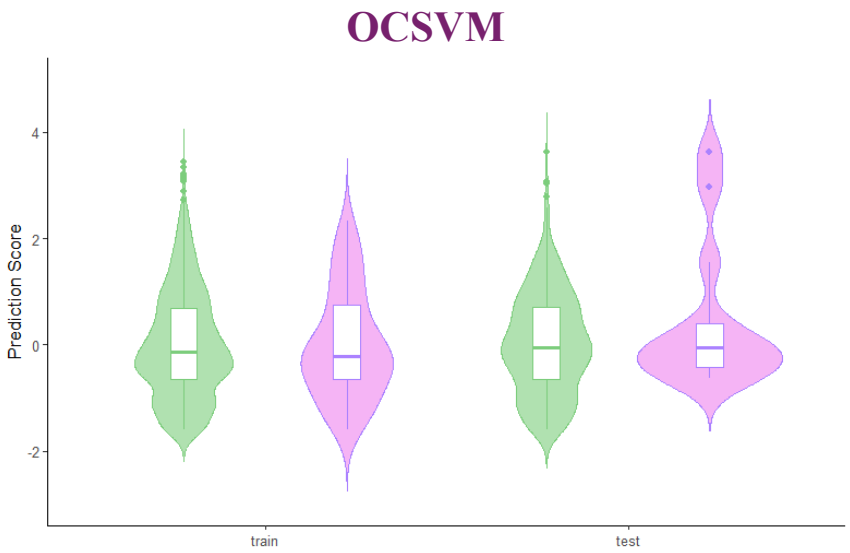


Figure 52. Violin–boxplots of OCSVM model predicted scores for the training and test sets. Green violins denote predictions for uninfected subjects, lilac violins denote predictions for CABSI infected subjects (CRBSI infections were excluded from this analysis).

Finally, a ranking of the features selected and included in the model was applied to evaluate those that were most important for characterizing the "majority class", i.e. the class of patients who did not undergo infection (Fig. 53). Drugs abuse, cirrhosis, patients arrival from ICU, active chemotherapy and type of ward resulted as the five most important variables that characterized healthy patients of the train set.

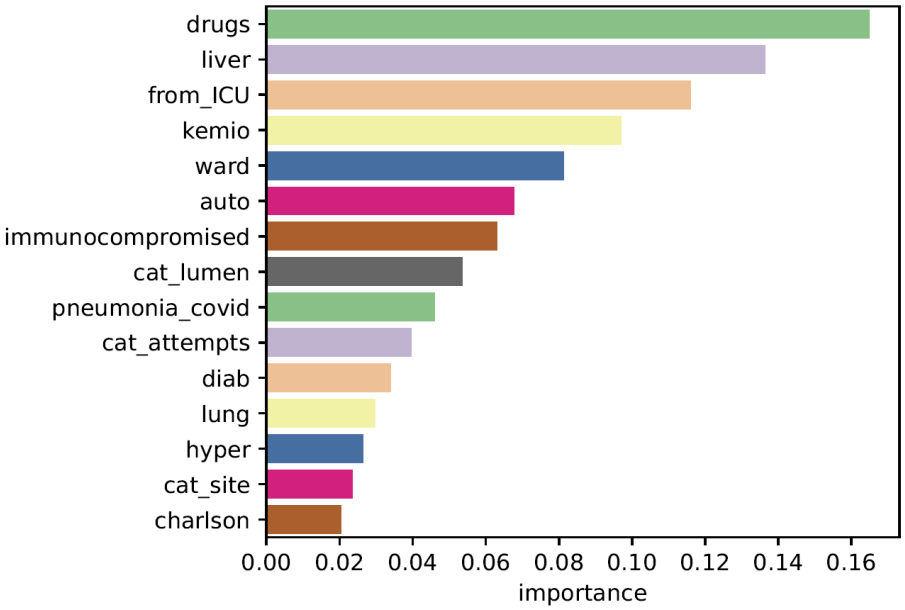


Figure 53. Ranked feature importances for variables selected via the Q-table graph-based method, with importance values computed for the majority class in the training set (uninfected patients).

3.5.1 Sensitivity analysis

The sensitivity analysis proceeded along two lines: first, we evaluated model performance without feature selection; second, we restricted the dataset to cases where catheter maintenance lasted at least a predefined threshold (7, 14 and 21 days).

3.5.1.1. All features

Using all clinical features of interest with no missing values in the dataset, we evaluated the model's performance and predictions in comparison with two benchmark models. The AUC value for the OC-Cat model demonstrated comparable performance to both iForest and OCSVM in the training set. In contrast, the performance of these two benchmark models deteriorated when using the complete feature test set (Figs. 54-60). The Mann-Whitney U test indicated statistically significant differences only for the OC-Cat prediction scores, in both the training and test sets. The Hellinger index was consistently higher for the OC-Cat model than for the benchmark models across all scenarios (Table 8).

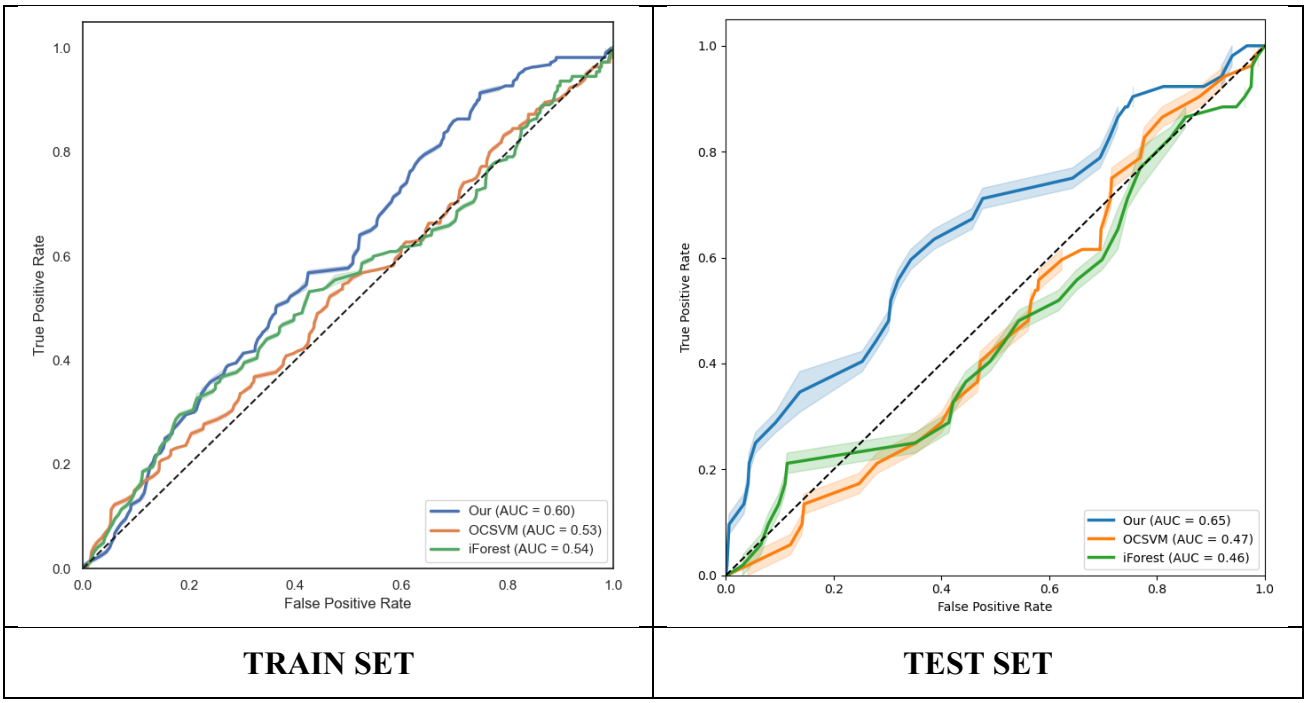


Figure 54. Receiver Operating Characteristic (ROC) curve of Train and Test sets for the three implemented models. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards. All the 33 original features were retained for analysis.

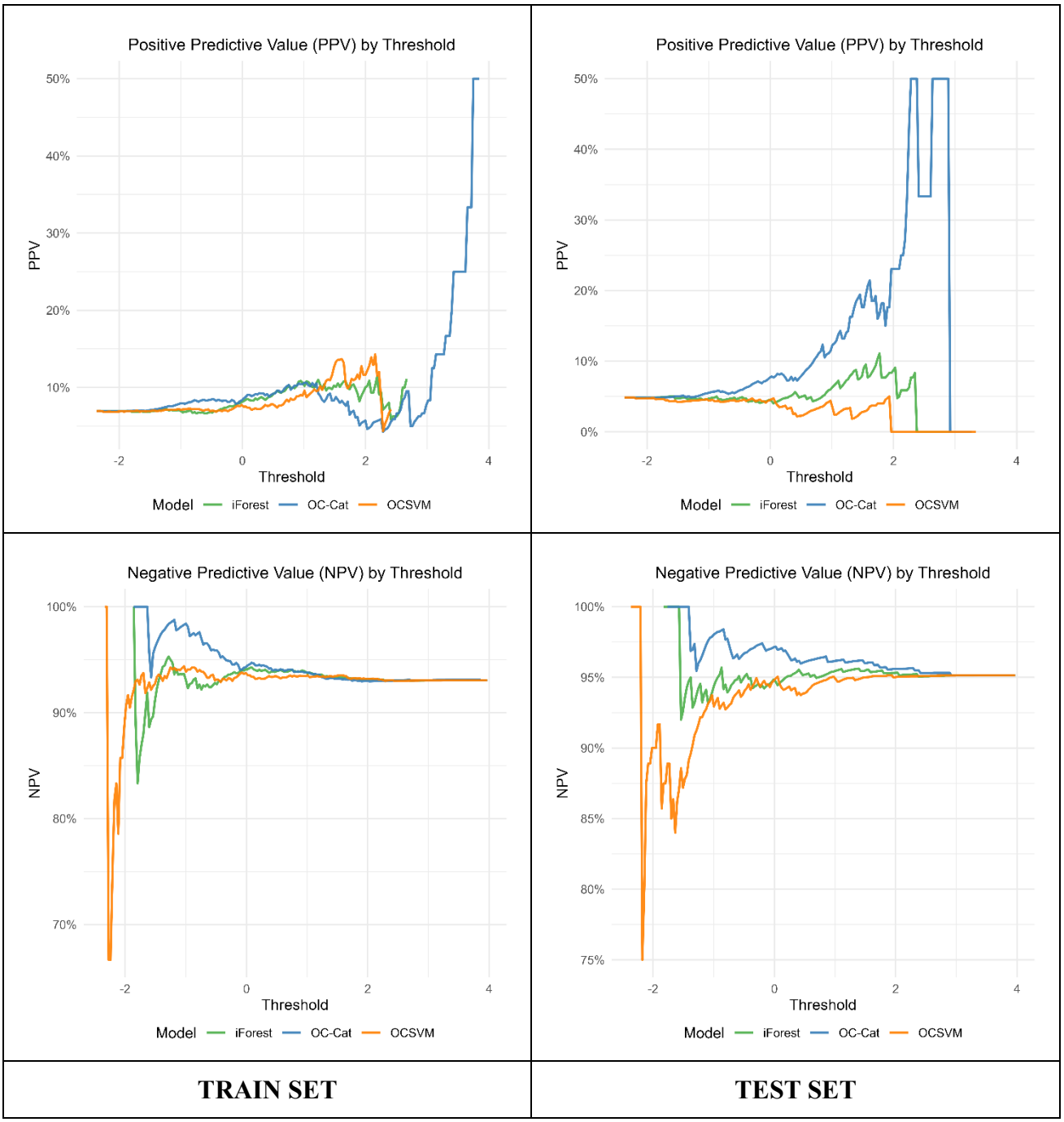


Figure 55. PPV and NPV across probability thresholds for infection risk models on Train and Test sets. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards. All the 33 original features were retained for analysis.

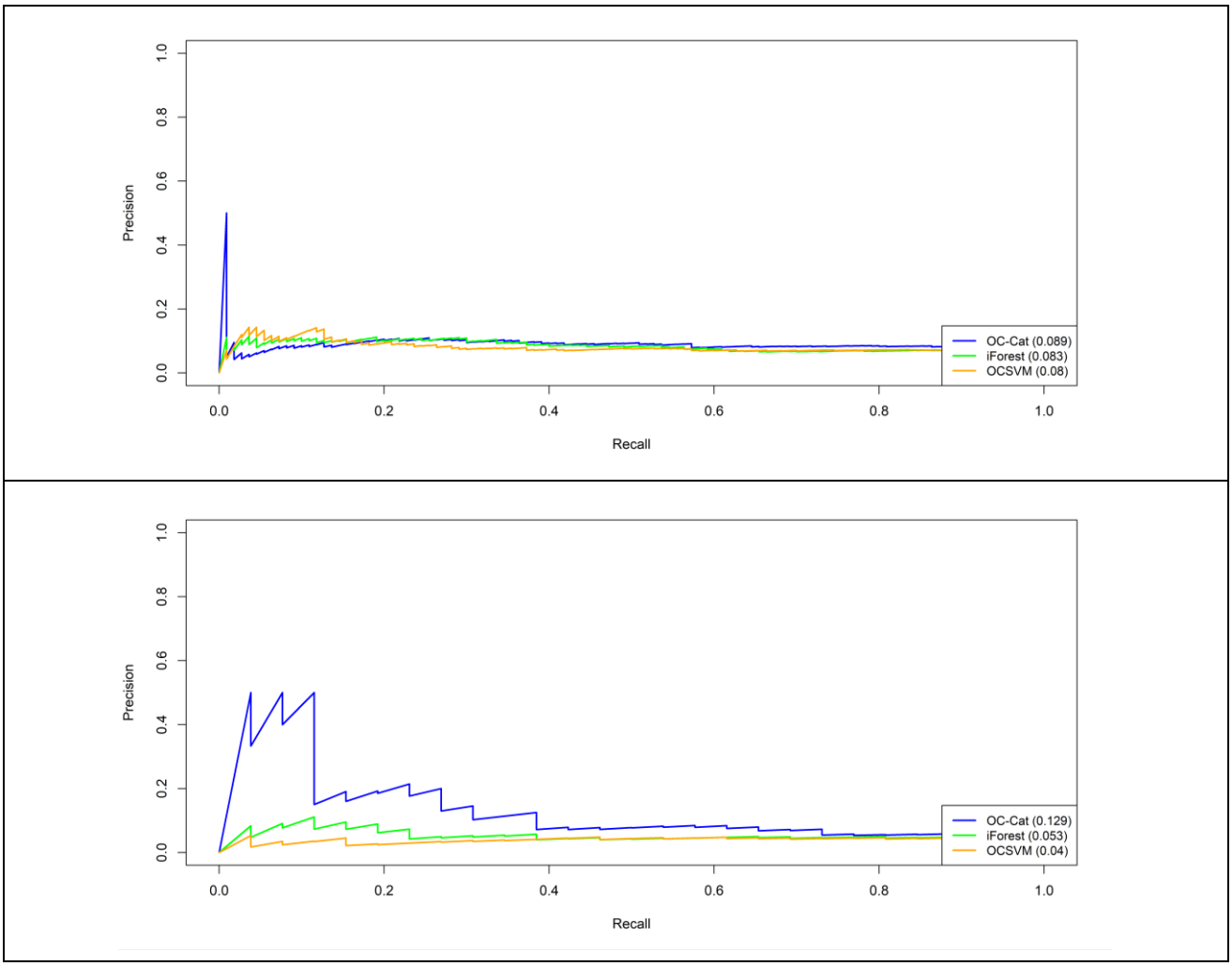


Figure 56. Precision-Recall curves on the Train and Test sets. The OC-Cat model is shown in blue, the iForest model in green, and the OCSVM model in orange. PRAUC values are reported in parentheses. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards. All the 33 original features were retained for analysis.

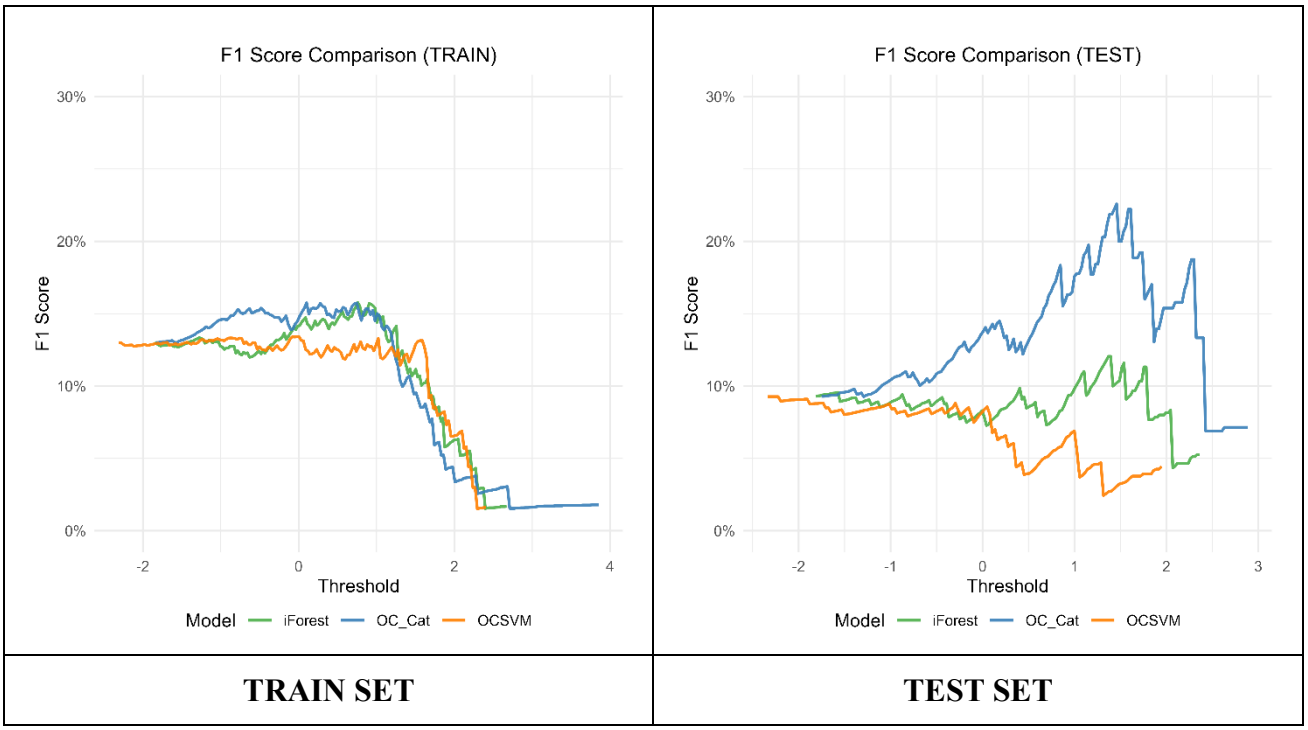


Figure 57. F1 Score curves for the Train and Test sets comparing OC-Cat (blue), iForest (green), and OCSVM (orange) across varying threshold values. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards. All the 33 original features were retained for analysis.

	Mann-Whitney *p-value*		**Hellinger distance**	
	Train	**Test**	**Train**	**Test**
OC-Cat	<0.001	0.009	0.14	0.23
iForest	0.151	0.811	0.11	0.11
OCSVM	0.372	0.196	0.09	0.13

Table 8. Comparison of predicted scores between infected and uninfected patients in training and test sets, assessed using the Mann–Whitney U test *(p-values*) and Hellinger distance. All the 33 original features were retained for analysis.

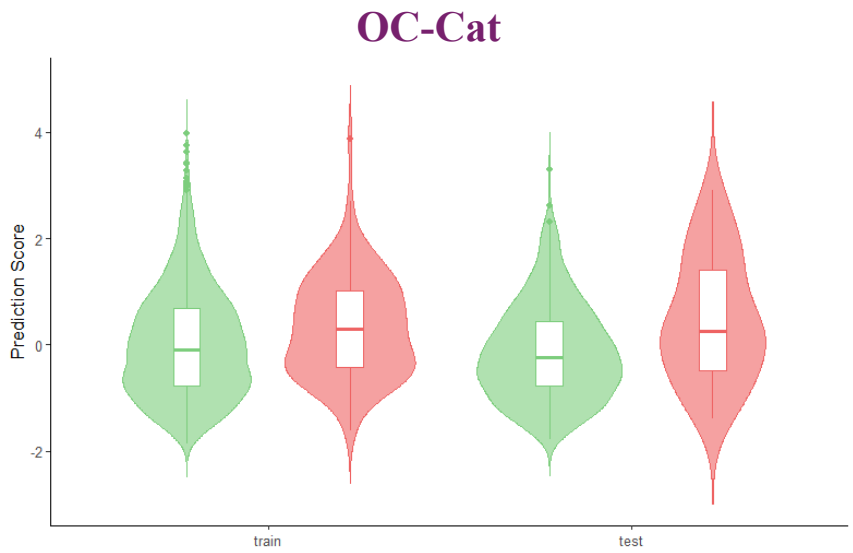


Figure 58. Violin–boxplots showing the OC-Cat model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects. All the 33 original features were retained for analysis.

Figure 59. Violin–boxplots showing the Isolation Forest model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects. All the 33 original features were retained for analysis.

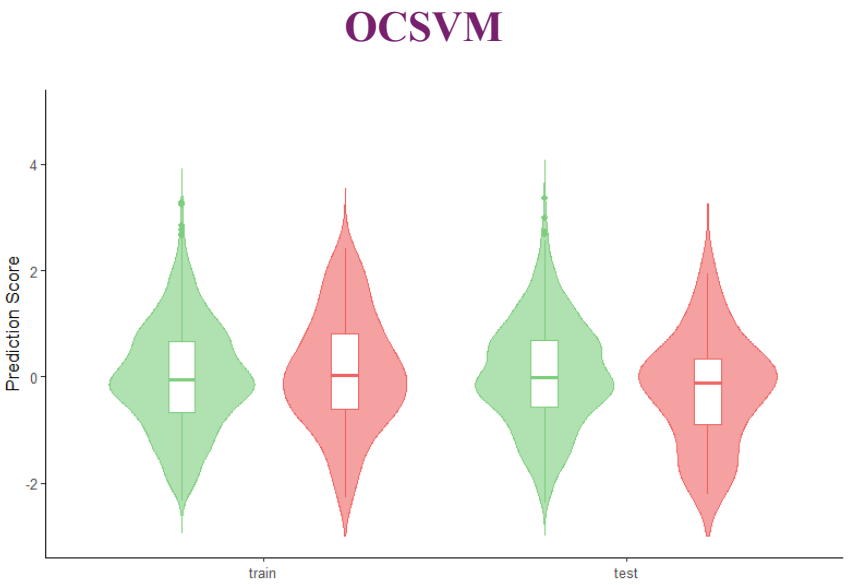


Figure 60. Violin–boxplots showing the OCSVM model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects. All the 33 original features were retained for analysis.

Considering only the CRBSI among infected subjects, the performances for the OC-Cat model were superior to those of the benchmark models (Figs. 61-67). The Hellinger distance was higher for the OC-Cat model than for the OCSVM model in both the training and test sets (Table 9).

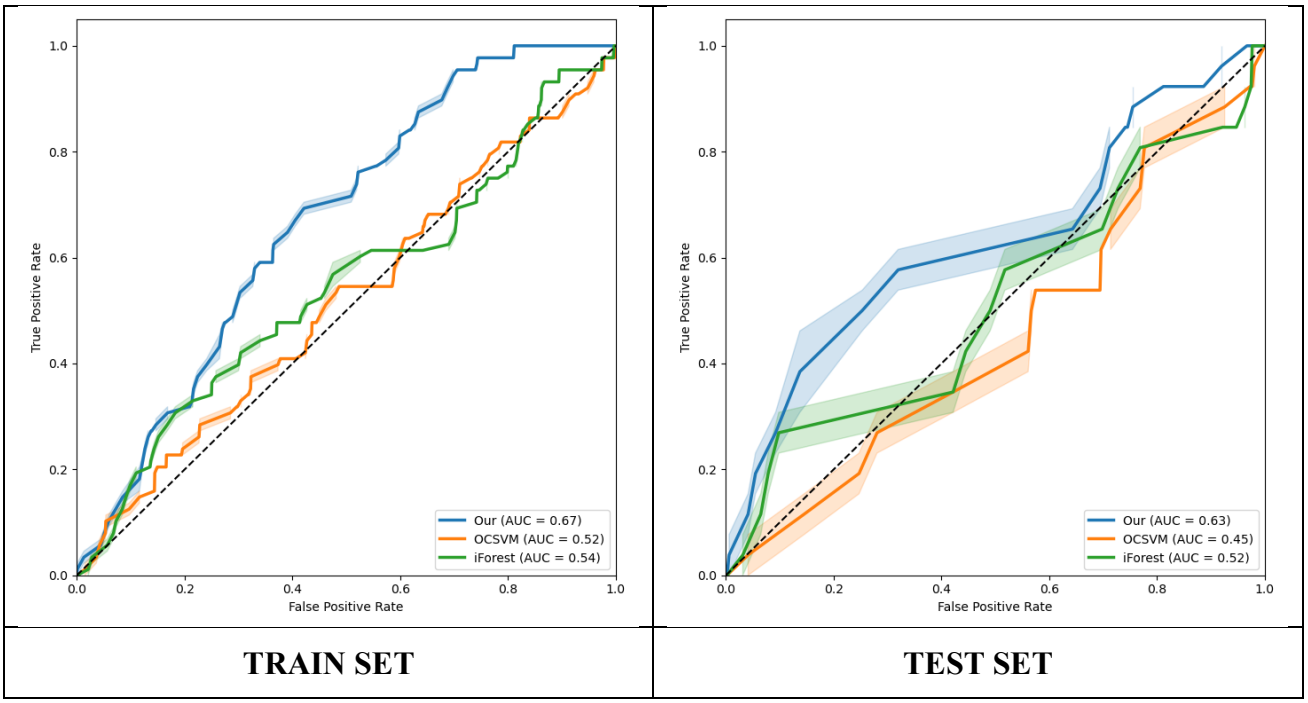


Figure 61. Receiver Operating Characteristic (ROC) curve of Train and Test sets for the three implemented models. Only CRBSI are considered, excluding CABSI events from predictions. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards. All the 33 original features were retained for analysis.

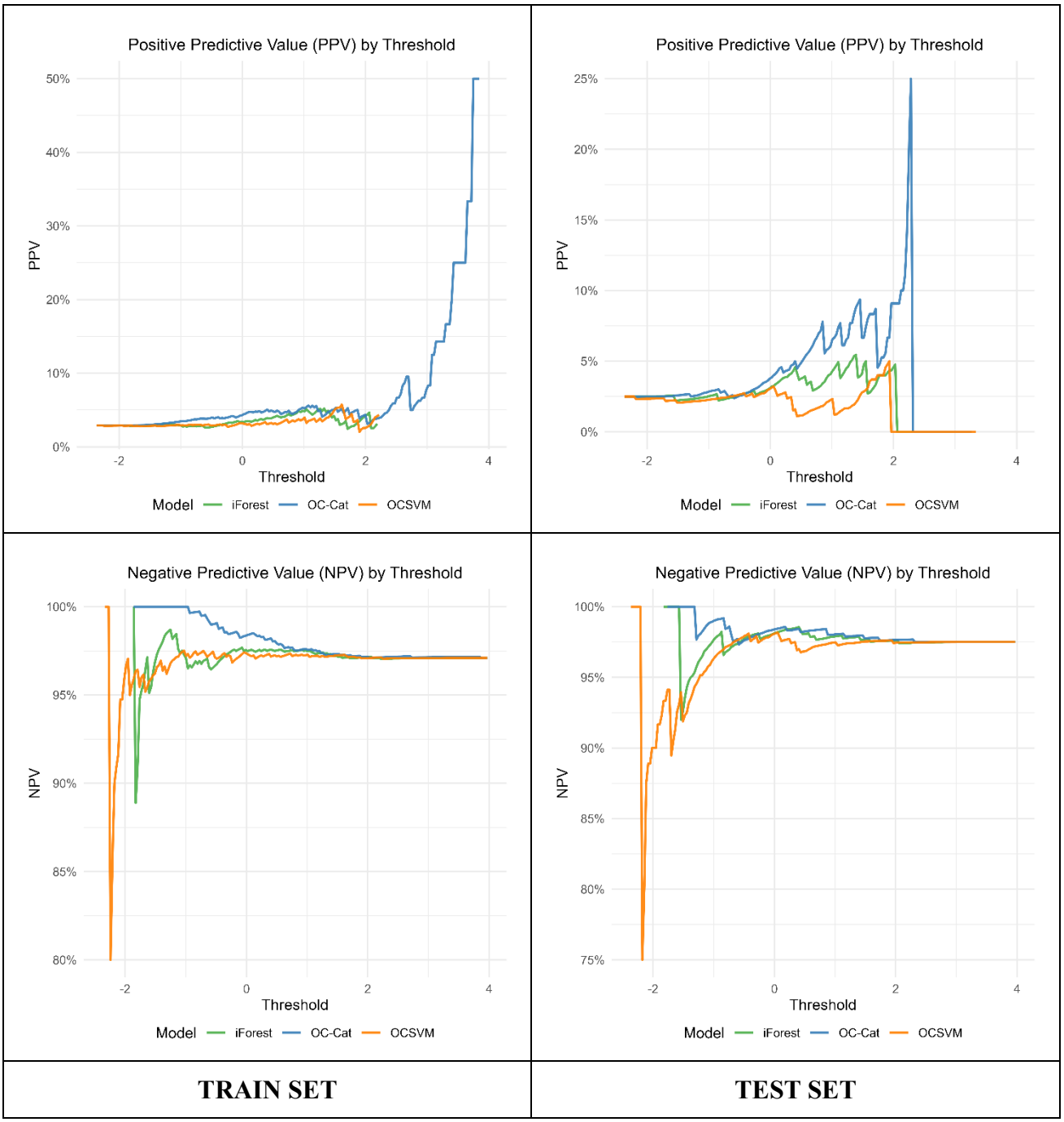


Figure 62. PPV and NPV across probability thresholds for infection risk models on Train and Test sets. Only CRBSI are considered, excluding CABSI events from predictions. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards. All the 33 original features were retained for analysis.

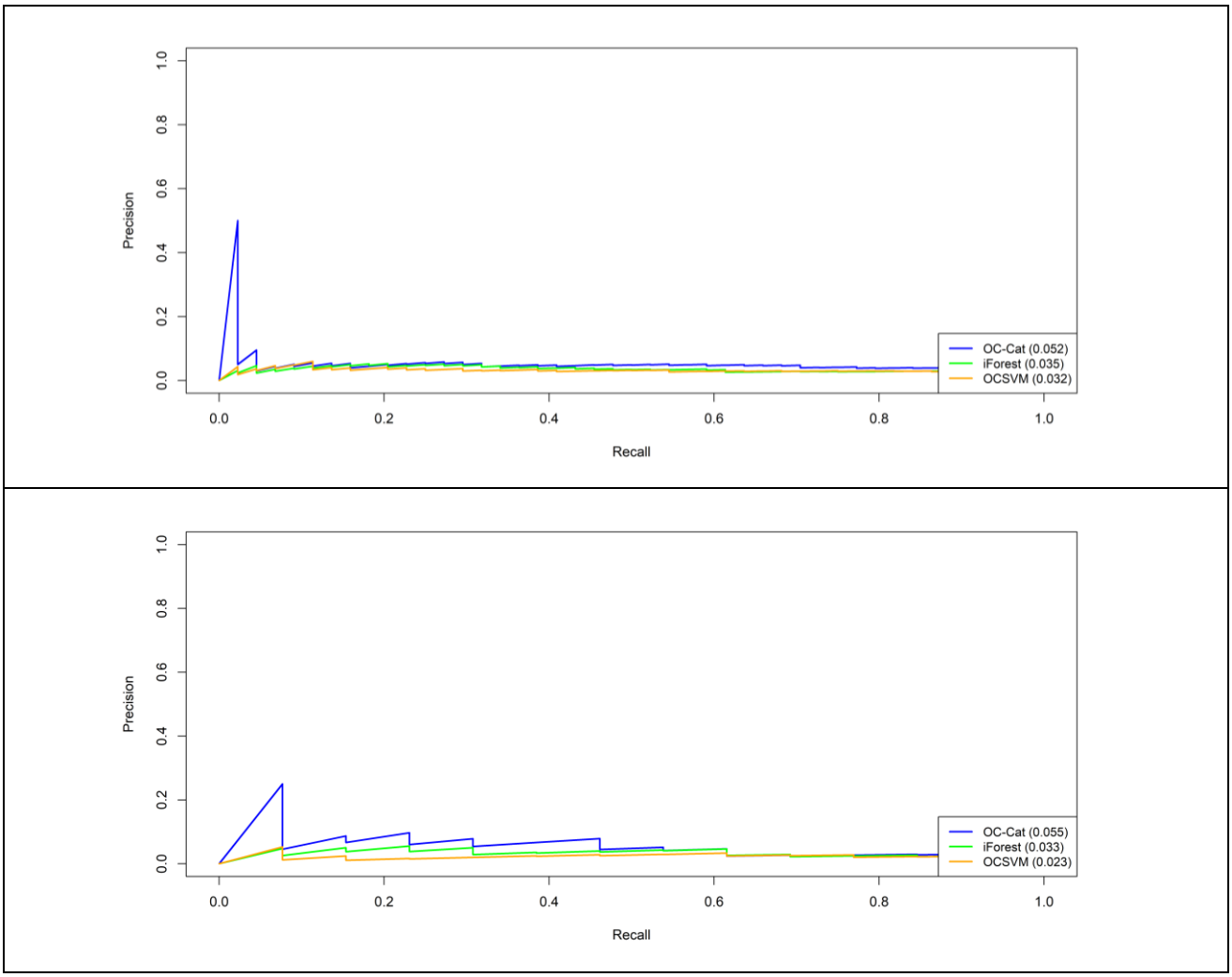


Figure 63. Precision-Recall curves on the Train and Test sets. Only CRBSI are considered, excluding CABSI events from predictions. The OC-Cat model is shown in blue, the iForest model in green, and the OCSVM model in orange. PRAUC values are reported in parentheses. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards. All the 33 original features were retained for analysis.

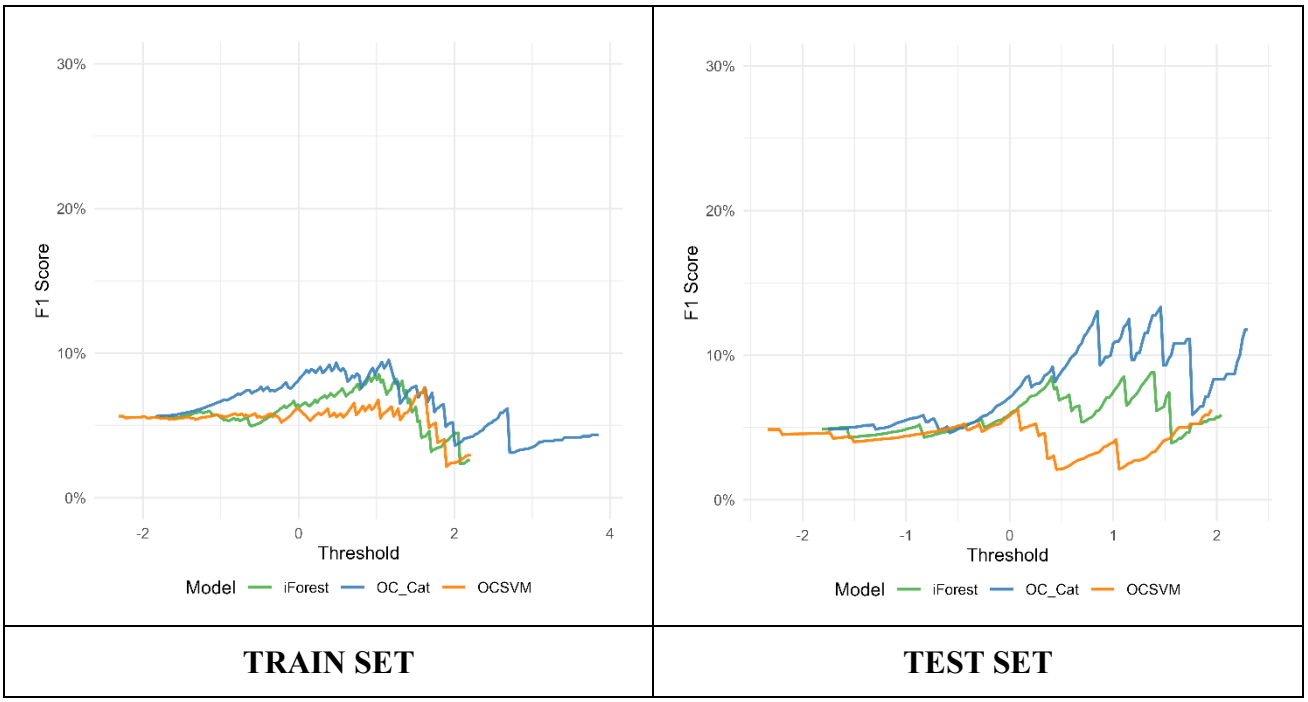


Figure 64. F1 Score curves for the Train and Test sets comparing OC-Cat (blue), iForest (green), and OCSVM (orange) across varying threshold values. Only CRBSI are considered, excluding CABSI events from predictions. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards. All the 33 original features were retained for analysis.

	Mann-Whitney *p-value*		Hellinger distance	
	Train	**Test**	**Train**	**Test**
OC-Cat	<0.001	0.100	0.22	0.22
iForest	0.319	0.400	0.12	0.18
OCSVM	0.691	0.731	0.09	0.21

Table 9. Comparison of predicted scores between infected and uninfected patients in training and test sets, assessed using the Mann–Whitney U test *(p-values*) and Hellinger distance. Only CRBSI are considered, excluding CABSI events from predictions. All the 33 original features were retained for analysis.

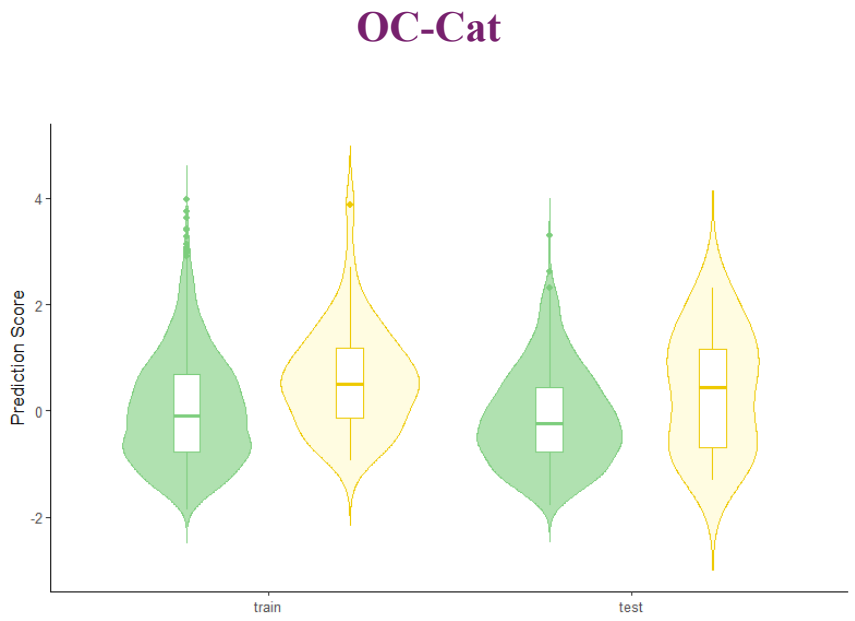


Figure 65. Violin–boxplots of OC-Cat model predicted scores for the training and test sets. Green violins denote predictions for uninfected subjects, while light yellow violins denote predictions for CRBSI infected subjects (CABSI infections were excluded from this analysis). All the 33 original features were retained for analysis.

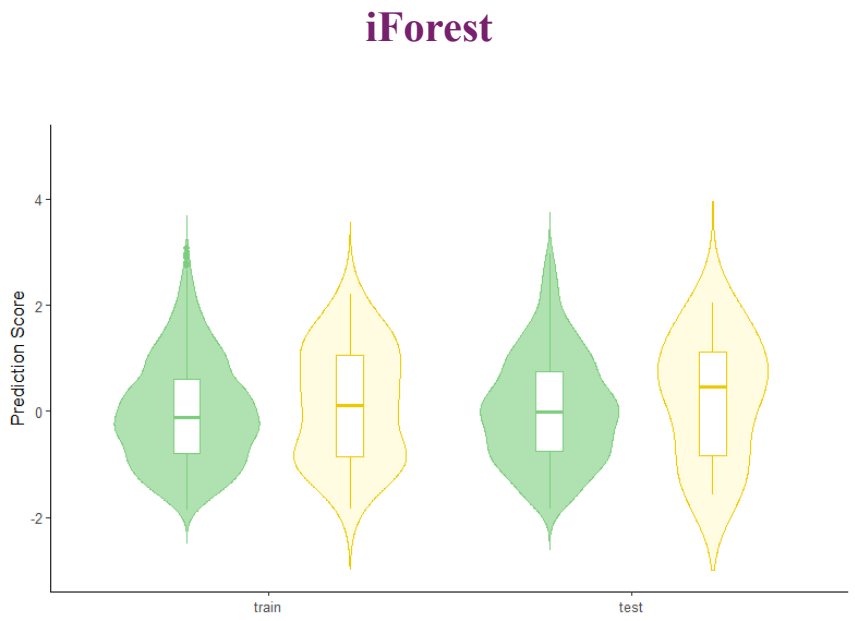


Figure 66. Violin–boxplots of iForest model predicted scores for the training and test sets. Green violins denote predictions for uninfected subjects, while light yellow violins denote predictions for CRBSI infected subjects (CABSI infections were excluded from this analysis). All the 33 original features were retained for analysis.

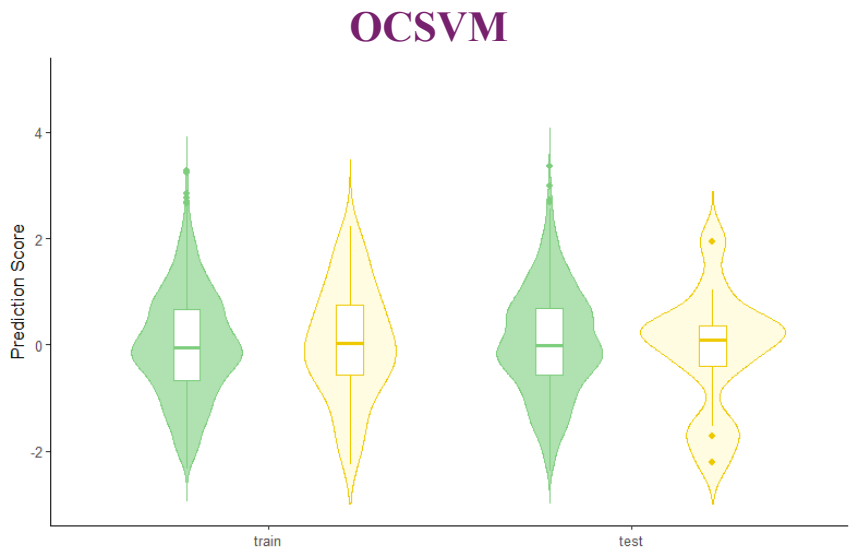


Figure 67. Violin–boxplots of OCSVM model predicted scores for the training and test sets. Green violins denote predictions for uninfected subjects, while light yellow violins denote predictions for CRBSI infected subjects (CABSI infections were excluded from this analysis). All the 33 original features were retained for analysis.

Considering only the CABSI among infected subjects, the OC-Cat model outperformed the benchmark models on the test set, while achieving comparable results to the other two models on the train set (Figs. 68-74). In both the training and test sets, the Hellinger distance for the OC-Cat model exceeded that of the OCSVM model and was comparable to that of iForest (Table 10).

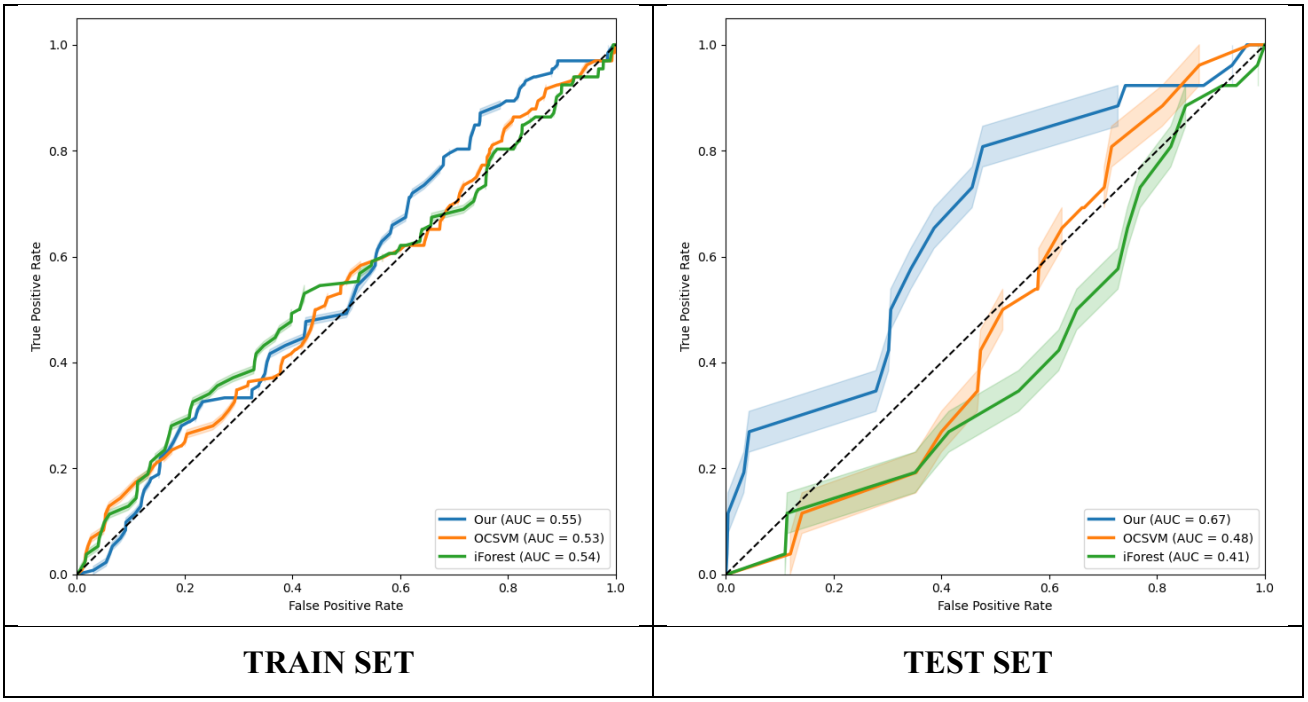


Figure 68. Receiver Operating Characteristic (ROC) curve of Train and Test sets for the three implemented models. Only CABSI are considered, excluding CRBSI events from predictions. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards. All the 33 original features were retained for analysis.

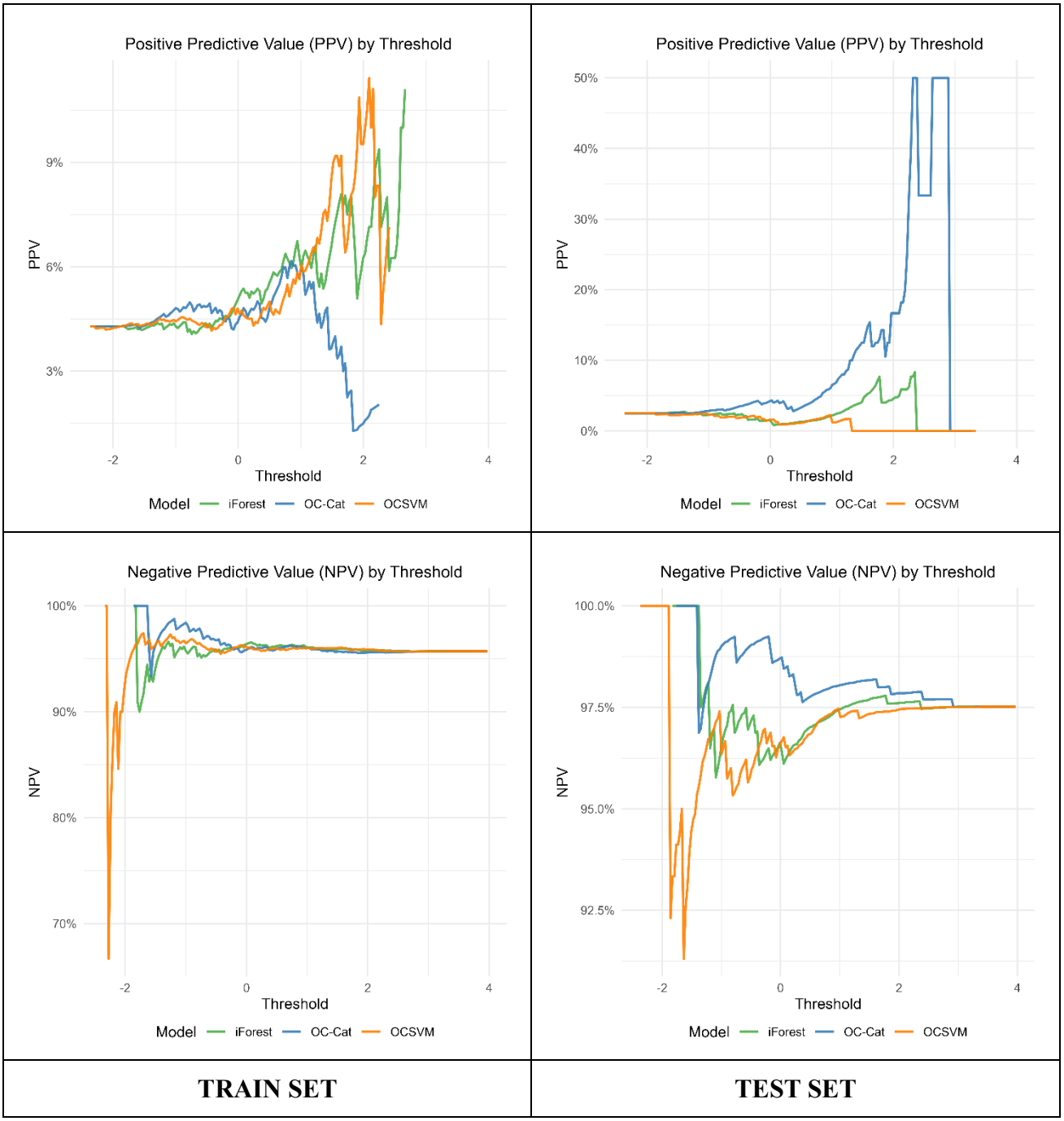


Figure 69. PPV and NPV across probability thresholds for infection risk models on Train and Test sets. Only CABSI are considered, excluding CRBSI events from predictions. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards. All the 33 original features were retained for analysis.

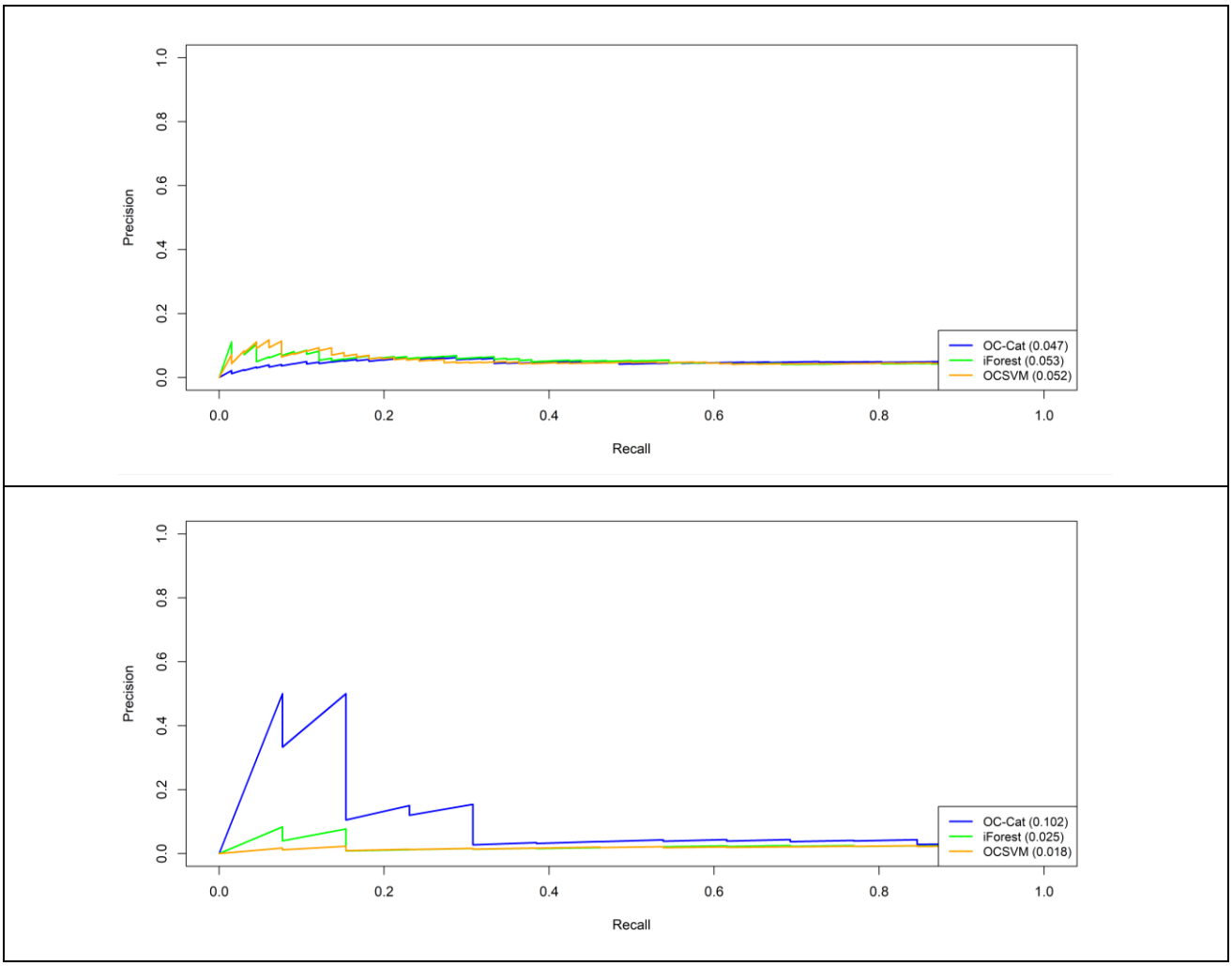


Figure 70. Precision-Recall curves on the Train and Test sets. Only CABSI are considered, excluding CRBSI events from predictions. The OC-Cat model is shown in blue, the iForest model in green, and the OCSVM model in orange. PRAUC values are reported in parentheses. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards. All the 33 original features were retained for analysis.

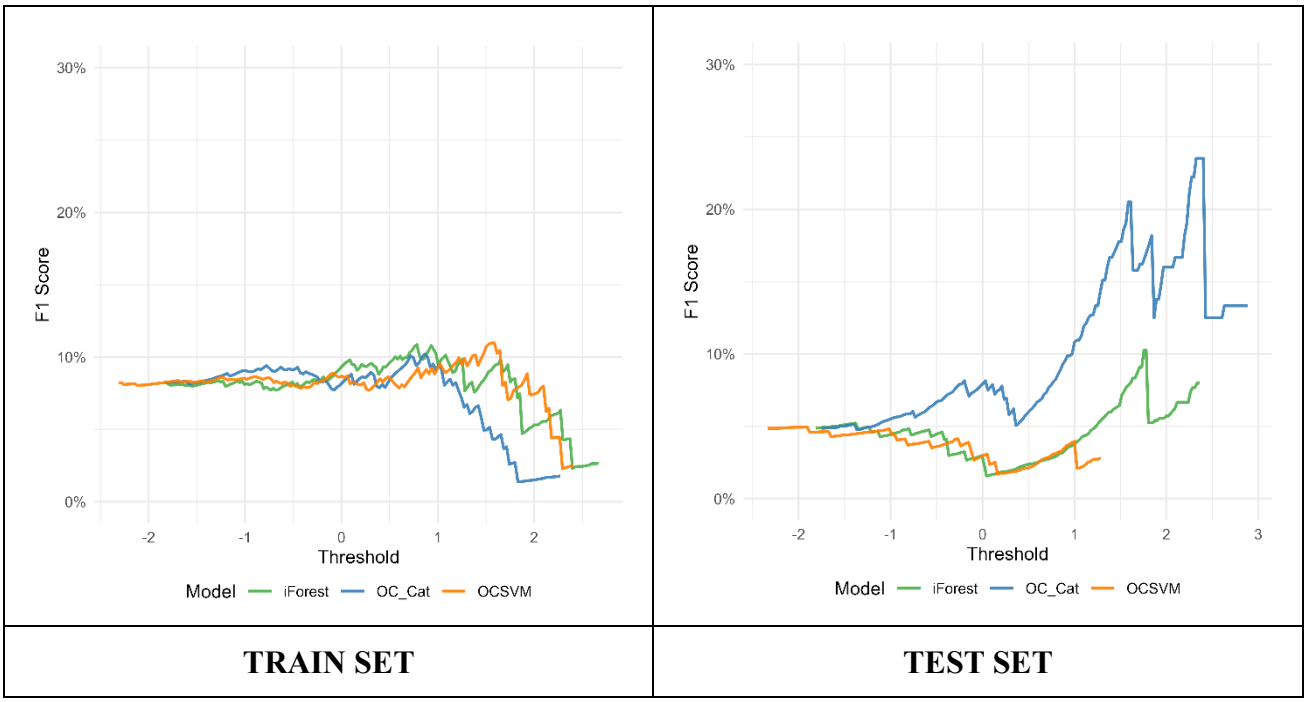


Figure 71. F1 Score curves for the Train and Test sets comparing OC-Cat (blue), iForest (green), and OCSVM (orange) across varying threshold values. Only CABSI are considered, excluding CRBSI events from predictions. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards. All the 33 original features were retained for analysis.

	Mann-Whitney *p-value*		**Hellinger distance**	
	Train	**Test**	**Train**	**Test**
OC-Cat	0.165	0.037	0.12	0.27
iForest	0.284	0.291	0.12	0.28
OCSVM	0.397	0.132	0.10	0.17

Table 10. Comparison of predicted scores between infected and uninfected patients in training and test sets, assessed using the Mann–Whitney U test *(p-values*) and Hellinger distance. Only CABSI are considered, excluding CRBSI events from predictions. All the 33 original features were retained for analysis.

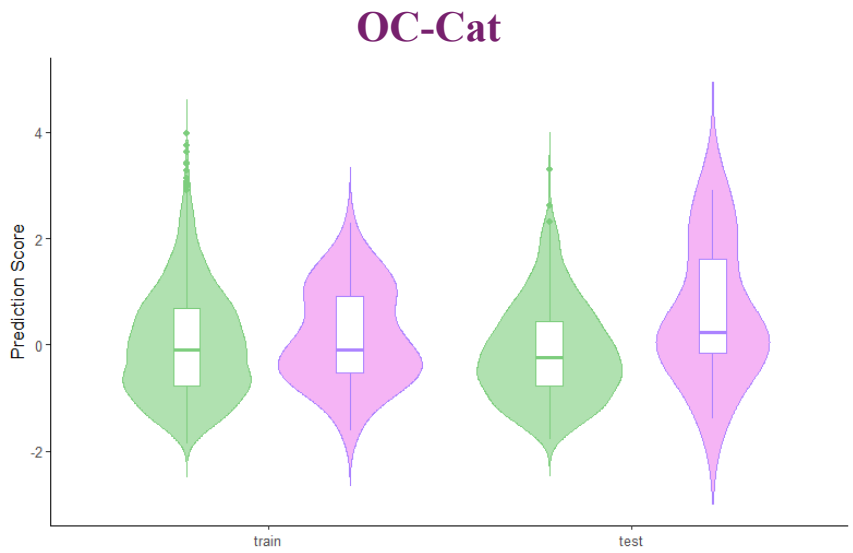


Figure 72. Violin–boxplots of OC-Cat model predicted scores for the training and test sets. Green violins denote predictions for uninfected subjects, while lilac violins denote predictions for CABSI infected subjects (CRBSI infections were excluded from this analysis). All the 33 original features were retained for analysis.

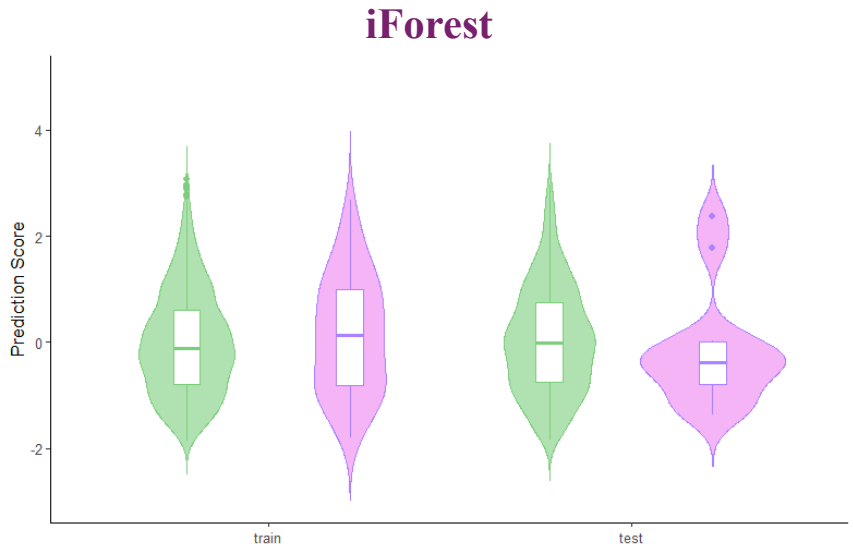


Figure 73. Violin–boxplots of iForest model predicted scores for the training and test sets. Green violins denote predictions for uninfected subjects, while lilac violins denote predictions for CABSI infected subjects (CRBSI infections were excluded from this analysis). All the 33 original features were retained for analysis.

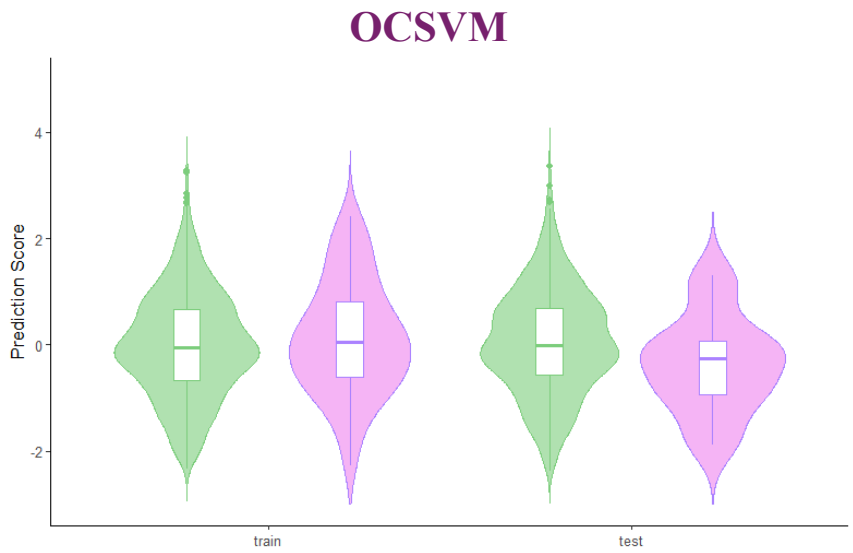


Figure 74. Violin–boxplots of OCSVM model predicted scores for the training and test sets. Green violins denote predictions for uninfected subjects, while lilac violins denote predictions for CABSI infected subjects (CRBSI infections were excluded from this analysis). All the 33 original features were retained for analysis.

The variable-importance ranking procedure was additionally applied to the complete set of 33 initial variables. Notably, the majority of variables identified by the graphical approach following Q-table construction ranked among the top positions in this new ordering, suggesting a high degree of consistency in their relevance to defining the majority class. Furthermore, the relative ordering of the variables when considered individually remained unchanged when these variables were incorporated into the global ranking alongside all 33 features. The occurrence of complications during catheter insertion, the patient's tracheostomy status, and a history of infection within the preceding 30 days constitute factors relevant to defining the majority class; however, these variables were not retained by the graphical approach (Fig. 75).

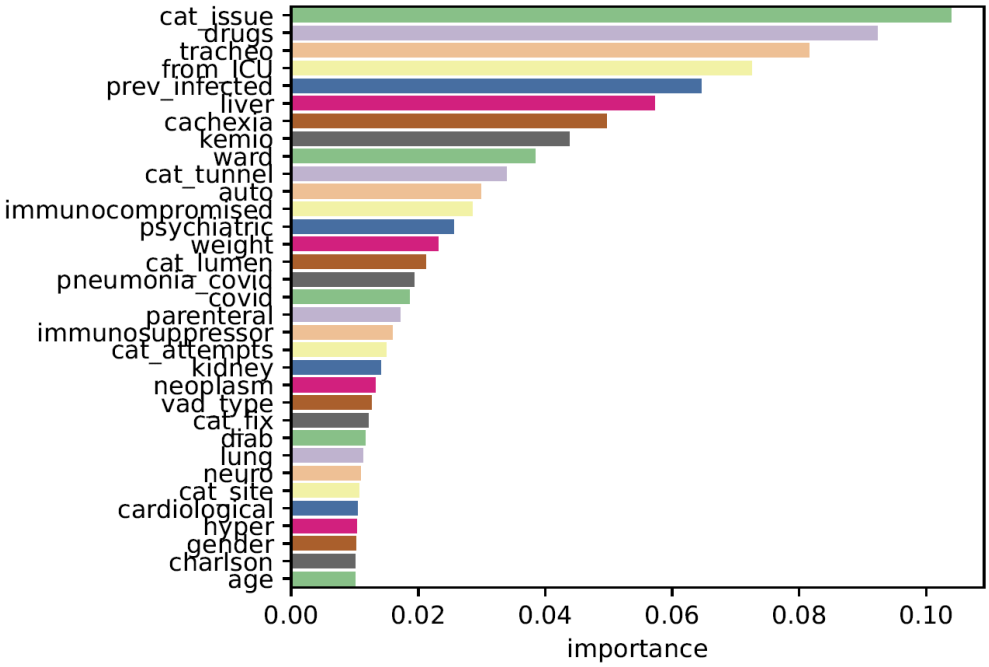


Figure 75. Ranked feature importances considering all the 33 starting features, with importance values computed for the majority class in the training set (uninfected patients).

3.5.1.2. Threshold

The three-steps approach was also applied to a dataset from which subjects retaining the catheter for fewer than seven days were excluded from the analyses. In this context, subjects who did not develop CRBSI or CABSI (considered jointly in these analyses) were classified as healthy, provided that no infection occurred within the first seven days of catheter retention. The same procedure was subsequently repeated using cut-off points of 14 and 21 days. The cut-off values were determined based on the distribution of catheter placement duration across all subjects, in conjunction with considerations of clinical relevance (Fig. 76).

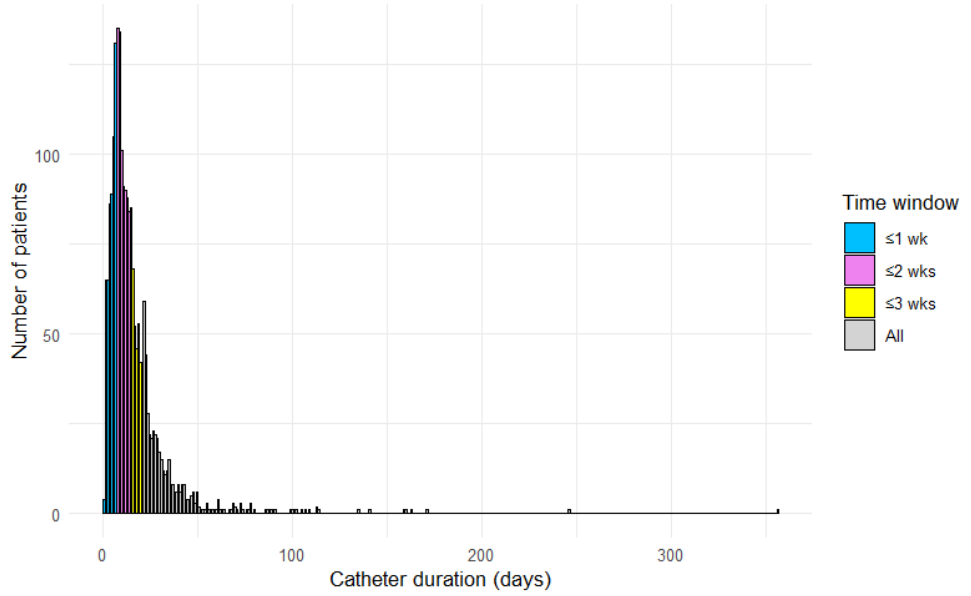


Figure 76. Bar plot showing overall VAD placement duration. Blue represents catheters retained for up to 7 days, pink represents those retained for between 7 and 14 days, yellow represents those retained for between 14 and 21 days, and grey represents all remaining cases.

The number of events decreased substantially after recoding, with infections representing 19.1%, 49.3%, and 72.8% of the total infections considered in the main analyses, respectively. The complete dataset included subsets corresponding to 81.7%, 47.3%, and 28.5% of the total initial dataset of 2,120 observations. (Tab. 11).

	No infection (N=1706)	**Infected (N=26)**	**CABSI (N=19)**	**CRBSI (N=7)**	**Overall (N=1732)**
	At least 7 days (sensitivity analysis)				
Train	1259 (73.8%)	20 (76.9%)	15 (78.9%)	5 (71.4%)	1279 (73.8%)
Test	447 (26.2%)	6 (23.1%)	4 (21.1%)	2 (28.6%)	453 (26.2%)
	No infection (N=936)	**Infected (N=67)**	**CABSI (N=42)**	**CRBSI (N=25)**	**Overall (N=1003)**
	At least 14 days (sensitivity analysis)				
Train	707 (75.5%)	55 (82.1%)	36 (85.7%)	19 (76.0%)	762 (76.0%)
Test	229 (24.5%)	12 (17.9%)	6 (14.3%)	6 (24.0%)	241 (24.0%)
	No infection (N=506)	**Infected (N=99)**	**CABSI (N=57)**	**CRBSI (N=43)**	**Overall (N=605)**
	At least 21 days (sensitivity analysis)				
Train	394 (77.9%)	81 (81.8%)	49 (86.0%)	33 (76.7%)	475 (78.5%)
Test	112 (22.1%)	18 (18.2%)	8 (14.0%)	10 (23.3%)	130 (21.5%)

Table 11. Number of subjects in each dataset, stratified by vascular access device (VAD) placement duration and by train *vs* test dataset.

Most of the selected features in the main analysis were also preserved in the sensitivity analysis. Table 12 lists the variables retained in each of them.

Feature	≥ 7 days	≥ 14 days	≥ 21 days	All
gender	✓			
age				
diab		✓	✓	✓
cardiological				
lung				✓
neuro			✓	
liver	✓	✓		✓
kidney		✓	✓	
auto	✓	✓	✓	✓
immunocompromised	✓	✓	✓	✓
neoplasm				
hyper	✓	✓	✓	✓
drugs	✓	✓	✓	✓
psychiatric			✓	
weigth				
charlson	✓	✓	✓	✓
prev_infected				
from_ICU	✓	✓	✓	✓
ward	✓	✓	✓	✓
covid				
pneumonia_covid	✓	✓	✓	✓
tracheo			✓	
cachexia			✓	
parenteral				
kemio	✓	✓	✓	✓
immunosuppressor				
vad_type				
cat_lumen	✓	✓		✓
cat_fix		✓	✓	
cat_tunnel				
cat_attempts	✓	✓		✓
cat_site	✓	✓	✓	✓

Table 12. List of features selected in the principal models and using the cut-offs analysis.

Adopting the 7-day cut-off, 14 variables out of the initial 33 were retained through graphical feature selection. Compared to the main analysis, gender was retained as additional variable, while diabetes and lung infections were discarded (Fig. 77). The performance of the OC-Cat model deteriorated slightly, particularly in the test set. In this case, the median prediction score for infected patients was lower than that for uninfected subjects, and a similar pattern was observed for both benchmarks (Figs 78-84) and Table 13).

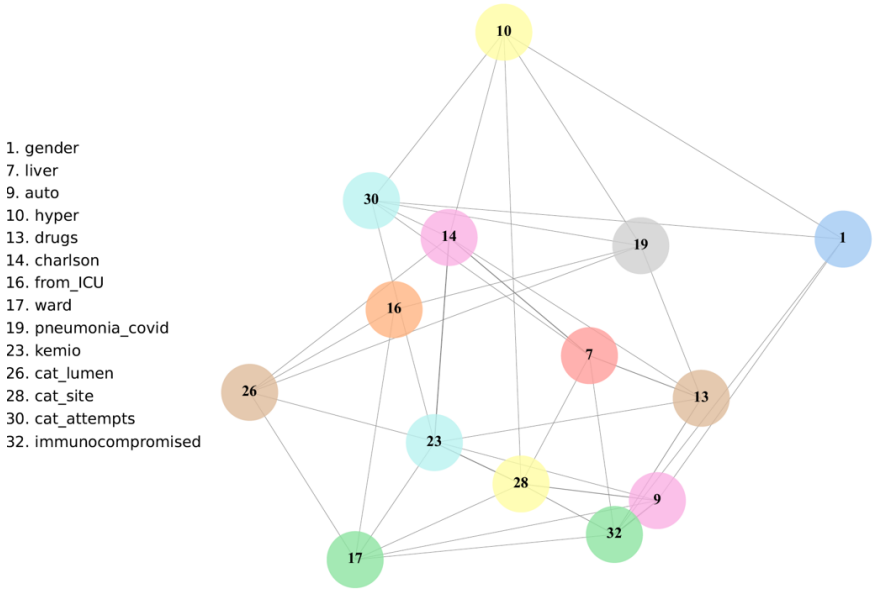


Figure 77. Network graph showing nodes and weighted connections among selected features of the train dataset. Subjects who retained the catheter for fewer than 7 days were excluded from the majority class, while patients who remained free of infection during the first 7 days following catheter placement were counted in the majority class.

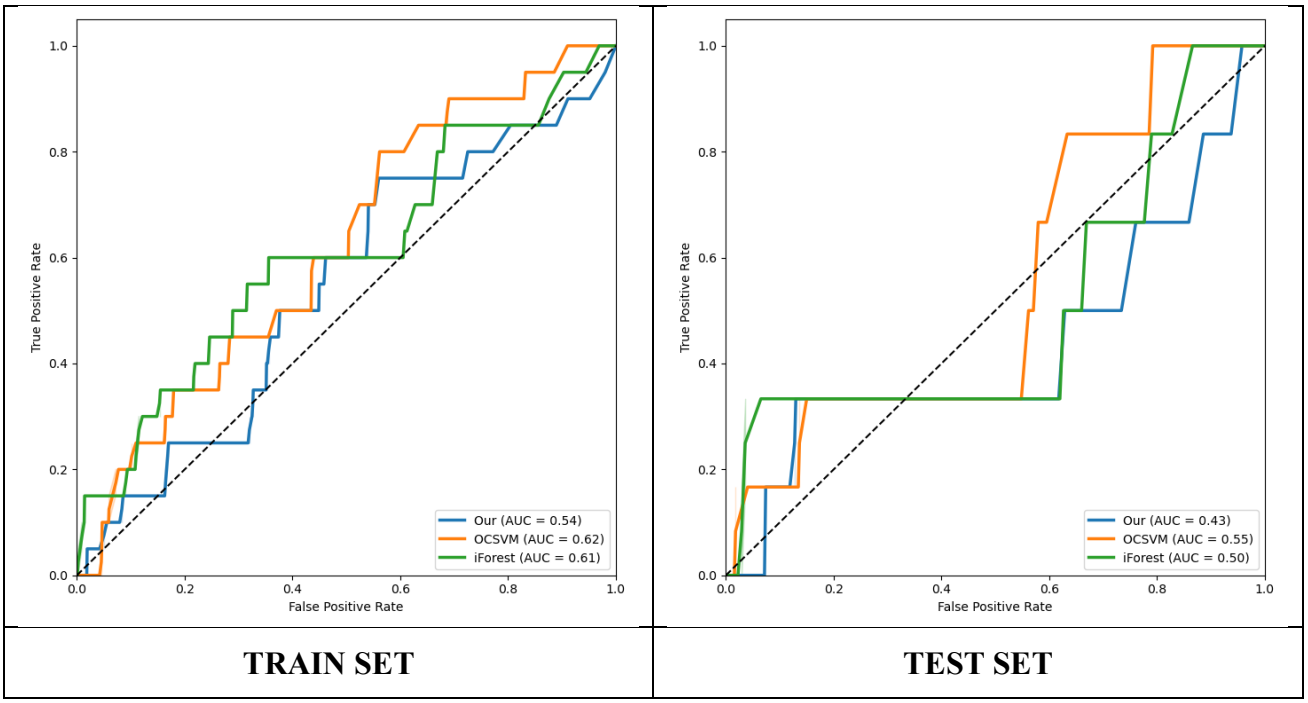


Figure 78. Receiver Operating Characteristic (ROC) curve of Train and Test sets for the three implemented models. Subjects who retained the catheter for fewer than 7 days were excluded from the analysis. Patients who remained free of infection during the first 7 days following catheter placement were classified as healthy. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

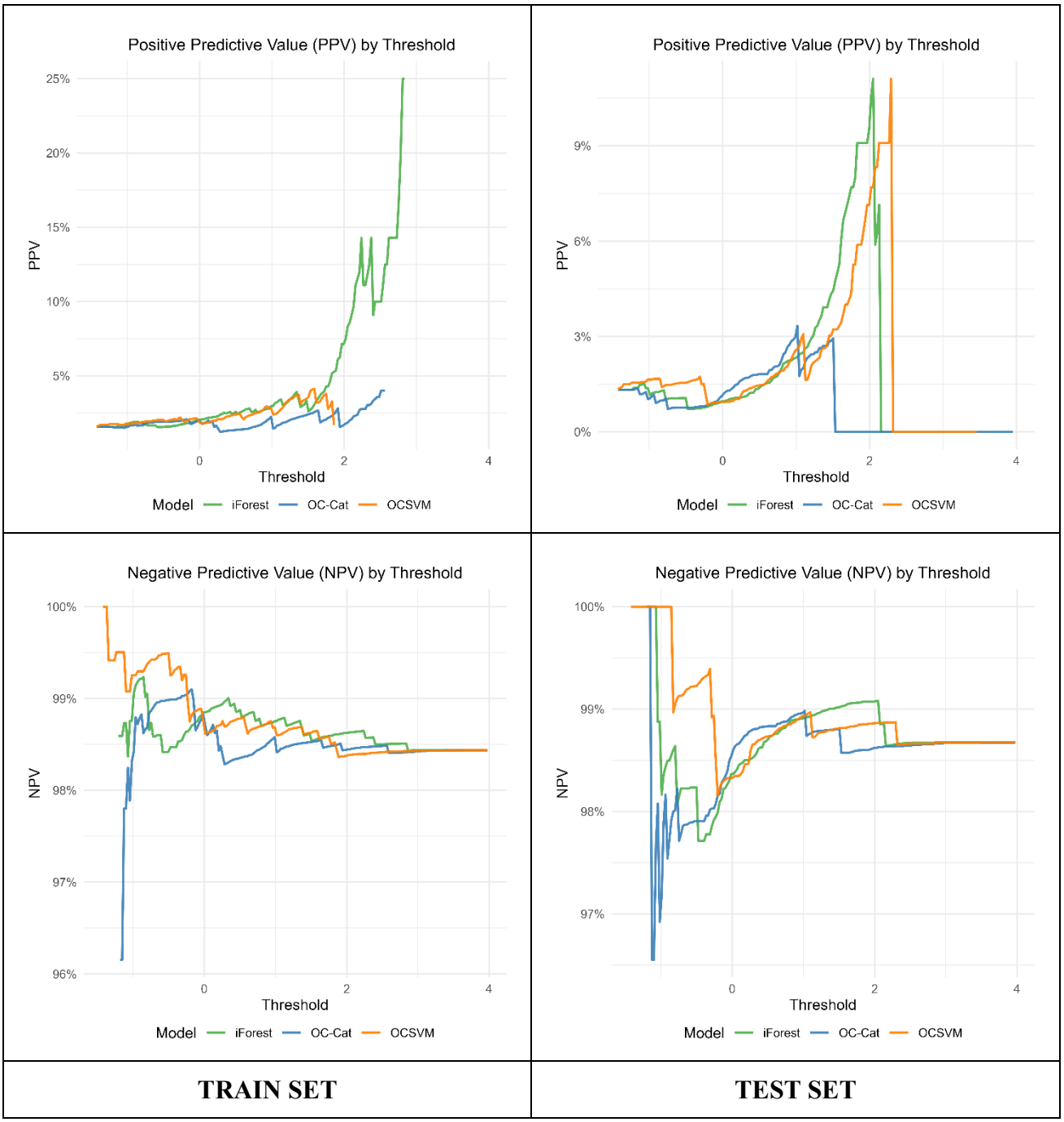


Figure 79. PPV and NPV across probability thresholds for infection risk models on Train and Test sets. Subjects who retained the catheter for fewer than 7 days were excluded from the analysis. Patients who remained free of infection during the first 7 days following catheter placement were classified as healthy. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

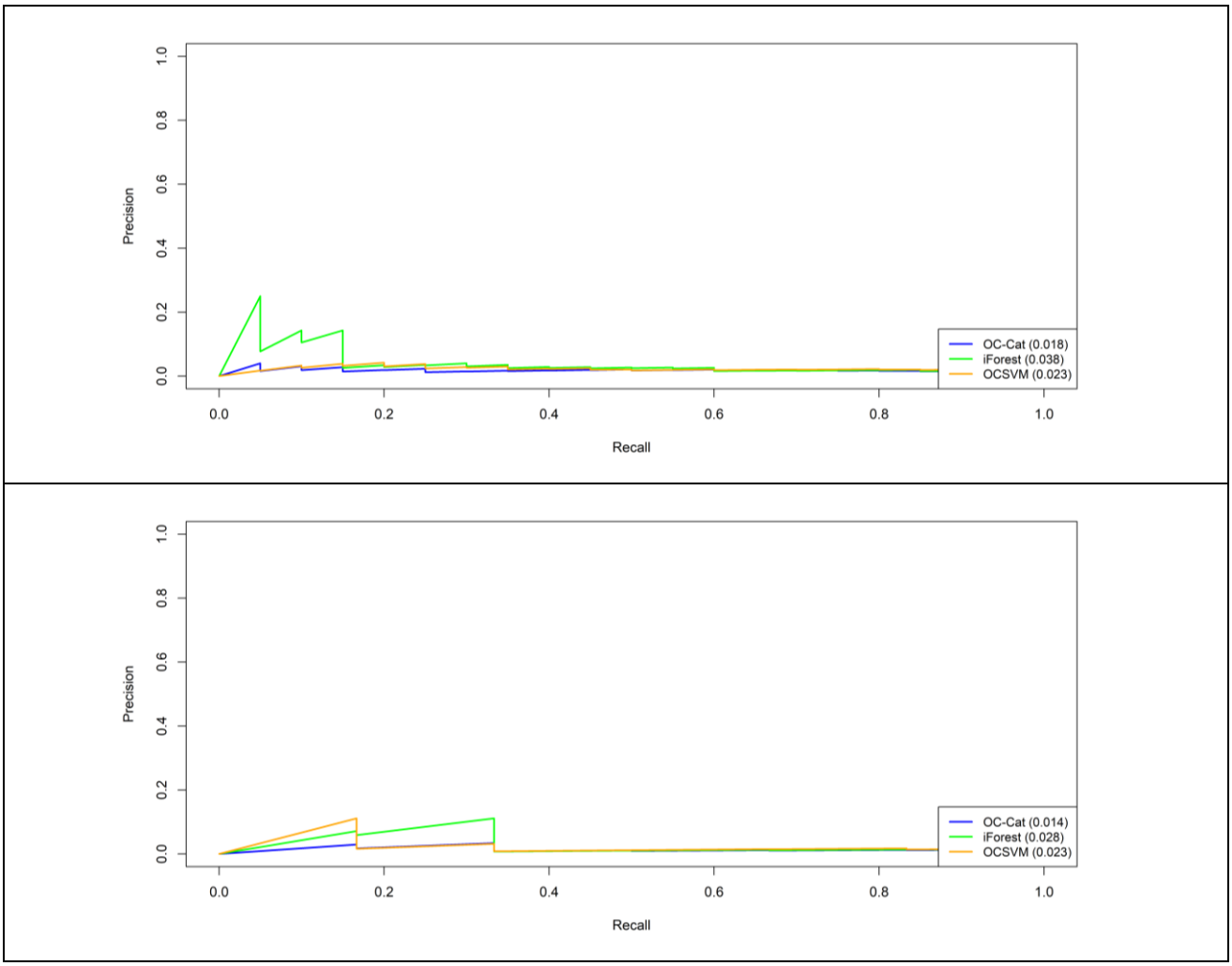


Figure 80. Precision-Recall curves on the Train and Test sets. Subjects who retained the catheter for fewer than 7 days were excluded from the analysis. Patients who remained free of infection during the first 7 days following catheter placement were classified as healthy. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

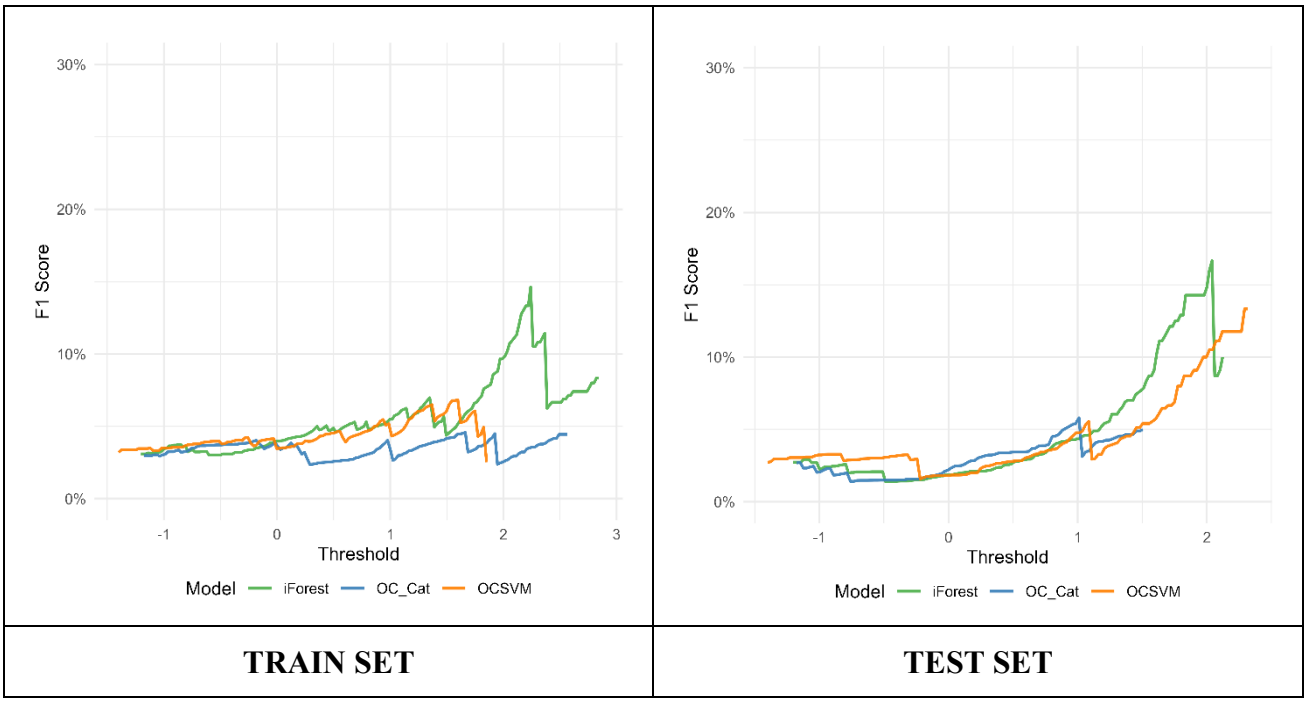


Figure 81. F1 Score curves for the Train and Test sets comparing OC-Cat (blue), iForest (green), and OCSVM (orange) across varying threshold values. Subjects who retained the catheter for fewer than 7 days were excluded from the analysis. Patients who remained free of infection during the first 7 days following catheter placement were classified as healthy. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

	Mann-Whitney *p-value*		Hellinger distance	
	Train	**Test**	**Train**	**Test**
OC-Cat	0.501	0.585	0.15	0.28
iForest	0.101	0.981	0.26	0.34
OCSVM	0.067	0.661	0.18	0.19

Table 13. Comparison of predicted scores between infected and uninfected patients in training and test sets, assessed using the Mann–Whitney U test *(p-values*) and Hellinger distance. Subjects who retained the catheter for fewer than 7 days were excluded from the analysis. Patients who remained free of infection during the first 7 days following catheter placement were classified as healthy.

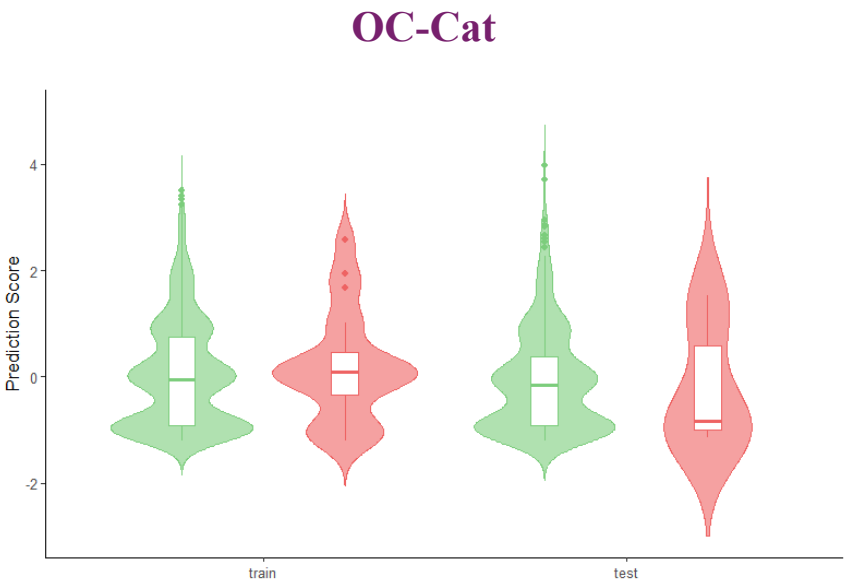


Figure 82. Violin–boxplots showing the OC-Cat model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects. Subjects who retained the catheter for fewer than 7 days were excluded from the analysis. Patients who remained free of infection during the first 7 days following catheter placement were classified as healthy.

Figure 83. Violin–boxplots showing the iForest model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects. Subjects who retained the catheter for fewer than 7 days were excluded from the analysis. Patients who remained free of infection during the first 7 days following catheter placement were classified as healthy.

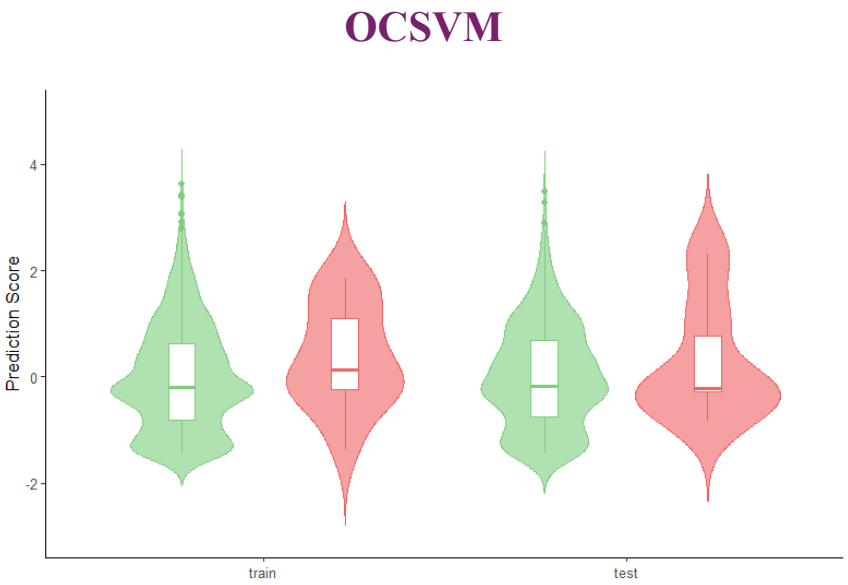


Figure 84. Violin–boxplots showing the OCSVM model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects. Subjects who retained the catheter for fewer than 7 days were excluded from the analysis. Patients who remained free of infection during the first 7 days following catheter placement were classified as healthy.

In the 14-day cut-off analysis, 16 variables out of the initial 33 were retained through graphical feature selection. Chronic kidney disease and type of catheter fixation were features retained within the majority class, whereas pulmonary disease was not (Fig. 85).

The OC-Cat model exhibited a decline in performance on both the training and test sets, accompanied by a corresponding improvement in the benchmark models (Figs 86-92 and Table 14).

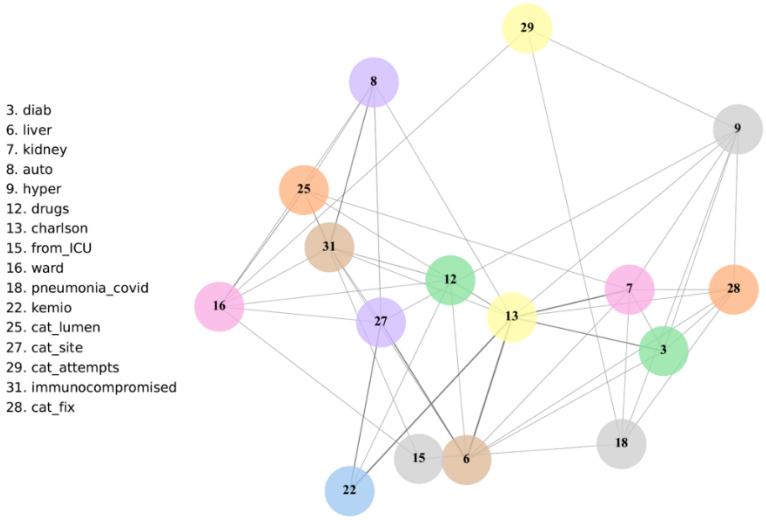


Figure 85. Network graph showing nodes and weighted connections among selected features of the train dataset. Subjects who retained the catheter for fewer than 14 days were excluded from the majority class, while patients who remained free of infection during the first 14 days following catheter placement were counted in the majority class.

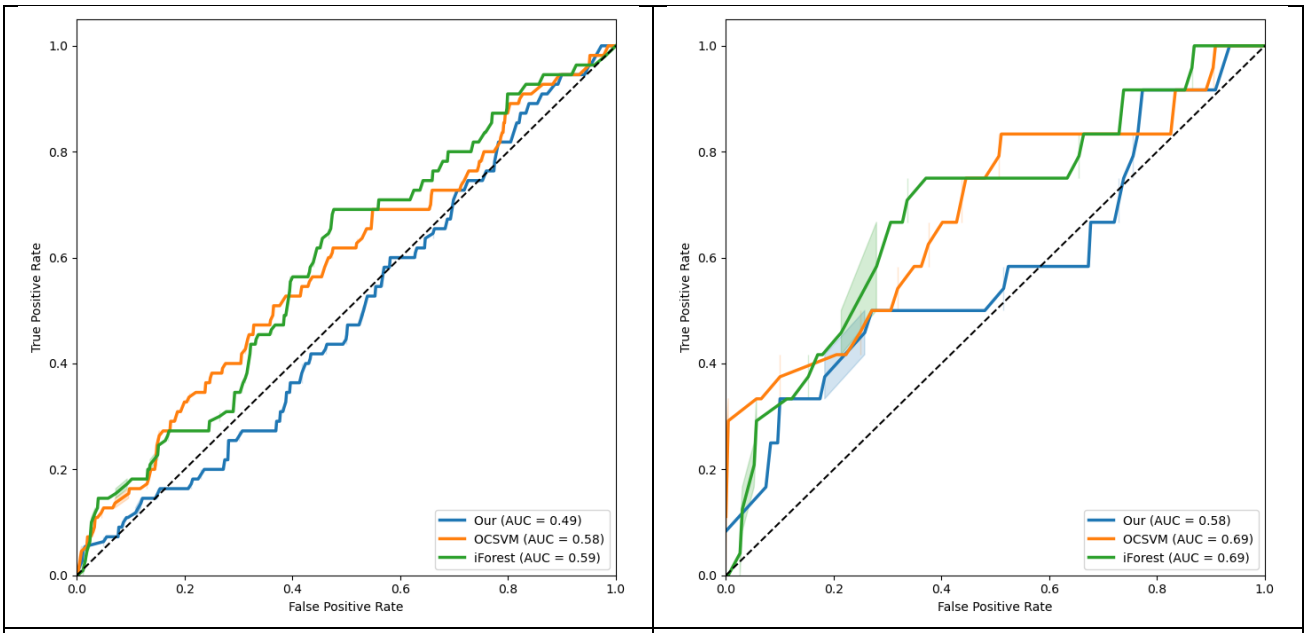


Figure 86. Receiver Operating Characteristic (ROC) curve of Train and Test sets for the three implemented models. Subjects who retained the catheter for fewer than 14 days were excluded from the analysis. Patients who remained free of infection during the first 14 days following catheter placement were classified as healthy. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

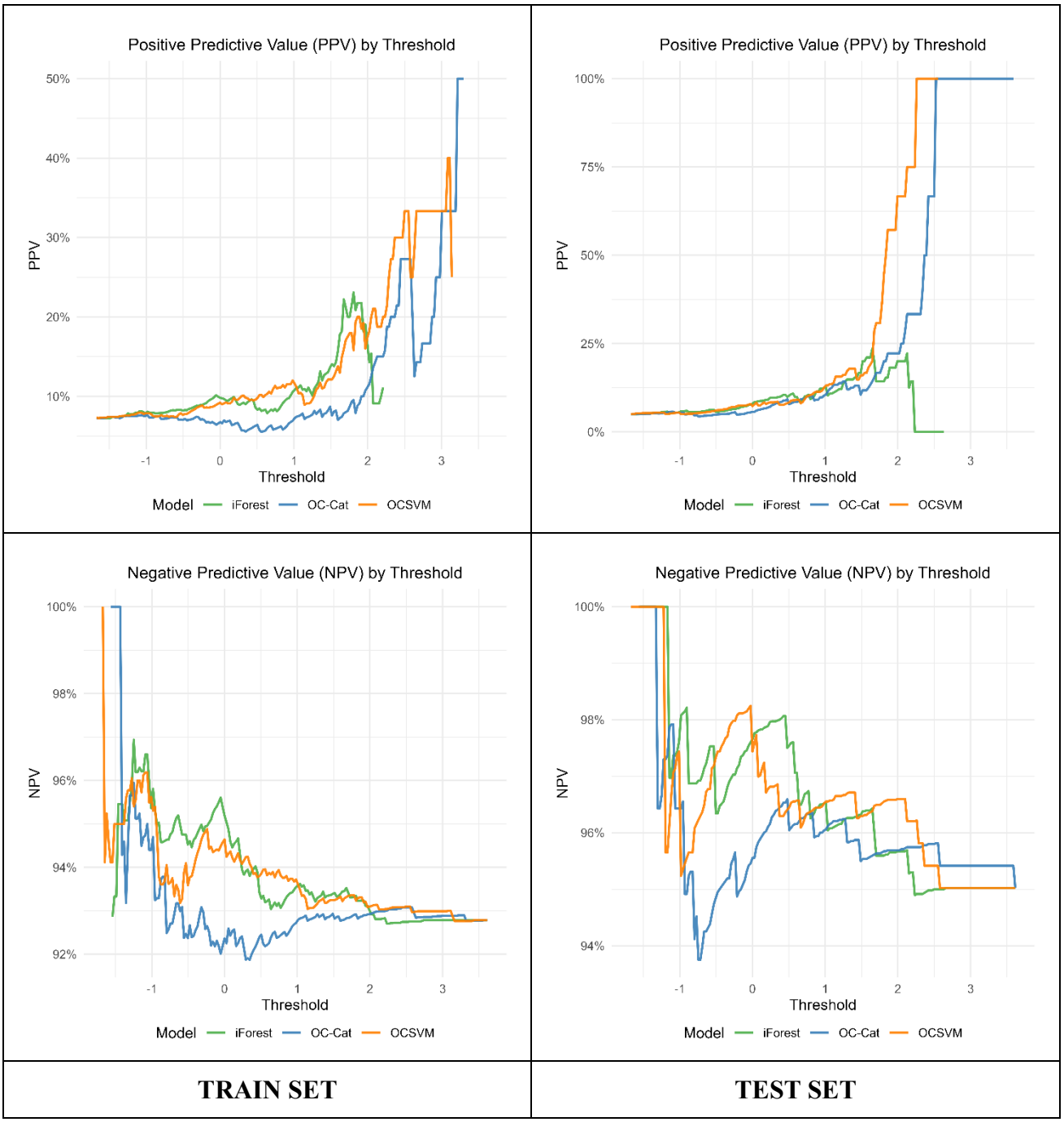


Figure 87. PPV and NPV across probability thresholds for infection risk models on Train and Test sets. Subjects who retained the catheter for fewer than 14 days were excluded from the analysis. Patients who remained free of infection during the first 14 days following catheter placement were classified as healthy. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

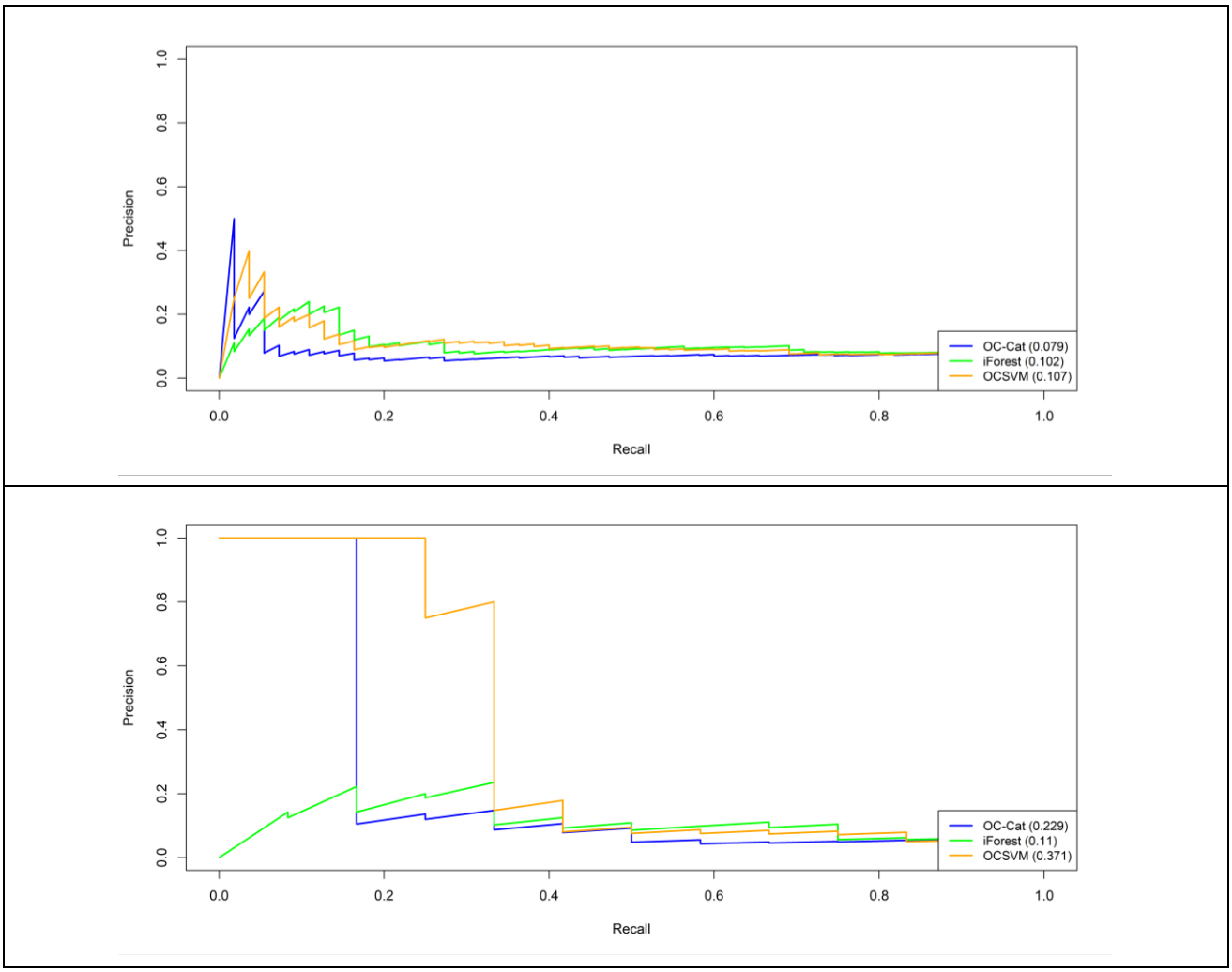


Figure 88. Precision-Recall curves on the Train and Test sets. Subjects who retained the catheter for fewer than 14 days were excluded from the analysis. Patients who remained free of infection during the first 14 days following catheter placement were classified as healthy. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

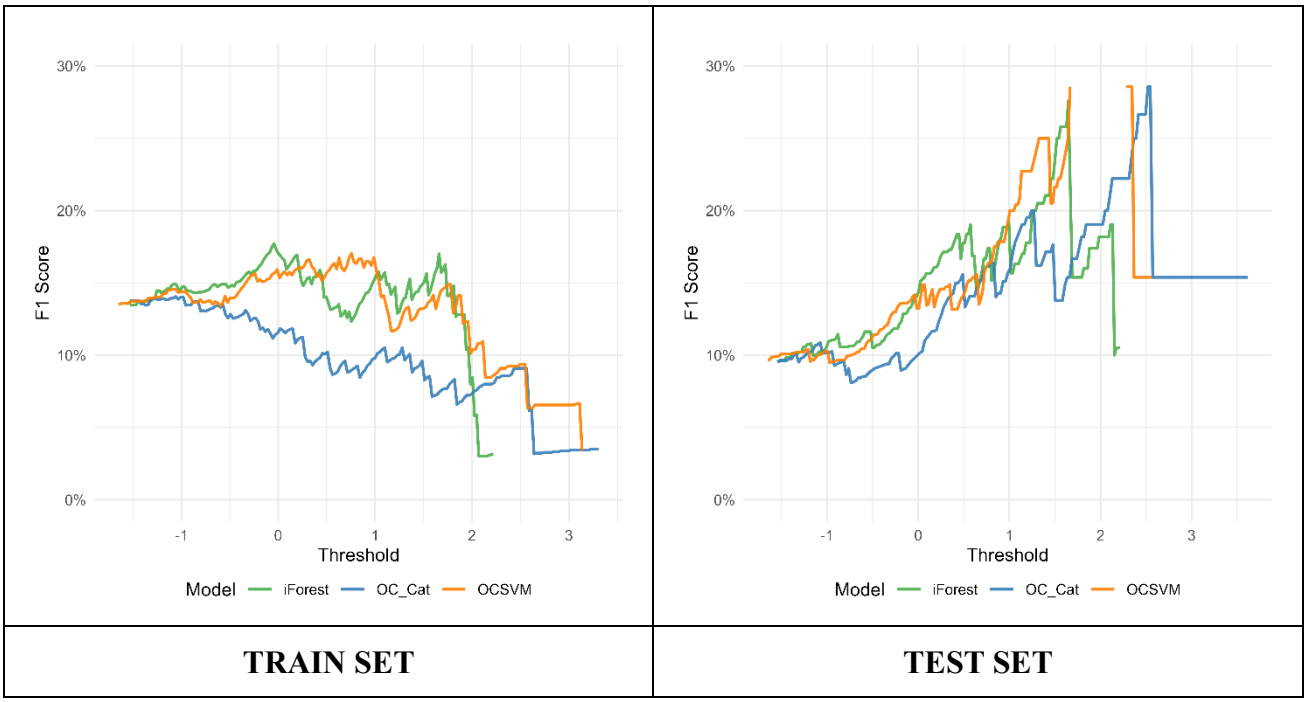


Figure 89. F1 Score curves for the Train and Test sets comparing OC-Cat (blue), iForest (green), and OCSVM (orange) across varying threshold values. Subjects who retained the catheter for fewer than 14 days were excluded from the analysis. Patients who remained free of infection during the first 14 days following catheter placement were classified as healthy. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

	Mann-Whitney *p-value*		Hellinger distance	
	Train	**Test**	**Train**	**Test**
OC-Cat	0.820	0.323	0.09	0.30
iForest	0.028	0.024	0.14	0.25
OCSVM	0.047	0.027	0.14	0.33

Table 14. Comparison of predicted scores between infected and uninfected patients in training and test sets, assessed using the Mann–Whitney U test *(p-values)* and Hellinger distance. Subjects who retained the catheter for fewer than 14 days were excluded from the analysis. Patients who remained free of infection during the first 14 days following catheter placement were classified as healthy.

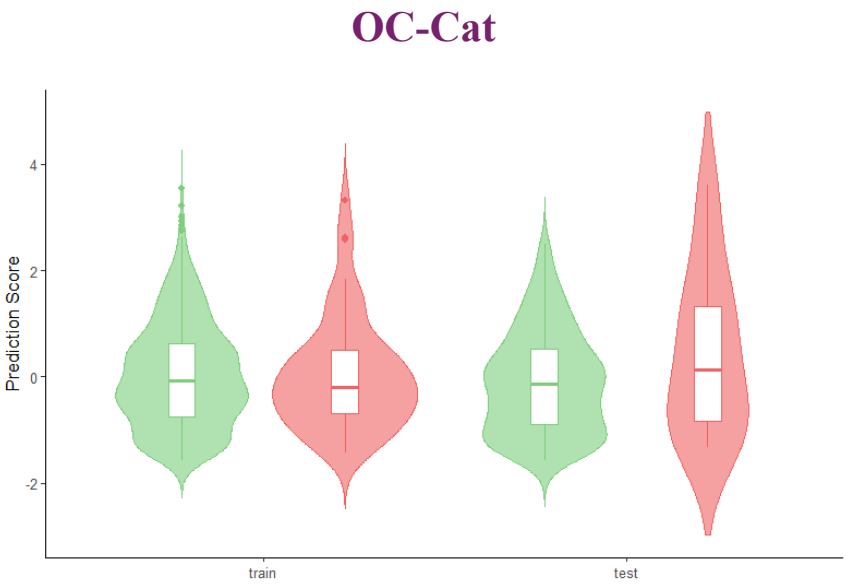


Figure 90. Violin–boxplots showing the OC-Cat model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects. Subjects who retained the catheter for fewer than 14 days were excluded from the analysis. Patients who remained free of infection during the first 14 days following catheter placement were classified as healthy.

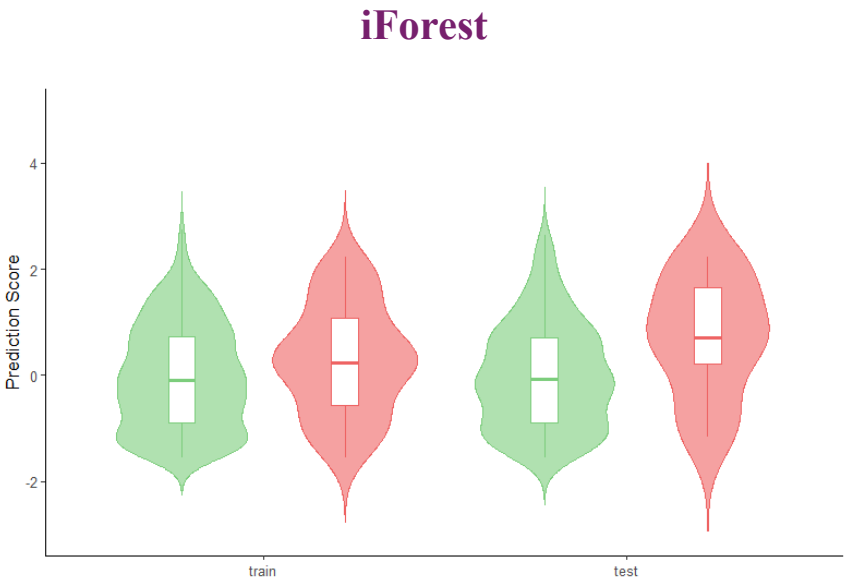


Figure 91. Violin–boxplots showing the iForest model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects. Subjects who retained the catheter for fewer than 14 days were excluded from the analysis. Patients who remained free of infection during the first 14 days following catheter placement were classified as healthy.

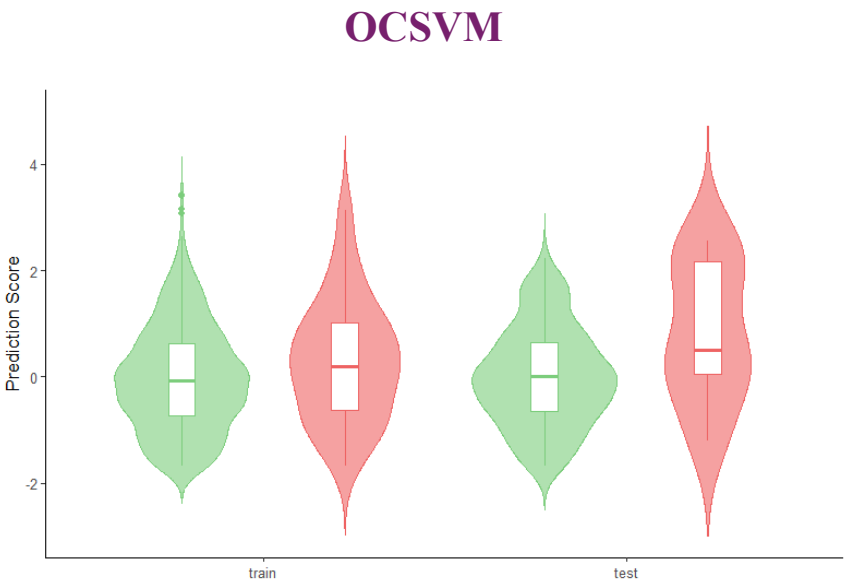


Figure 92. Violin–boxplots showing the OCSVM model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects. Subjects who retained the catheter for fewer than 14 days were excluded from the analysis. Patients who remained free of infection during the first 14 days following catheter placement were classified as healthy.

In the analysis using the 21-day cut-off, 17 of the original 33 variables were retained through graphical feature selection. Relative to the main analysis, cachexia, tracheostomy, psychiatric disorders, and neurological diseases were preserved, whereas pulmonary disease, cirrhosis, the number of catheter lumens, and the number of insertion attempts were excluded (Fig. 93). The performance of the OC-Cat model declined in the training set, whereas, in the test set, its ROC values exceeded those of the other two models (Fig. 94). In both the training and test sets, the OC-Cat model exhibited larger differences between the median values of the prediction distributions than those observed for the 14-day cut-off model, iForest, and OCSVM (Table 15 and Figs. 95-100).

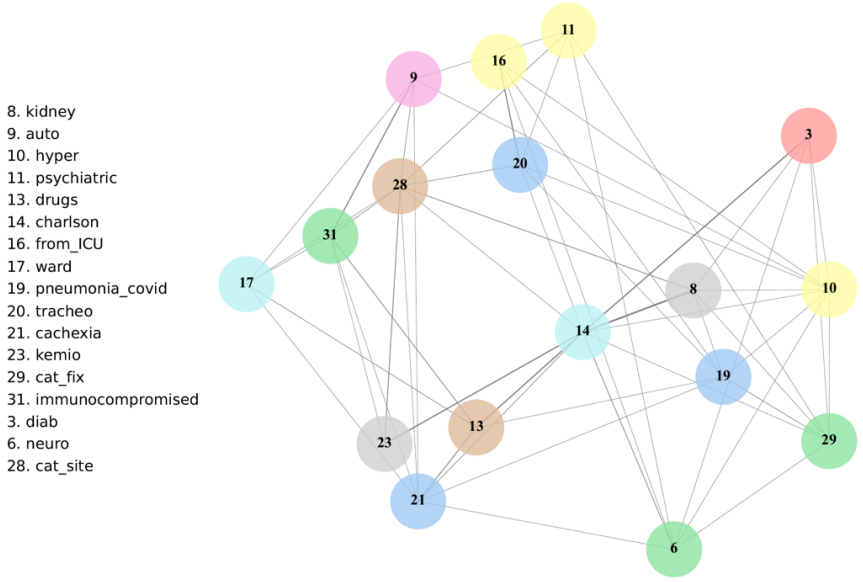


Figure 93. Network graph showing nodes and weighted connections among selected features of the train dataset. Subjects who retained the catheter for fewer than 21 days were excluded from the majority class, while patients who remained free of infection during the first 21 days following catheter placement were counted in the majority class.

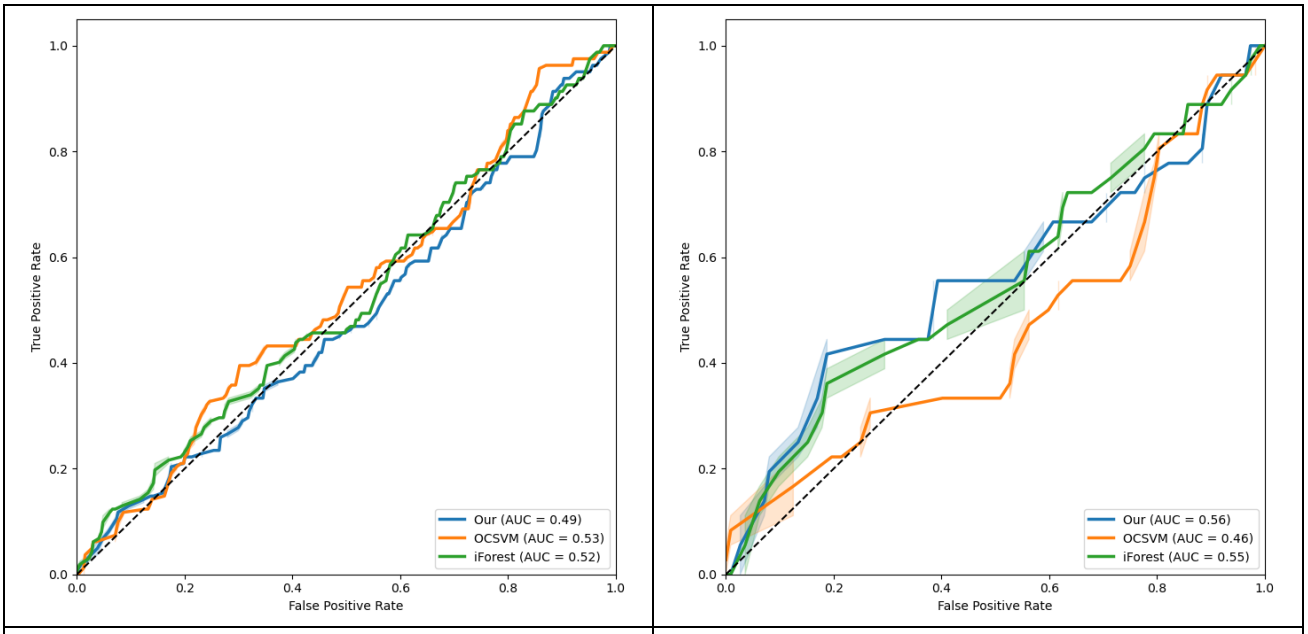


Figure 94. Receiver Operating Characteristic (ROC) curve of Train and Test sets for the three implemented models. Subjects who retained the catheter for fewer than 21 days were excluded from the analysis. Patients who remained free of infection during the first 21 days following catheter placement were classified as healthy..The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

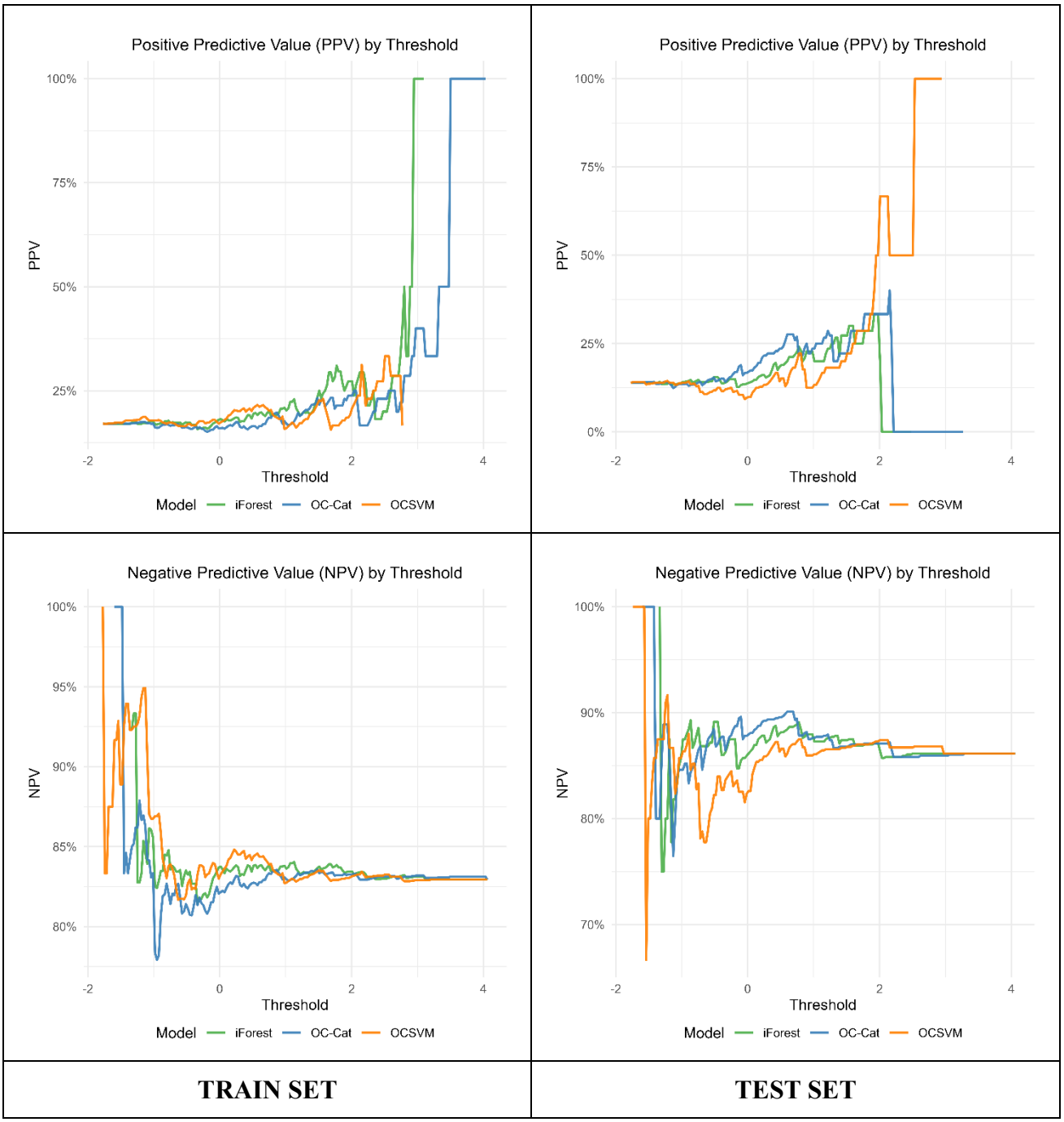


Figure 95. PPV and NPV across probability thresholds for infection risk models on Train and Test sets. Subjects who retained the catheter for fewer than 21 days were excluded from the analysis. Patients who remained free of infection during the first 21 days following catheter placement were classified as healthy. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

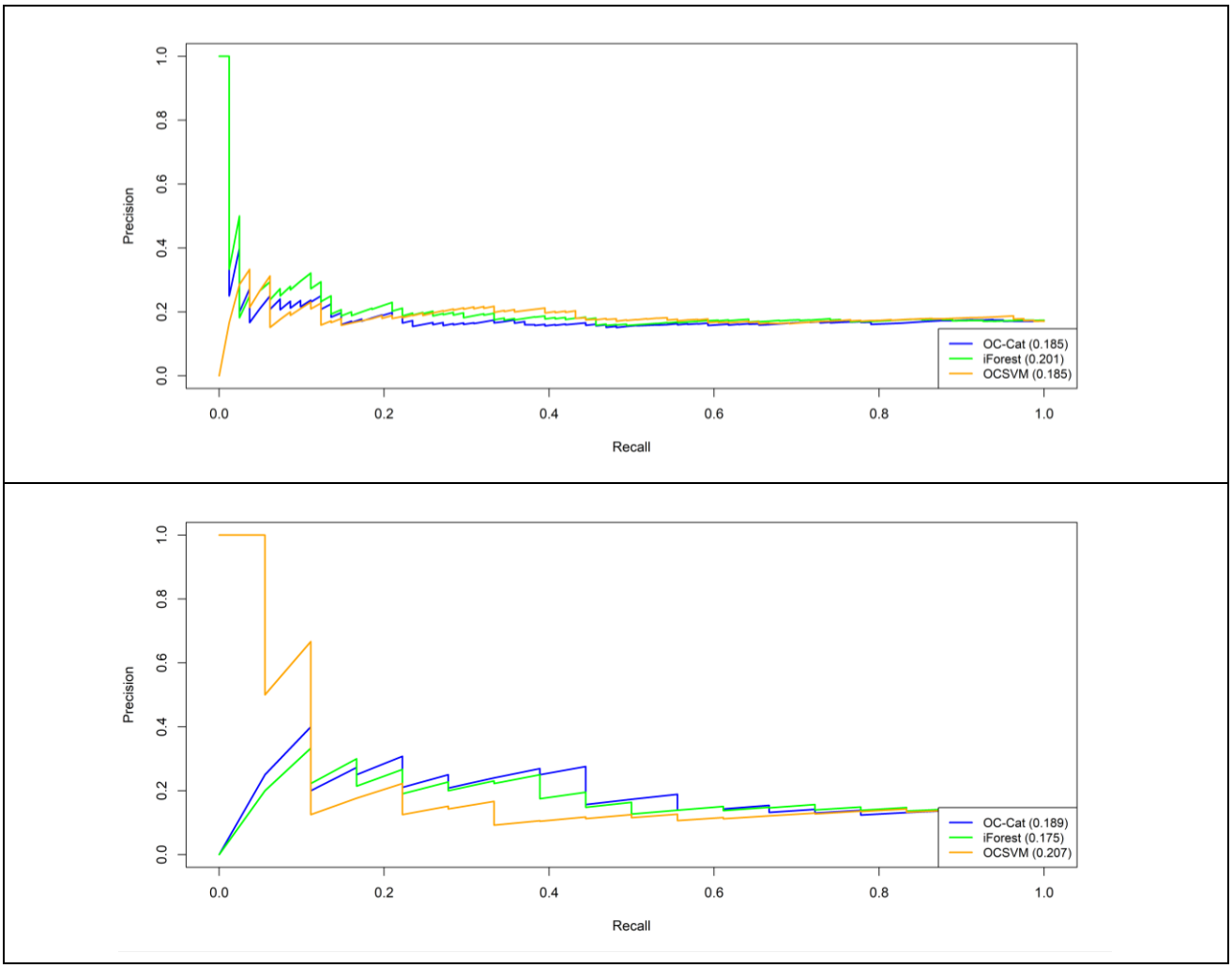


Figure 96. Precision-Recall curves on the Train and Test sets. Subjects who retained the catheter for fewer than 21 days were excluded from the analysis. Patients who remained free of infection during the first 21 days following catheter placement were classified as healthy. The OC-Cat model is represented with a blue line, the iForest model with a green line and the OCSVM with an orange one. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

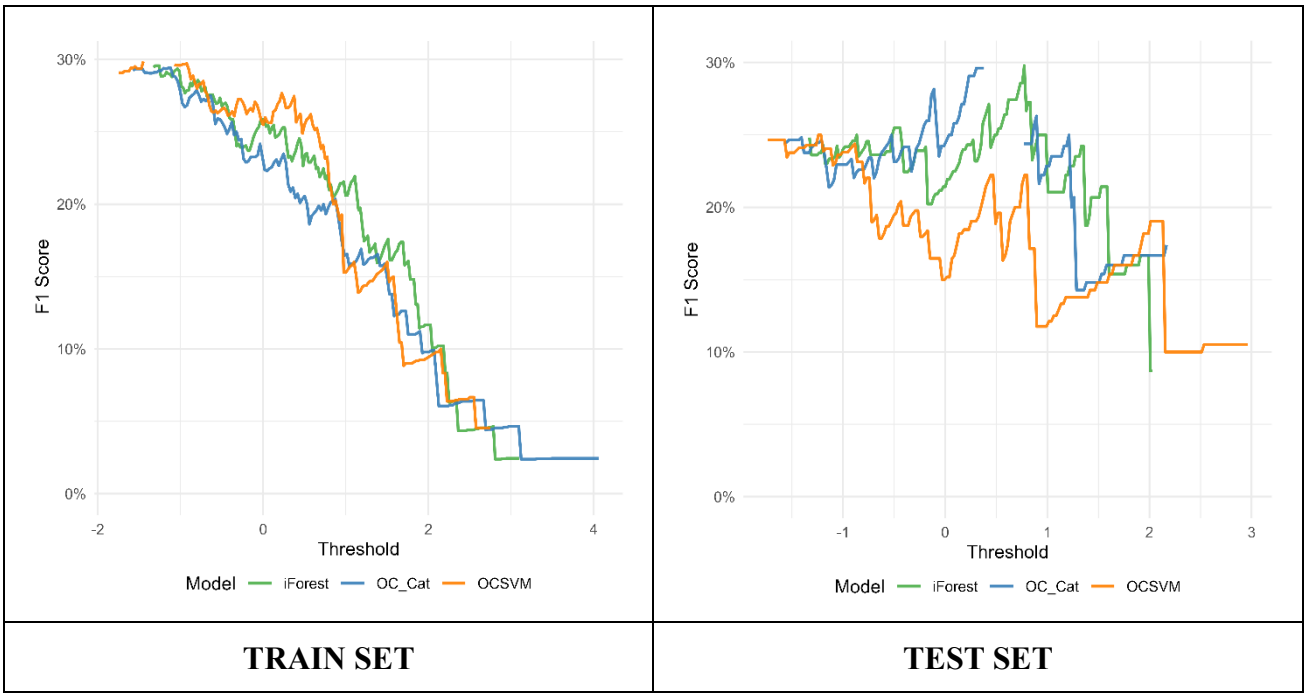


Figure 97. F1 Score curves for the Train and Test sets comparing OC-Cat (blue), iForest (green), and OCSVM (orange) across varying threshold values. Subjects who retained the catheter for fewer than 21 days were excluded from the analysis. Patients who remained free of infection during the first 21 days following catheter placement were classified as healthy. The train set includes all observations up to August 23, while the testing dataset includes observations from that date onwards.

	Mann-Whitney *p-value*		**Hellinger distance**	
	Train	**Test**	**Train**	**Test**
OC-Cat	0.703	0.411	0.08	0.20
iForest	0.616	0.479	0.09	0.14
OCSVM	0.445	0.613	0.07	0.17

Table 15. Comparison of predicted scores between infected and uninfected patients in training and test sets, assessed using the Mann–Whitney U test *(p-values*) and Hellinger distance. Subjects who retained the catheter for fewer than 21 days were excluded from the analysis. Patients who remained free of infection during the first 21 days following catheter placement were classified as healthy.

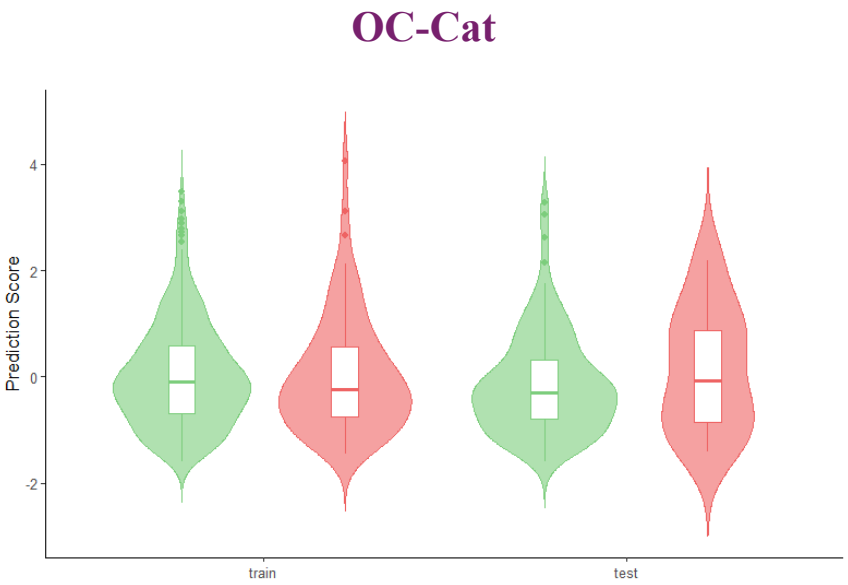


Figure 98. Violin–boxplots showing the OC-Cat model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects. Subjects who retained the catheter for fewer than 21 days were excluded from the analysis. Patients who remained free of infection during the first 21 days following catheter placement were classified as healthy.

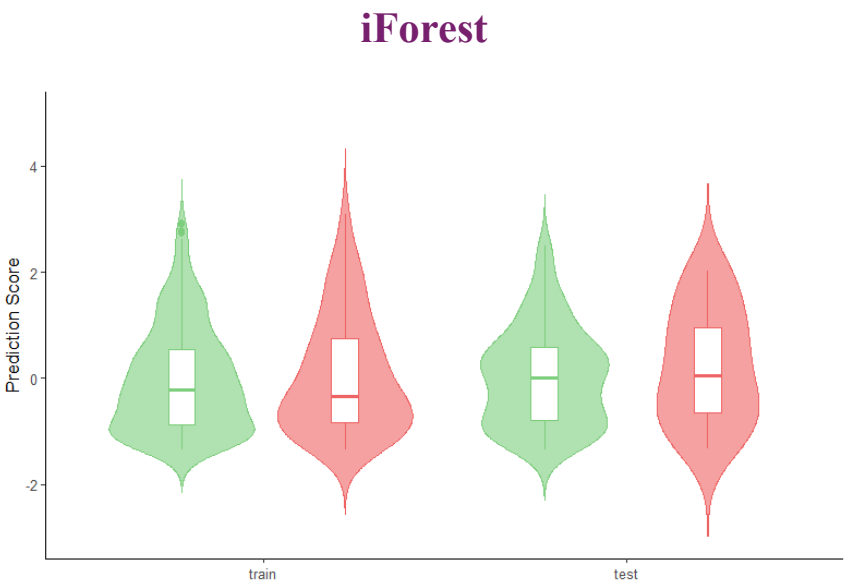


Figure 99. Violin–boxplots showing the iForest model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects. Subjects who retained the catheter for fewer than 21 days were excluded from the analysis. Patients who remained free of infection during the first 21 days following catheter placement were classified as healthy.

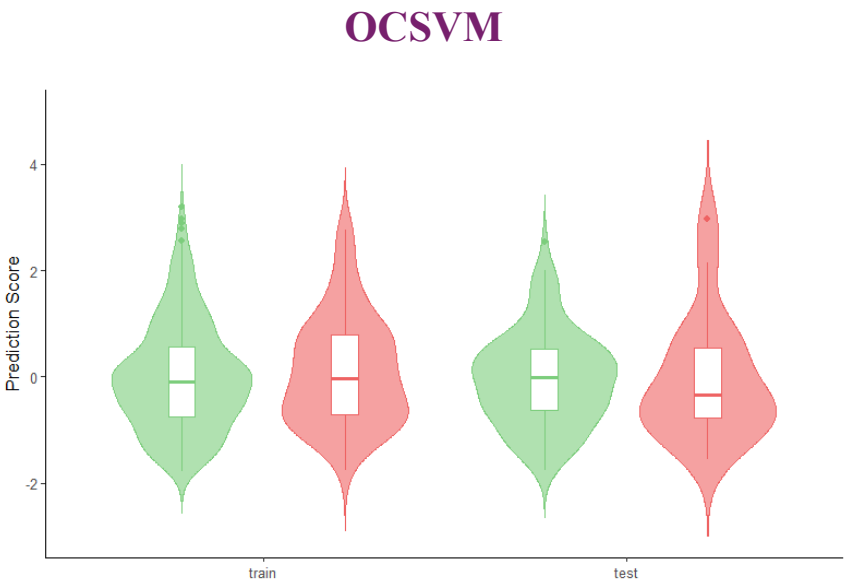


Figure 100. Violin–boxplots showing the OCSVM model's predicted scores for the training and test sets. All scores are normalized. Green violins indicate predictions for uninfected subjects, and red violins indicate predictions for infected subjects. Subjects who retained the catheter for fewer than 21 days were excluded from the analysis. Patients who remained free of infection during the first 21 days following catheter placement were classified as healthy.

Finally, here is the ranking of the features selected for the three different cut-offs. For the 7- and 14-day cut-offs, the ordering of variables remained virtually unchanged relative to the main analyses (Figs. 101, 102). In the 21-day cut-off scenario, transthoracic surgery and cachexia (variables not selected in the main analyses) ranked highly in terms of importance (Fig. 103).

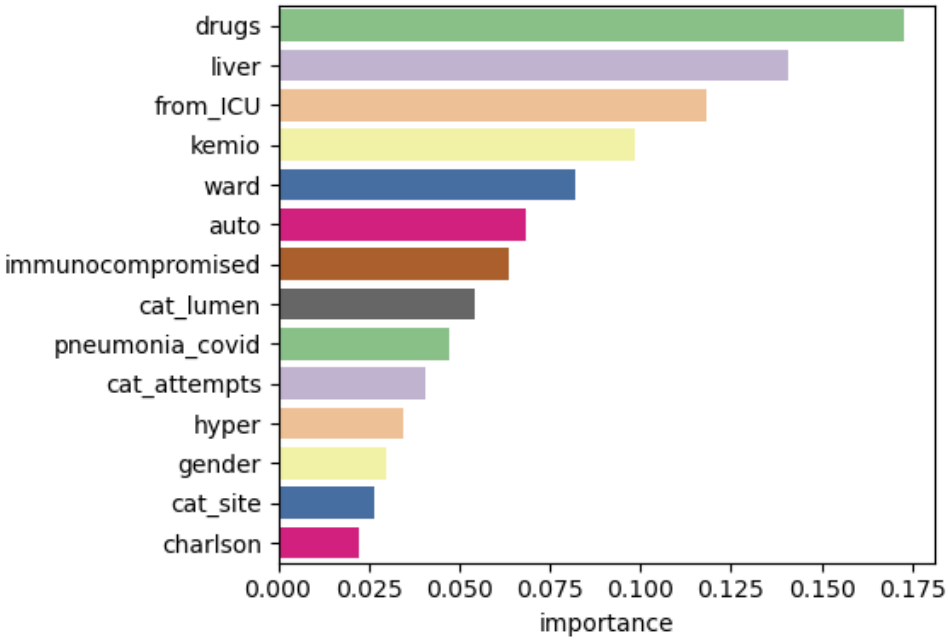


Figure 101. Ranked feature importances considering selected features in the 7-days cut-off scenario, with importance values computed for the majority class in the training set (uninfected patients).

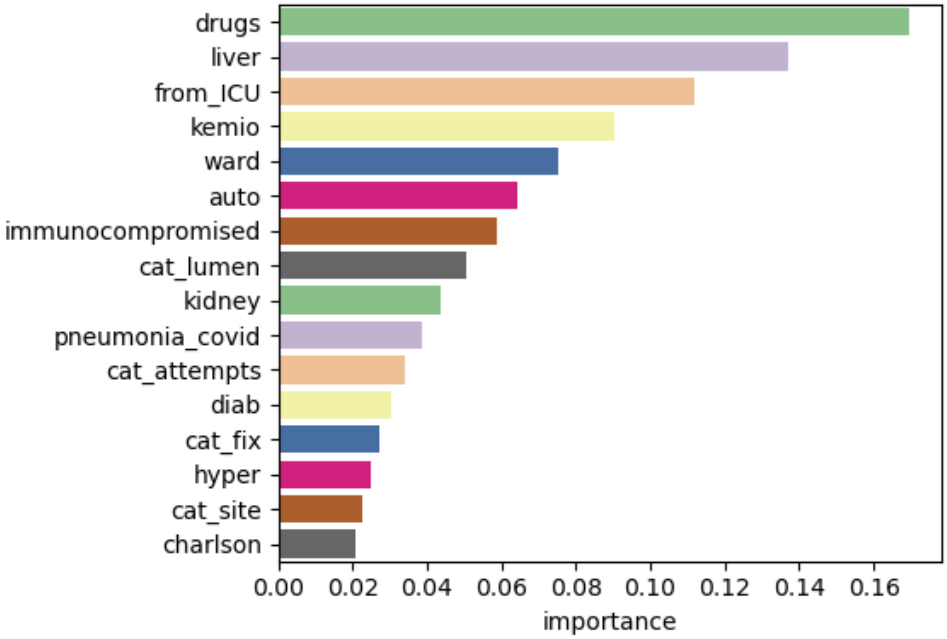


Figure 102. Ranked feature importances considering selected features in the 7-days cut-off scenario, with importance values computed for the majority class in the training set (uninfected patients).

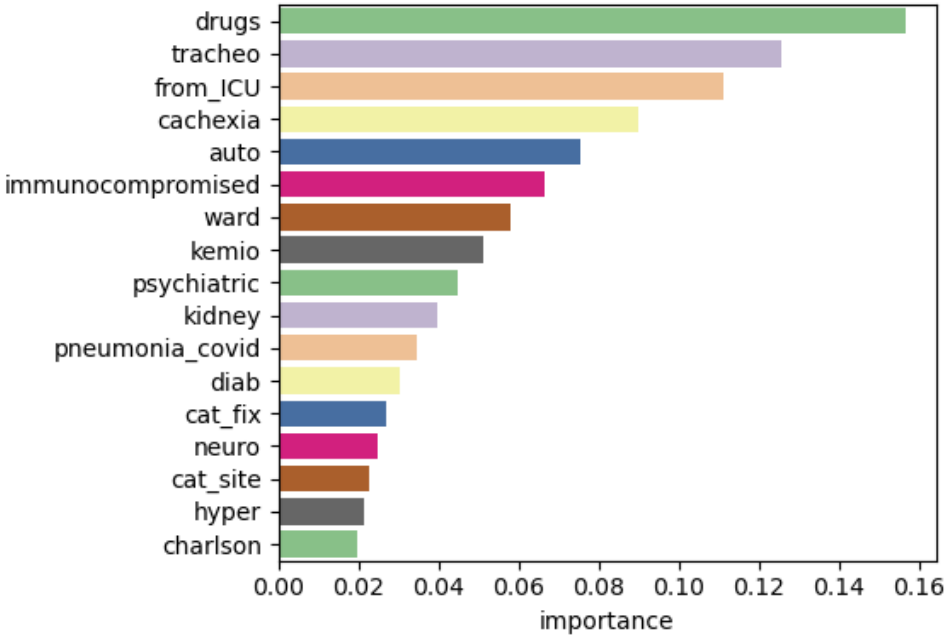


Figure 103. Ranked feature importances considering selected features in the 7-days cut-off scenario, with importance values computed for the majority class in the training set (uninfected patients).

4. Discussion

In this study, the aim is to address the challenges of analysing difficult clinical data using traditional statistical models by instead applying machine learning techniques. From an interpretative perspective, while data related to catheter infections can offer more practical insights from a clinical standpoint, the data concerning Versalis, due to its structure, did not provide adequate interpretability.

The initial project envisaged using not only data on workers' health, but also information on participation in company welfare programs, commuting time, daily physical activity, work tasks, and eating habits. These variables would be used to create clusters through an unsupervised approach. Reports of work-related accidents or near-misses could then be linked to the identified clusters, and the results compared with those obtained from supervised models.

Due to privacy constraints, it was not possible to complete the initial plan for the company research project, which involved applying innovative statistical models directly to the medical record data of ENI workers. Consequently, the available data were aggregated, focusing on employee participation in health monitoring and campaigns conducted by Versalis (an ENI branch) in 2022 and 2023. This constituted a limitation for both the assessment of temporal effects and the detailed interpretation of test types across centers. Because the performance indication codes were frequently vague or ambiguous, a fine-grained interpretation of the resulting clusters was not feasible. Consequently, the analysis was limited to comparing community-based models with a reference benchmark (hierarchical clustering).

Characterizing diagnostic tests proved difficult, as the same test could have been repeated multiple times for the same individual throughout the year. Therefore, values exceeding 100% were capped at 100% to ensure a reasonable and solid approximation. Aggregation of diagnostic tests was intentionally avoided, as linking tests could have artificially reinforced the identified patterns, especially given the scarcity of data and the large number of zero values in the dataset.

Although not all diagnostic procedures were performed in both years (even though they comprised only a minority), the analyses conducted separately for each year yielded a partial overlap of results. Centers not present in 2022 were excluded because their data

were incomplete and of inadequate quality, which would have compromised the direct comparability of the two years under investigation. The analyses for 2022 proved to be more consistent than those for the following year, exhibiting higher modularity values across models and a greater number of models achieving performance values greater than or equal to zero. The matrix computed using the Jaccard method emerged as the most robust across the various models examined. This finding is not unexpected, given that the high discriminative power of this index has been demonstrated in other studies[43,99,100]. The Jaccard index is indeed well suited to dichotomous data, so we discretized our values based on whether or not a site participated in a diagnostic test[27]. This approach, however, could limit the ability to distinguish between sites with consistently high participation and those with only sporadic participation[101]. Such discrimination could be improved by adjusting the participation threshold or by using a weighted approach, as done the weighted Jaccard matrix, where zero values were excluded, giving different weight to intersections with varying levels of participation[101].

In general, performance appeared better when fewer communities were identified. Specifically, the largest and most cohesive community corresponded to the Italian facilities, leaving out the other foreign industrial plants. This may have several explanations: on one hand, higher participation in activities promoted in Italy, the company's home market; on the other, a larger proportion of personnel employed in Italy, accounting for more than 60-70% percent of the workforce considered. The most dispersed group was formed by the foreign sites, where participation rates were relatively low or, in some locations, the number of employees involved was fewer than ten, suggesting that many workers were contract employees not subject to the company's health surveillance protocols.

It is worth emphasizing that in most of the best six network models (distinguished by comparatively higher modularity values), the Sardinian Sarroch plant, for both years, appeared as part of the same community of foreign sites. Moreover, the Mexican plant, in some models related to 2022, showed no connection with the other sites, leading to its exclusion from the plots. Since there is no detailed public information on the production typology of the Sardinian plant, it is not possible to venture definitive interpretative hypotheses on the reason for this different behaviour compared to other Italian sites. Indicatively, we know that some sites are exclusively logistics, while others are actual

refining facilities. The industrial configuration could be one of the major factors explaining the observed results, in addition to geographical location. Moreover, at foreign sites, many employees are on temporary lease and may have had examinations at Italian headquarters either on their return or before departure, potentially explaining their clustering with Italian sites despite geographical distance.

The type of available data certainly does not provide sufficient information to directly assess the clinical impact of occupational health surveillance across the various production sites. However, it could serve as a useful tool for quickly classifying and distinguishing differences between sites in a fast and comprehensive manner. One classic alternative, in fact, would be to describe (including graphically) the participation percentages by site and exam. While this approach can offer a detailed analysis of individual items to address, it does not allow for a quick, at-a-glance identification of common characteristics or critical issues. Obviously, models with a better modularity values and well-defined communities would also allow for an approach that is not only descriptive, as applied in this work, but also inferential. This represents a potential future extension of this work, contingent upon the availability of more detailed information regarding plant site characteristics

In this preliminary analysis, we did not proceed further, as the robustness of the models would not have ensured adequate predictive strength. As indicated by Clauset et al. (2004)[42], modularity values above zero represent deviations from randomness, suggesting a possible community pattern. However, only values above about 0.3 are generally considered a good indicator of significant community structure in a network[56]. The modularity values obtained from our models never exceeded this threshold. It should be noted that in the presented scenario, Walktrap, Infomap and Fluid (especially with >2 predefined groups) algorithms do not directly optimize modularity; nonetheless, to enable comparison between models, this parameter was still calculated and used for ranking purposes.

Arguably, the data at our disposal may have been insufficient in both quality and dimensionality. However, the fact that relatively distinct clusters were obtained may indicate that this approach is genuinely useful for identifying characteristics shared by different production plants, even when using data that, at first glance, appear difficult to interpret on a global scale.

In a paper comparing "traditional" models (k-means, hierarchical clustering, and latent profile analysis) with community detection models based on similarity and dissimilarity matrices, Agelink van Rentergem et al. (2022)[102] found that, on simulated psychiatric data, similarity-based models performed moderately well, but classic latent class models showed greater reliability. They also reported that community detection models were adversely affected by a reduction in the number of features considered. Overall, the Authors concluded that, in their simulations, the arbitrary choice of matrix and model had a substantial impact on the stability of the results. Indeed, our analysis suggest that the choice of similarity/dissimilarity matrix strongly influences both the clustering structure and especially model performance. Jaccard and Dice estimated matrix consistently outperformed other metrics, achieving the highest modularity values.

One persistent challenge, as noted by Fortunato at al. (2016)[103], is the identification of adequate, universal benchmarking metrics for community detection algorithms. The Rand Index, which measures concordance between predicted and ground-truth communities, requires labelled data and thus could not be directly calculated in the present scenario. The only available reference parameter was nationality (foreign vs Italian), which proved both limiting and approximate. Furthermore, it suffers from resolution limits[104].

The second scenario analyses catheter-related infections collected from the "Luigi Sacco" Hospital in Milan. Specifically, this study presents the application of a novel novelty detection model developed as part of this project.

The cost associated with vascular catheter infections, both in economic terms[72,73] and in terms of patients' quality of life[74–76], is substantial, in some cases reaching tens of thousands of dollars per patient. This underscores the importance of developing predictive methods, including innovative modelling approaches, to identify patients at increased risk of catheter infection. Although the impact of such infections is considerable, they are fortunately relatively rare events. The ultimate goal is to reduce their occurrence to zero, or at least to anticipate potential cases through timely catheter removal or appropriate management.

As previously emphasized, the infection rate associated with catheter placement is relatively low[67–69]. This rarity, as noted in recent systematic evidence, poses persistent challenges for model development and statistical power across PICC-CRBSI studies (AUC range 0.67–0.93), where underpowered retrospective designs often limit

generalisability[65].Consequently, identifying risk factors linked to such events can be challenging, as the rarity of occurrence may limit the statistical power of the estimates. In this context, the three-state approach presented in this study has proven to be a valuable tool for deriving both a risk score and a set of parameters that should be monitored during catheter insertion. In practical terms, when the PICC team performs a placement, the model can assist in evaluating, based on variables available from the patient's clinical history, which individuals are at higher or lower risk of developing an infection either directly related to, or occurring concurrently with, the catheter.

With regard to the characterisation of the majority class, the feature selection process retained variables that are both of interest and readily associated with catheter-related infection, such as diabetes, drug use, the age-adjusted Charlson Comorbidity Index, a history of previous admission to the ICU, ongoing chemotherapy, autoimmune diseases or immunosuppression, as well as intrinsic characteristics of catheter insertion, including the insertion site, the number of insertion attempts, and the number of lumens. These align closely with predictors consistently identified in recent nomograms and ML models for PICC-CRBSI (diabetes, dwell time, insertion attempts, lumens) and related devices (e.g., venous access ports, haemodialysis CVCs), underscoring their clinical relevance despite dataset imbalance[65,77,105–109]. As might easily be expected, other variables were instead excluded from this selection, as they were likely to be strongly associated with one another, both from a statistical perspective and from a clinical perspective, with each other. For example, neoplasia and cachexia are conditions that are often associated with a state of secondary immunocompromise, even if only temporary. Likewise, certain catheter characteristics may be redundant with one another, such as the number of lumens, the type of catheter (tunnelled or not), and the site of placement.

The excess over independence metric, akin to χ^2-based dependency measures, effectively captured these redundancies while prioritising informative features, similar to LASSO or random forest importance ranking, as adopted in recent studies[65,107,108]. Through this first step, it was therefore possible to reduce the number of parameters to be included in the models, thereby making the computation less demanding from a computational standpoint. The appropriateness of this selection is demonstrated by the fact that the ROC, Precision-Recall and F1-score curves from the main analyses on the dataset with selected features returned values comparable to those of the models run on the full set of available

features. For the OC-Cat model, in fact, the AUC values were almost identical for predictions on the training set (0.61 *vs* 0.60). They were, however, slightly lower in the analyses on the test set when considering fewer features (0.61 *vs* 0.65). This mirrors findings from studies conducted in neonatal and cancer cohorts, where feature reduction (via Random Forest or LASSO) preserved or even improved test AUC values (0.88–0.98). It should be noted, however, that in those studies the class imbalance was considerably less pronounced than in the present analysis (14% to >20% positive observations)[106,107]. Regarding our analysis, it is worth noting that the other two models used as benchmarks actually performed considerably worse, particularly in the test set, when using the full set of features considered, almost as if they were adversely affected by the redundancy of information available (33 variables) compared with having fewer. Our model, on the other hand, while benefiting from a reduction in the number of variables on which it was trained, demonstrated robustness under this change in conditions. It is also worth noting that, in both the condition with all variables and that with only the 15 selected variables, when CRBSI infections were considered, the distributions of predictions for healthy and infected subjects were markedly better than when only CABSI cases were considered.

Among the possible explanations for this finding is the fact that, by definition, CRBSI is an infection in which the pathogen detected in the blood culture (bacteraemia) is the same as that identified at the catheter tip, thereby providing a confirmed associative link. This stricter causality (vs. CABSI's broader surveillance definition) could explain superior model performance in CRBSI cohorts, as seen in other studies. CABSI, indeed, may represent a poorly defined situation in which the detection of the pathogen in the blood is occasional and may be entirely independent of the patient's clinical condition, instead depending on other factors that were not considered here[110]. For example, from a clinical perspective, a significant source of bias could lie in the administration of empirical antibiotic therapy, either concomitant with or preceding catheter insertion, which could have resolved a pre-existing infectious condition while leaving behind contaminations or bacteraemias unrelated to the catheter itself. Furthermore, the relationship between ward staffing levels and infection was not investigated. In the hospital in question, a previous management analysis had observed that the ratio of healthcare staff to beds per ward was broadly correlated with the type of ward ─ clinical or surgical (variables considered here)[67]. However, this association was not examined in detail in the present analysis due

to the lack of precise monthly or quarterly data, a limitation stemming from various staff reallocations in the post COVID-19 period and the opening or closure of new wards.

It should be observed that the age-adjusted CCI was retained in the analyses, together with certain pathologies that contribute to its composition. As an index, it is, of course, inclusive of a range of clinical conditions; indeed, age (contributing up to three points), cardiovascular and renal diseases, and the occurrence of neoplasia (in this case contributing as many as six points) exert a substantial influence on the score. These variables, however, were not retained in the principal analysis. Within our cohort, approximately three-quarters of patients were aged over sixty-five years, corresponding to additional CCI scores of between two and four points. Diabetes and hepatic disease, by contrast, were retained, most probably because they convey supplementary information beyond their contribution to the CCI calculation. Liver diseases and diabetes have also been identified as significant predictors in several independent investigations, further corroborating their relevance within infection-risk stratification frameworks[109,111].

Within the feature ranking of the selected variables, the CCI was positioned at the lower end, as though indicating that, whilst it was indeed relevant for characterisation, it was less so than other variables. It must be borne in mind that almost all healthy subjects presented with advanced-stage liver disease and more than seventy per cent had diabetes; it is therefore understandable that, in characterising the majority class through our approach, these variables would be highly descriptive of it. The CCI, by contrast, which indeed appeared at the bottom of the feature selection ranking, displayed a more balanced distribution between the classes (approximately fifteen to thirty per cent in each). Interestingly, in our study parenteral nutrition, although not ranked among the lowest of the 33 evaluated features, was removed during the feature-reduction process. This may partly reflect its strong correlation with cachexia, also removed during feature reduction, but closely associated with drug use and CCI, both of which were retained among the 15 selected variables. Moreover, evidence from a French cohort suggests that its role as an independent risk factor, while not entirely excluded, appears less prominent than previously assumed, and furthermore in other studies it has also been excluded as a variable by automatic selection[108,112]

Naturally, this type of selection method ─ whilst effective for markedly imbalanced classes, particularly when new observations are strongly associated with other

combinations of features ─ may prove limiting in circumstances where a more traditional classification approach performs well, namely where there is a balanced ratio between classes. We deliberately refrained from conducting any descriptive association tests between the target variable and the covariates considered, as such procedures did not form part of the methodology underpinning either the feature selection process or the model employed.

An interesting aspect is that the performance of OC-Cat model tended, in general, to be superior to that of the iForest model, whilst, in those instances where it did not surpass it, the results were more comparable to those of the OCSVM model. This may be explained, at least in part, by the manner in which the models operate. Conceptually, the OC-Cat model is more akin to the OCSVM, in the sense that both seek to delineate the majority class, albeit by different means. Naturally, a high degree of redundancy in non-informative features will tend to impair the OCSVM, as it will struggle more to delineate and bound the majority class (the "healthy" group, so to speak) through a separating hyperplane. The iForest, by contrast, is based on a divisive approach, whereby the more rapidly a value is isolated in a terminal leaf, the more readily it is considered an outlier. When such a model is trained on healthy subjects, the splitting points are learned from this group; anomalous values must therefore be more extreme than the splitting point that defined the leaf during training, which may prove limiting. Additionally, the hyperparameters of the benchmark models (iForest and OCSVM) were not deliberately optimized. This choice reflects the fact that the OC-Cat model proposed by Leoni et al. does not require hyperparameter tuning, and therefore offers the practical advantage of being usable 'as-is' even by researchers who may not have the technical expertise needed to properly optimize more complex machine-learning algorithms.

A point for reflection arises with regard to the sensitivity analyses. These analyses were conducted more with the aim of assessing the robustness of the clinical conclusions than that of the model itself, given that the models in question did not take into account the timing of the event, but only a binary yes/no characterisation. Indeed, a subject might not have developed an infection because they were censored by catheter removal ─ whether for necessity or because it was no longer required ─ before an infection could occur, despite having predisposing conditions, in comparison with those who did in fact become infected. The variables that characterised the majority class in the training set were,

broadly speaking, preserved, both in number and in order of importance in step three of the feature ranking. The greatest variations were observed particularly when applying the 21-day cut-off, in which we substantially reduced (indeed, more than halved) the cohort of healthy subjects. This naturally led to a difference in the characterisation of the majority class that can hardly be described as unexpected. The cohort of infected patients, by contrast, remained more consistent and similar to that of the main analyses. In the analysis with the 7-day cut-off, the cohort of healthy subjects was instead augmented by all those whose catheter became infected after seven days, whilst excluding around one hundred subjects who had their catheter removed before an infection could have occurred. Test-set performance was lower for all models, owing to the very small number of infected subjects, and the ROC estimate was consequently less stable. Nevertheless, the median prediction scores ─ largely overlapping between healthy and infected subjects, and in some cases even inverted in the test set ─ would seem to indicate that the clinical characterisation was not robust, suggesting that the excluded subjects who had not shown infection in the first few days would, in fact, have remained healthy.

From a clinical perspective, the results may not appear particularly striking; however, the use of these cut-offs, with the attendant variation in the dataset (infected subjects recoded as healthy, and healthy subjects with short catheter dwell times excluded), may nonetheless have yielded an interesting secondary finding in relation to the OC-Cat model. Specifically, since the performance of the other two models ─ which had not been impressive in the main analyses ─ remained stable in the sensitivity analyses, this may suggest that these models were not performing well under any of the conditions, and would therefore have been unlikely to identify an anomalous observation had one arisen. The OC-Cat model, by contrast, particularly in the analyses using the full set of 33 features rather than the selected subset, demonstrated superior performance in anomaly detection, indicative of a sound characterisation of the healthy class. The loss of performance observed in the sensitivity analyses could, in fact, reinforce the notion that the model had been working well, given that we had deliberately altered an observed condition ─ namely, we had recoded as healthy certain subjects who had in reality become ill, albeit after a considerable number of days. One might therefore surmise that these individuals did indeed possess conditions predisposing them to infection, and that the OC-Cat model was the one that best detected them.

As we have already noted, the model also appears not to be unduly affected by redundancy of information; on the contrary, a greater combination of features evidently allows for better calibration of the weights estimated via Hamming distance, which the classifier uses to define the likelihood parameter for each new observation.

With regard to the estimation of model performance, the use of the ROC curve and the AUC value would appear to be the most comprehensive parameters, albeit based on the concept of varying the decision threshold. The Mann-Whitney test, whilst useful for providing a measure of statistical significance, may not be entirely appropriate, as it considers only the distribution of ranks and does not take into account the potential presence of a high number of false negatives in the prediction of infected patients (anomalies). This could have significant clinical consequences, given that the purpose of developing the model was to identify, at baseline, the probability that a subject would be more likely to develop a catheter-related infection, given their condition at the time of insertion. The Hellinger index, on the other hand, by measuring the distance between the prediction distributions, may prove weak in the presence of scattered values in the predictions for infected patients, compared with a tighter clustering of predictions for healthy subjects, even if, overall, the absolute prediction values are not markedly different.

Given the pronounced class imbalance, ROC curves alone could be insufficient and potentially misleading. Therefore, we calculated also positive predictive value (PPV), negative predictive value (NPV), precision-recall AUC (PR-AUC), and F1-score, which are suited for evaluating performance in unbalanced datasets where negative cases predominate. Across primary analyses, all three models exhibited suboptimal overall performance, a predictable consequence of the fundamental challenges in rare-event prediction. Importantly, PPV, NPV, precision-recall and F1-score curves were generated using standardized model scores to enable direct comparability across architectures, a pragmatic methodological choice among several possible approaches, despite introducing a potential limitation. Specifically, while the OC-Cat model natively outputs calibrated event probabilities, the benchmark models produce uncalibrated scores requiring transformation. This standardization may have disadvantaged OC-Cat's inherent probabilistic framework, but it consistently outperformed both benchmarks, demonstrating superior PPV (i.e. better identification of true CRBSI cases among

predicted positives) and higher PR-AUC (demonstrating a stronger precision-recall trade-off across thresholds, critical for clinical decision-making). Performance again showed degradation across test set and sensitivity analyses, attributable to reduced statistical power and cohort shift. This underscores the critical need for external validation in other hospital settings. All models exhibited excellent NPV (almost > 90%), being able to confidently rule out CRBSI/CABSI in low-risk patients. This supports safe clinical deployment for negative predictions, minimizing unnecessary interventions (catheter removal, cultures, antibiotics) while focusing resources on the smaller PPV-constrained positive predictions.

Finally, the presentation of the feature ranking should not be interpreted as being at odds with the graphical feature selection approach. The two approaches employ different methods: one favouring independence between features, the other the contribution of entropy that each feature provides to the characterisation of the majority class. In a certain sense, the feature ranking ─ whose purpose was to provide an explanation of the extent to which each individual variable contributed to defining the majority class, and thus indirectly to the OC-Cat model ─ could, at the same time, be transformed into a feature selection method, should one define an arbitrary cut-off value or apply the elbow rule (using the inflection point to decide which variables to retain).

5. Conclusion

The work was conducted along two parallel lines, both of which, however, were underpinned by a common idea: namely, to identify, through a graphical approach combined with the concept of information redundancy (similarity or dissimilarity), characteristics shared by the values present in the original dataset.

In the first approach, the principal difficulty lay in having access only to aggregated data. Whilst this limitation inevitably affected the scope for analysis and, above all, the interpretation of results, it nonetheless suggested that the community detection model approach could serve as a means of identifying groupings within a dataset in which the sheer number of variables might otherwise constrain a global descriptive view of the potential connections present.

In the second approach, the challenge instead arose from the low number of events in the target class. Here, the nodes, rather than representing production sites, represented the variables in the dataset that described individual records, and the connections between them were weighted according to their degree of "independence".

The study has thus demonstrated the potential of representing indirect connections graphically as a useful tool in the healthcare and global health domains for addressing specific questions, moving beyond the traditional direct graphical representation exemplified, for instance, by Directed Acyclic Graphs (DAGs).

The possibility of linking the groups obtained from the community detection analysis to an organisational outcome ─ such as the number of workplace injuries or other performance parameters ─ represents the natural next step for this approach. As for the OC-Cat model, its application to other highly imbalanced clinical datasets, in which the target variable is less dependent on time and the number of initial covariates is considerably greater, constitutes the next step towards strengthening the evidence observed thus far.

6. References

1. Dhanda, S. S. *et al.* Advancement in public health through machine learning: a narrative review of opportunities and ethical considerations. *Journal of Big Data* 12, 154 (2025).
2. Panteli, D. *et al.* Artificial intelligence in public health: promises, challenges, and an agenda for policy makers and public health institutions. *The Lancet Public Health* 10, e428–e432 (2025).
3. Pinto, A. D. *et al.* Machine Learning Applications in Population and Public Health: Guidelines for Development, Testing, and Implementation. *JMIR Public Health and Surveillance* 11, e68952 (2025).
4. Xu, D. & Xu, Z. Machine learning applications in preventive healthcare: A systematic literature review on predictive analytics of disease comorbidity from multiple perspectives. *Artificial Intelligence in Medicine* 156, 102950 (2024).
5. Hairol Anuar, S. H., Abas, Z. A., Mukhtar, M. F. & Miswan, N. H. Community Detection in Practice: A Review of Real-World Applications Across Six Themes. *IJARBSS* 14, Pages 953-996 (2024).
6. Tokala, S., Enduri, M. K., Lakshm, T. J. & Hajarathaiah, K. Evaluating Community Detection Algorithms: A Focus on Effectiveness and Efficiency. *JSR* 14, 62–74 (2025).
7. Eze, P. U., Geard, N., Mueller, I. & Chades, I. Anomaly Detection in Endemic Disease Surveillance Data Using Machine Learning Techniques. *Healthcare* 11, 1896 (2023).
8. Chunaev, P. Community detection in node-attributed social networks: A survey. *Computer Science Review* 37, 100286 (2020).

9. Rostami, M., Oussalah, M., Berahmand, K. & Farrahi, V. Community Detection Algorithms in Healthcare Applications: A Systematic Review. *IEEE Access* 11, 30247–30272 (2023).
10. Khawaja, F. R., Zhang, Z., Memon, Y. & Ullah, A. Exploring community detection methods and their diverse applications in complex networks: a comprehensive review. *Soc. Netw. Anal. Min.* 14, 115 (2024).
11. Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Physics Reports* 659, 1–44 (2016).
12. Newman, M. *Networks*. (Oxford University Press, 2018). doi:10.1093/oso/9780198805090.001.0001.
13. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* 69, 026113 (2004).
14. Carlsson, G., & Mémoli, F. Characterization, stability and convergence of hierarchical clustering methods. *Journal of Machine Learning Research, 11,* Article 1425-1470 (2010).
15. Mantrach, A. *et al.* The Sum-over-Paths Covariance Kernel: A Novel Covariance Measure between Nodes of a Directed Graph. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 1112–1126 (2010).
16. Luxburg, U. von. A tutorial on spectral clustering. *Stat Comput* 17, 395–416 (2007).
17. Li, J. *et al.* A comprehensive review of community detection in graphs. *Neurocomputing* 600, 128169 (2024).
18. Kojaku, S., Radicchi, F., Ahn, Y.-Y. & Fortunato, S. Network community detection via neural embeddings. *Nat Commun* 15, 9446 (2024).

19. Pinheiro, D; Hartman, R; Romero, E; Menezes, R; Cadeiras, M (2020). Network-Based Delineation of Health Service Areas: A Comparative Analysis of Community Detection Algorithms. University of Exeter.
20. Han, J., Wan, N., Horns, J. J. & McCrum, M. L. Application of Community Detection Methods to Identify Emergency General Surgery–Specific Regional Networks. *JAMA Netw Open* 7, e2439509 (2024).
21. Faisandier, L. *et al.* A network based approach or surveillance of occupational health exposures. *arXiv*.
22. Friesen, M. C. *et al.* Using hierarchical cluster models to systematically identify groups of jobs with similar occupational questionnaire response patterns to assist rule-based expert exposure assessment in population-based studies. *Ann Occup Hyg* 59, 455–466 (2015).
23. Contreras-Valenzuela, M. R. & Martínez-Ibanez, C. A. Hierarchical clustering analysis of musculoskeletal stress factors and their risk level in cardboard manufacturing: research from PLIBEL. *J Occup Health* 66, uiae008 (2024).
24. Liu, Z. An operational risk assessment method for petrochemical plants based on deep learning. *Front. Energy Res.* 12, (2024).
25. Shrivastava, P. Comparison of Cosine, Euclidean Distance and Jaccard Distance. *International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET)*, 3(8), 712-715 (2017).
26. Afzali, M. & Kumar, S. Comparative Analysis of Various Similarity Measures for Finding Similarity of Two Documents. *IJDTA* 10, 23–30 (2017).
27. Travieso, G., & Benatti, A. An Analytical Approach to the Jaccard Similarity Index. *arXiv* (2024).

28. Luciano da Fontoura, C. Further Generalizations of the Jaccard Index. *arXiv* (2021).
29. Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter/Kongelige Danske Videnskabernes Selskab*, 5(4), 1-34.
30. Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* 26, 297–302 (1945).
31. Daniel, W. W. *Applied Nonparametric Statistics*. (PWS-KENT Pub., 1990).
32. Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications*, 19(4), 497-515 (2010).
33. Paninski, L. Estimation of Entropy and Mutual Information. *Neural Computation* 15, 1191–1253 (2003).
34. Bavaud, F., Chappelier, J.-C. & Kohlas, J. An introduction to information theory and applications. *University of Lausanne* (2004).
35. Mahmoudi, A. & Jemielniak, D. Proof of biased behavior of Normalized Mutual Information. *Sci Rep* 14, 9021 (2024).
36. Bradley, E. L. Overlapping Coefficient. in *Encyclopedia of Statistical Sciences* (John Wiley & Sons, Ltd, 2006).
37. Shawe-Taylor, J., & Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press (2004).
38. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7821–7826 (2002).
39. Freeman, L. C. A Set of Measures of Centrality Based on Betweenness. *Sociometry* 40, 35–41 (1977).

40. Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* 45, 167–256 (2003).
41. Brandes, U. A faster algorithm for betweenness centrality*. *The Journal of Mathematical Sociology* 25, 163–177 (2001).
42. Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* 70, 066111 (2004).
43. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008, P10008 (2008).
44. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 9, 5233 (2019).
45. Peixoto, T. P. *Descriptive vs. Inferential Community Detection in Networks: Pitfalls, Myths and Half-Truths*. (Cambridge University Press, Cambridge, 2023).
46. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A.* 105, 1118–1123 (2008).
47. Pons, P. & Latapy, M. Computing Communities in Large Networks Using Random Walks. in *Computer and Information Sciences - ISCIS 2005* (eds Yolum, pInar, Güngör, T., Gürgen, F. & Özturan, C.) 284–293 (Springer, Berlin, Heidelberg, 2005).
48. Reichardt, J. & Bornholdt, S. Statistical Mechanics of Community Detection. *Phys. Rev. E* 74, 016110 (2006).
49. Reichardt, J. & Bornholdt, S. Detecting Fuzzy Community Structures in Complex Networks with a Potts Model. *Phys. Rev. Lett.* 93, 218701 (2004).
50. Traag, V. A., Dooren, P. V. & Nesterov, Y. Narrow scope for resolution-limit-free community detection. *Phys. Rev. E* 84, 016114 (2011).

51. Parés, F. *et al.* Fluid Communities: A Competitive, Scalable and Diverse Community Detection Algorithm. *arXiv* (2017).

52. Ward Jr., J. H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58, 236–244 (1963).

53. Murtagh, F. & Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J Classif* 31, 274–295 (2014).

54. Sohil, F., Sohali, M. U. & Shabbir, J. An introduction to statistical learning with applications in R: by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7. *Statistical Theory and Related Fields* 6, 87–87 (2022).

55. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, 53–65 (1987).

56. Newman, M. E. J. Fast algorithm for detecting community structure in networks. *Phys. Rev. E* 69, 066133 (2004).

57. Salmi, M., Atif, D., Oliva, D., Abraham, A. & Ventura, S. Handling imbalanced medical datasets: review of a decade of research. *Artif Intell Rev* 57, 273 (2024).

58. Ishwaran, H. & O'Brien, R. Commentary: The Problem of Class Imbalance in Biomedical Data. *J Thorac Cardiovasc Surg* 161, 1940–1941 (2021).

59. López, V., Fernández, A., García, S., Palade, V. & Herrera, F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences* 250, 113–141 (2013).

60. Kaur, H., Pannu, H. S. & Malhi, A. K. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. *ACM Comput. Surv.* 52, 79:1-79:36 (2019).
61. Luu, J. *et al.* Practical guide to building machine learning-based clinical prediction models using imbalanced datasets. *Trauma Surg Acute Care Open* 9, e001222 (2024).
62. Emir, Ş. An Investigation of Anomaly Detection Methods in Machine Learning for High Dimensional Datasets. in *Global Studies on Management Information Systems* 227–254 (Istanbul University Press, 2023).
63. Tabassum, M. *et al.* Anomaly-based threat detection in smart health using machine learning. *BMC Med Inform Decis Mak* 24, 347 (2024).
64. Breunig, M. M., Kriegel, H.-P., Ng, R. T. & Sander, J. LOF: identifying density-based local outliers. *SIGMOD Rec.* 29, 93–104 (2000).
65. Cao, L. *et al.* Models for predicting the risk of bloodstream infections associated with peripherally inserted central venous catheters: A scoping review. *PLOS ONE* 20, e0333466 (2025).
66. C, S. *et al.* Central Line Associated Bloodstream Infection Prediction Using Deep Attention Nets in the Healthcare Field. *InformingSciJ* 28, 013 (2025).
67. Borgonovo, F. *et al.* The CONSIDER study: Assessing the risk of catheter-associated bloodstream infections beyond the intensive care setting. *Am J Infect Control* 53, 950–956 (2025).
68. Wittekamp, B. H. *et al.* Catheter-related bloodstream infections: a prospective observational study of central venous and arterial catheters. *Scand J Infect Dis* 45, 738–745 (2013).

69. Haddadin, Y., Annamaraju, P. & Regunath, H. Central Line–Associated Blood Stream Infections. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).

70. Herzer, K. R., Niessen, L., Constenla, D. O., Ward, W. J. & Pronovost, P. J. Cost-effectiveness of a quality improvement programme to reduce central line-associated bloodstream infections in intensive care units in the USA. *BMJ Open* 4, e006065 (2014).

71. Martelin, A. *et al.* Cost-effectiveness of a new multi-lumen infusion device to reduce central-venous-line-associated bloodstream infections in neonates. *Journal of Hospital Infection* 152, 114–121 (2024).

72. Hollenbeak, C. S. The cost of catheter-related bloodstream infections: implications for the value of prevention. *J Infus Nurs* 34, 309–313 (2011).

73. Zhang, Y. *et al.* Incidence Rate, Pathogens and Economic Burden of Catheter-Related Bloodstream Infection: A Single-Center, Retrospective Case-Control Study. *Infect Drug Resist* 16, 3551–3560 (2023).

74. Christiaans, C. H. H., van Veen, F. E. E., Scheepe, J. R. & Blok, B. F. M. Patient satisfaction, quality of life, and catheter-related complications in long-term urinary catheter users: a nationwide survey. *World J Urol* 43, 470 (2025).

75. Krein, S. L. *et al.* Patient-reported complications related to peripherally inserted central catheters: a multicentre prospective cohort study. *BMJ Qual Saf* 28, 574–581 (2019).

76. Brown, R. & Burke, D. The hidden cost of catheter related blood stream infections in patients on parenteral nutrition. *Clin Nutr ESPEN* 36, 146–149 (2020).

77. Lafuente Cabrero, E. *et al.* Risk factors of catheter- associated bloodstream infection: Systematic review and meta-analysis. *PLoS ONE* 18, e0282290 (2023).

78. Marks, K. T. *et al.* Risk factors for central line-associated bloodstream infection in the pediatric intensive care setting despite standard prevention measures. *Infect Control Hosp Epidemiol* 1–9 (2024).
79. Haddad, A. *et al.* Risk factors for catheter-related bloodstream infections in a high-risk cancer patient population. *Infection Control & Hospital Epidemiology* 46, 692–695 (2025).
80. Cheng, S. *et al.* Risk Factors of Central Venous Catheter-Related Bloodstream Infection for Continuous Renal Replacement Therapy in Kidney Intensive Care Unit Patients. *Blood Purif* 48, 175–182 (2019).
81. Li, J. *et al.* The relationship between catheter-related bloodstream infection and multi-drug resistant bacteria: a five-year retrospective study. *BMC Infect Dis* 25, 988 (2025).
82. WHO. Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycaemia: Report of a WHO/IDF Consultation. (World Health Organization, Geneva, Switzerland, 2006).
83. KDIGO. Chapter 1: Definition and classification of CKD. *Kidney Int Suppl (2011)* 3, 19–62 (2013).
84. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 40, 373–383 (1987).
85. CDC - NHSN. Bloodstream Infections. (2025).
86. Nickel, B. *et al.* Infusion Therapy Standards of Practice, 9th Edition. *J Infus Nurs* 47, S1–S285 (2024).

87. Liu, F. T., Ting, K. M. & Zhou, Z.-H. Isolation Forest. in *2008 Eighth IEEE International Conference on Data Mining* 413–422 (2008). doi:10.1109/ICDM.2008.17.

88. Liu, F. T., Ting, K. M. & Zhou, Z.-H. Isolation-Based Anomaly Detection. *ACM Trans. Knowl. Discov. Data* 6, 3:1-3:39 (2012).

89. Aggarwal, C. C. *Outlier Analysis*. (Springer International Publishing, Cham, 2017). doi:10.1007/978-3-319-47578-3.

90. Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J. & Platt, J. Support Vector Method for Novelty Detection. in *Advances in Neural Information Processing Systems* vol. 12 (MIT Press, 1999).

91. Amer, M., Goldstein, M. & Abdennadher, S. Enhancing one-class support vector machines for unsupervised anomaly detection. in *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description* 8–15 (Association for Computing Machinery, New York, NY, USA, 2013).

92. Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J. & Platt, J. C. Support Vector Method for Novelty Detection.

93. Du, K.-L., Jiang, B., Lu, J., Hua, J. & Swamy, M. N. S. Exploring Kernel Machines and Support Vector Machines: Principles, Techniques, and Future Directions. *Mathematics* 12, 3935 (2024).

94. Fassio, F. *et al.* A Novel One-Class Classification Framework for Highly Imbalanced Binary Outcomes: the OC-Cat Approach. *Epidemiology, Biostatistics, and Public Health* https://doi.org/10.54103/2282-0930/29381 (2025).

95. Leoni, J., Fassio, F., Colaneri, M. & Breschi, V. OCCat: a novel framework for One-class Classification with Categorical data.

96. Maron, M. E. & Kuhns, J. L. On Relevance, Probabilistic Indexing and Information Retrieval. *J. ACM* 7, 216–244 (1960).

97. Bellman, R. On a routing problem. *Quart. Appl. Math.* 16, 87–90 (1958).

98. Hamming, R. W. Error Detecting and Error Correcting Codes. *Bell System Technical Journal* 29, 147–160 (1950).

99. Todeschini, R. *et al.* Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *J. Chem. Inf. Model.* 52, 2884–2901 (2012).

100. Wang, M., Wang, C., Yu, J. X. & Zhang, J. Community Detection in Social Networks: An In-depth Benchmarking Study with a Procedure-Oriented Framework.

101. Fortunato, S. Community detection in graphs. *Physics Reports* 486, 75–174 (2010).

102. Agelink van Rentergem, J. A., Bathelt, J. & Geurts, H. M. Clinical subtyping using community detection: Limited utility? *Int J Methods Psychiatr Res* 32, e1951 (2023).

103. Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Physics Reports* 659, 1–44 (2016).

104. Lancichinetti, A. & Fortunato, S. Community detection algorithms: A comparative analysis. *Phys. Rev. E* 80, 056117 (2009).

105. Huang, H., Chang, Q., Zhou, Y. & Liao, L. Risk factors of central catheter bloodstream infections in intensive care units: A systematic review and meta-analysis. *PLoS One* 19, e0296723 (2024).

106. Zeng, F., Li, W., Ye, H., Xie, C. & Zhong, H. A predictive model for catheter-related bloodstream infection in neonates with peripherally inserted central catheter. *Front. Med.* 12, 1665068 (2025).

107. Wang, F. *et al.* Machine learning risk prediction model for bloodstream infections related to totally implantable venous access ports in patients with cancer. *Asia-Pacific Journal of Oncology Nursing* 11, 100546 (2024).

108. Guo, Q. *et al.* Development and validation of a nomogram for predicting PICC catheter-related bloodstream infection among patients with hematologic malignancies. *BMC Infect Dis* 25, 1370 (2025).

109. Wu, S., Dai, F., Wen, Y., Luo, C. & Wu, C. Development and validation of a nomogram for predicting CRBSI in hemodialysis: a retrospective cohort study. *BMC Nephrol* 26, 255 (2025).

110. Guide to Preventing Catheter-Associated Bloodstream Infections in Adults (2025). *APIC* https://apic.org/implementation_guide/new-guide-to-preventing-catheter-associated-bloodstream-infections-in-adults-2025/.

111. Martin, K. *et al.* Clinical Outcomes and Risk Factors for Tunneled Hemodialysis Catheter-Related Bloodstream Infections. *Open Forum Infect Dis* 7, ofaa117 (2020).

112. Taieb, J. *et al.* Incidence of catheter-related infection in cancer patients receiving parenteral nutrition: A retrospective cohort study of French administrative claims data. *Medicine* 104, e42704 (2025).