Infect Dis Ther (2025) 14:1529–1545
https://doi.org/10.1007/s40121-025-01171-w

ORIGINAL RESEARCH

Deep Learning for the Early Diagnosis of Candidemia

Daniele Roberto Giacobbe · Sabrina Guastavino · Anna Razzetta · Cristina Marelli · Sara Mora · Chiara Russo · Giorgia Brucci · Alessandro Limongelli · Antonio Vena · Malgorzata Mikulska · Alessio Signori · Antonio Di Biagio · Anna Marchese · Ylenia Murgia · Marco Muccio · Nicola Rosso · Michele Piana · Mauro Giacomini · Cristina Campi · Matteo Bassetti

Received: April 28, 2025 / Accepted: May 23, 2025 / Published online: June 23, 2025

ABSTRACT

Introduction: Candidemia carries a heavy burden in terms of mortality, especially when presenting as septic shock, and its early diagnosis remains crucial.
Methods: We assessed the performance of a deep learning model for the early differential diagnosis between candidemia and bacteremia. The model was trained on a large dataset of automatically extracted laboratory features.
Results: A total of 12,483 episodes of candidemia (1275; 10%) or bacteremia (11,208; 90%) were included. For recognizing candidemia, a deep learning model showed sensitivity 0.80, specificity 0.59, positive predictive value (PPV) 0.18, weighted PPV (wPPV) 0.88, and negative predictive value (NPV) 0.96 on the training set (area under the curve [AUC] 0.69), and sensitivity 0.70, specificity 0.58, PPV 0.16, wPPV 0.87, and NPV 0.95 on the test set (AUC 0.64). Then, the learned discriminatory ability was tested in the subgroup of patients with available serum β-D-glucan (BDG) and procalcitonin (PCT) values to explore additive or

Daniele Roberto Giacobbe, Sabrina Guastavino, Cristina Campi, and Matteo Bassetti contributed equally to this manuscript.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40121-025-01171-w.

D. R. Giacobbe (✉) · C. Russo · G. Brucci · A. Limongelli · A. Vena · M. Mikulska · A. Di Biagio · M. Bassetti
Department of Health Sciences (DISSAL), University of Genoa, Via A. Pastore 1, 16132 Genoa, Italy
e-mail: danieleroberto.giacobbe@unige.it

D. R. Giacobbe · C. Russo · G. Brucci · A. Limongelli · A. Vena · M. Mikulska · A. Di Biagio · M. Muccio · M. Bassetti
Clinica Malattie Infettive, IRCCS Ospedale Policlinico San Martino, Genoa, Italy

S. Guastavino · M. Piana · C. Campi
Department of Mathematics (DIMA), University of Genoa, Genoa, Italy

A. Razzetta
School of Mathematics, University of Genoa, Genoa, Italy

C. Marelli
Oncostat, CESP, Inserm U1018, Université Paris-Saclay, Labeled Ligue Contre le Cancer, Gustave Roussy, Villejuif, France

C. Marelli
Institut Curie - INSERM U1331, Team statistics applied to personalized medicine, Paris, France

S. Mora · N. Rosso
UO Information and Communication Technologies, IRCCS Ospedale Policlinico San Martino, Genoa, Italy

A. Signori
Section of Biostatistics, Department of Health Sciences (DISSAL), University of Genoa, Genoa, Italy

synergistic effects with these more specific markers. Both feature selection and transfer learning did not improve the diagnostic performance of a model based on BDG and PCT only.

Conclusions: A deep learning model trained on nonspecific laboratory features showed some discriminatory ability to differentiate candidemia from bacteremia, highlighting the ability of deep learning to exploit complex patterns within nonspecific laboratory data. However, the learned patterns did not improve the diagnostic performance of more specific markers. Further exploration of candidemia prediction using laboratory features through machine learning techniques remains a promising area of research, serving as a valuable complement to the development of large-scale models that also incorporate clinical features.

Keywords: *Candida*; Machine learning; Artificial intelligence; Biomarker; Neural networks

Key Summary Points

Why carry out this study?
Candidemia has been associated with mortality in hospitalized patients, and its early diagnosis remains crucial.
Exploiting complex patterns within nonspecific laboratory data could be explored as a mean to improve the early recognition of candidemia.
What was learned from the study?
Deep learning models can exploit complex patterns within nonspecific laboratory data to help differentiate candidemia from bacteremia.
Further exploring early recognition of candidemia through machine learning techniques based on laboratory features remains interesting as a complement to the development of large-scale models also including clinical features.

A. Marchese
Department of Surgical Sciences and Integrated Diagnostics (DISC), University of Genoa, Genoa, Italy

A. Marchese
Microbiology Unit, IRCCS Ospedale Policlinico San Martino, Genoa, Italy

Y. Murgia · M. Giacomini
Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Genoa, Italy

M. Piana · C. Campi
Life Science Computational Laboratory (LISCOMP), IRCCS Ospedale Policlinico San Martino, Genoa, Italy

INTRODUCTION

Candidemia, i.e., bloodstream infection (BSI) caused by *Candida* spp., is among the most frequently encountered BSI in hospitalized patients, and has been associated with mortality rates up to > 50%, especially when presenting as septic shock [1–3].

In patients presenting with signs and symptoms consistent with bloodstream infection (BSI) and a severe clinical condition warranting empirical antimicrobial therapy while awaiting blood culture results, the decision to initiate empirical antifungal therapy is typically based on predictive clinical scores for candidemia and/or the presence of more or less specific laboratory markers [4–11]. In this context, the use of machine learning models has started to be investigated in the past few years in an attempt to improve accuracy in defining the risk of candidemia at the onset of consistent signs and symptoms, with the consequent aim of improving empirical decisions regarding early antifungal administration [12–18].

In the present study, we assessed the performance of a deep learning model for the early differential diagnosis between candidemia and bacteremia, trained on a large dataset of automatically extracted laboratory features.

METHODS

Setting and Previous Phases of the AUTO-CAND Project

The present retrospective study represents the third and final phase of the AUTO-CAND

project, and was conducted at IRCCS Ospedale Policlinico San Martino, a 1200-bed teaching hospital in northern Italy.

In the first phase of the AUTO-CAND project, a computerized extraction system of laboratory features from candidemia and bacteremia episodes was manually validated across a random subset of episodes to guarantee automated (i.e., computerized and non-manual) extraction accuracy > 99% [19]. The validation allowed the automated extraction, over a few days, of laboratory features from 15,752 episodes of candidemia and bacteremia (with 10% prevalence of candidemia, of which 2% mixed episodes of candidemia and bacteremia), i.e., all the episodes occurred at IRCCS Ospedale Policlinico San Martino from January 1, 2011 to December 31, 2019, a task that would have taken much longer if performed manually. The complete technical details of the automated extraction system and of its validation have been previously published [19, 20]. The first phase of the AUTO-CAND project served to provide a large dataset on which to explore the diagnostic performance of machine learning models for the diagnosis of candidemia based on laboratory features (the complete list of automatically extracted features is available as Supplementary Table S1).

This means that clinical features (e.g., presence of central venous catheter, total parenteral nutrition, previous broad-spectrum antibiotic therapy) were not automatically extracted. While, in principle, this can be expected to lead to lower diagnostic accuracy, it also cannot be known a priori whether training of machine learning models on a large amount (in terms both of features and of training examples) of laboratory data alone could provide useful complementary information to understand how to improve early diagnosis of candidemia through machine learning techniques.

In the second phase of the AUTO-CAND project, we assessed the diagnostic performance for the early diagnosis of candidemia (as a prediction task), based on the dataset of laboratory features extracted in the first phase, of three different supervised statistical/machine learning models: (i) penalized logistic regression with L1-norm; (ii) penalized logistic regression with L2-norm; (iii) random forest [13]. To this aim, we first considered that, whenever empirical treatment is required in clinical practice in patients with signs and symptoms consistent with BSI, early antibacterial therapy is usually administered, independent of the need for concomitant early antifungal therapy. This is because of the overall higher risk of bacteremia than candidemia, if only in terms of baseline prevalence. Consequently, we deemed the binary prediction task of clinical interest to be candidemia vs. no candidemia, with the former including both isolated candidemia and mixed candidemia/bacteremia, and the latter isolated bacteremia [13]. Second, we addressed the inherent and expected issue of missing values in real-world data. Owing to the large percentage of missing sequential data, we considered for each laboratory test only one result (i.e., the closest one to the onset of candidemia or bacteremia in the interval from within 2 days before to 12 h after the onset). Then, after excluding episodes and features with a proportion of missing values of more than 50% (in order to apply multiple imputations with sufficient reliability), the final dataset for the analysis was composed of 12,483 episodes (of which 1275 and 11,208 candidemia and bacteremia episodes, respectively) and 29 features (see Table 1 and Supplementary Table S2).

Notably, both serum beta-D-glucan (BDG) and serum procalcitonin (PCT), the combination of which is usually considered within our clinical reasoning for the differential diagnosis between candidemia and bacteremia [5]), were initially removed from the features considered for the analysis, owing to a high proportion of missing values (89% and 72%, respectively). The resulting dataset, not including serum BDG and serum PCT, was randomly split into a training set and a test set, including 70% and 30% of episodes, respectively, while retaining the original 10% prevalence of candidemia in both the training and test sets. Then, missing values were imputed through multiple imputation with chained equations, using a nearest neighbors method, and continuous features were standardized by subtraction of means and division by standard deviations [21]. Considering that we aimed to improve the baseline diagnostic performance

Table 1 Causative agents of candidemia and bacteremia in the study population

Causative agents	No./total (%)
Candidemia[a]	1275/12483 (10)
Candida albicans	524/1275 (41)
Candida parapsilosis	454/1275 (36)
Candida glabrata (*Nakaseomyces glabrata*)	106/1275 (8)
Candida tropicalis	57/1275 (4)
Candida krusei (*Pichia kudriavzevii*)	12/1275 (1)
Other	122/1275 (10)
Bacteremia[b]	11,208/12483 (90)
Escherichia spp.	2114/11208 (19)
Staphylococcus epidermidis	1484/11208 (13)
Staphylococcus aureus	1455/11208 (13)
Enterococcus spp.	1252/11208 (11)
Klebsiella spp.	1004/11208 (9)
Streptococcus spp.	722/11208 (6)
Pseudomonas spp.	595/11208 (5)
Staphylococcus hominis	385/11208 (3)
Enterobacter spp.	319/11208 (3)
Staphylococcus haemolyticus	261/11208 (2)
Proteus spp.	249/11208 (2)
Acinetobacter spp.	116/11208 (1)
Serratia spp.	111/11208 (1)
Bacteroides spp.	108/11208 (1)
Staphylococcus capitis	91/11208 (1)
Stenotrophomonas spp.	70/11208 (1)
Corynebacterium spp.	61/11208 (1)
Other	811/11208 (7)

Table 1 continued

[a]Of which, mixed candidemia/bacteremia episodes were 240/1275 (19%), all categorized as candidemia for study analyses in line with the purpose of the study (see "Methods" section). Some episodes caused by more than one *Candida* species occurred and were correctly identified and classified as candidemia by the automated extraction systems (as per manual validation [19]) for the study analyses, although for summary descriptive representation (as in the present table) the current version of the system only allowed to retrieve the first identified species

[b]Only bacteremia, no mixed candidemia/bacteremia episodes. Some episodes caused by more than one bacterial genus occurred and were correctly identified and classified as bacteremia by the automated extraction systems (as per manual validation [19]) for the study analyses, although for summary descriptive representation (as in the present table) the current version of the system only allowed to retrieve the first identified genus

for early candidemia compared to that of laboratory markers usually employed in our center (serum BDG and serum PCT, as discussed above), we decided to set the sensitivity of serum BDG previously measured in our center (around 60% based on internal data) as the threshold of minimum sensitivity to be required for the trained machine learning models. In addition, we set the rule for sensitivity to be higher than specificity. Indeed, we ranked the potential consequences of lower sensitivity, i.e., lack of early antifungal treatment in true cases of candidemia, as more clinically relevant than those of lower specificity, i.e., early antifungal treatment in false cases of candidemia (indeed, in such a case early antifungal treatment could then be discontinued at the time of blood culture results, while continuing antibacterial therapy). Consequently, for the training of machine learning models, we deemed any chosen τ value (defined as the threshold to assign an episode to 0 or 1 according to the probability predicted by the evaluated classifier) to fulfill the following condition in the training set: sensitivity $\geq$ specificity ≥ 0.60 [13].

Overall, the random forest model achieved the best diagnostic performance for candidemia, with 98% sensitivity and 65% specificity in the training set (with $\tau = 0.1$, based on a user-defined score function built to maximize the true skill

statistic [TSS], also known as Youden Index) and 74% sensitivity and 57% specificity in the test set (for more details on the complete calculation of the model diagnostic performance see [13]). Then, we exploited a permutation feature importance approach to identify the most influential features in predicting candidemia. This resulted in 12 selected features (eosinophil count, platelet count, neutrophil cell count, hematocrit, uric acid, monocyte cell count, hemoglobin, urea, albumin, lymphocyte cell count, white cell count, and prothrombin time) that were reemployed, together with serum BDG and PCT, to train a new random forest classifier for predicting candidemia among the 1165 episodes (of which 177 were candidemia, 15%) with available serum BDG and PCT results. This allowed us to leverage what was learned on the much larger entire dataset of 12,483 episodes without serum BDG and PCT results for the subsequent training of a smaller dataset in which it was possible to consider these two markers. Eventually, the diagnostic performance for candidemia of the model including the twelve selected features plus serum BDG and PCT was apparently numerically better than that of a model including only serum BDG and PCT for the majority of the evaluated diagnostic performance measures (e.g., in terms of TSS, sensitivity, weighted positive predictive value [wPPV] and negative predictive value [NPV]), although the differences were small and unlikely to be clinically significant [13].

Interpretation of Phase 2 Results and Rationale for Phase 3

Although small, the numerical trend towards better diagnostic performance when the model was trained on serum BDG and PCT plus the selected less specific 12 features vs. training on serum BDG and PCT alone could merit further investigation. In particular, should a true improvement exist, two non-mutually exclusive possibilities may deserve to be explored: (*i*) whether employing an alternative machine learning-based predictive model able to catch more complex interactions between the above less specific laboratory values could further improve the diagnostic performance of serum BDG and PCT plus nonspecific laboratory features over serum BDG and PCT alone; (*ii*) whether adding also clinical predictors of candidemia (e.g., previous broad-spectrum antibiotic therapy, presence of central venous catheter, total parenteral nutrition) to large datasets of laboratory features could further improve the early diagnosis of candidemia through ML techniques.

In the third phase of the AUTO-CAND project, we specifically explored the first of these two possibilities by training a deep learning model on the same dataset of laboratory features used in the second phase. The AUTO-CAND project was approved by the pertinent local ethics committee (Liguria Region Ethics Committee, registry number 71/2020). The requirement for informed consent for this study was waived due to the retrospective nature of the analyses.

Phase 3 Methods

Predictive Model Characteristics

We implemented a feed-forward neural network (NN) with one input layer, three hidden layers, and one output layer (Fig. 1): the first hidden layer contained 100 neurons, and the number of neurons was halved in each subsequent hidden layer. The activation functions were rectified linear unit (ReLU) for all layers, except for the last one, where the sigmoid function was used to provide the output probability of the event of interest (candidemia) being positive. For the training phase, we used a loss function called score-oriented loss (SOL) function that optimizes the TSS, and is particularly suitable for evaluating performance on imbalanced datasets [22]. We used Adam optimization [23], with the learning rate set to 10^{-5} and a batch size of 256. We applied early stopping with a patience of 1000 epochs by monitoring the validation loss in order to stop the training process for preventing overfitting. The hyperparameters of the NN were set by an

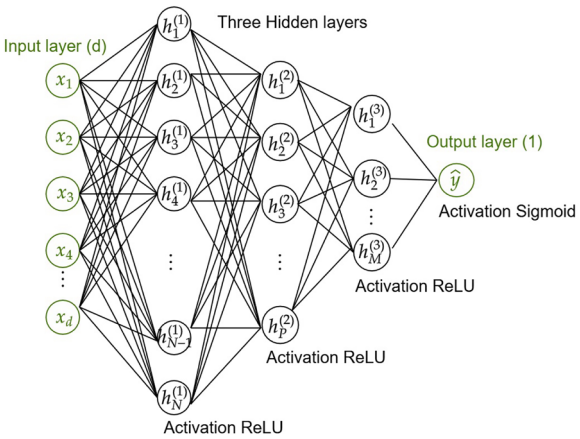

Fig. 1 Architecture of the feed-forward neural network trained for the early prediction of candidemia. The neural network architecture consists of the following components: the input layer contains d-neurons (corresponding to the number of features), the three hidden layers have 100, 50, and 25 neurons, respectively, and the output layer has a single neuron. ReLU, rectified linear unit

empirical trial-and-error optimization process on several experiments.

Dataset Preparation

We employed the same dataset of phase 2 [13]. As reported above, the dataset consisted of 12,483 observations, with a candidemia prevalence of 10% (1275/12,483), and 29 features (see Table 1 and Supplementary Table S2). We considered 16 different splits of this dataset in training-validation-test sets with the following percentage: we randomly assigned the 30% of the dataset to the test set, while the remaining 70% of the observations were divided, again randomly, between the training (85%) and validation (15%) sets. The only criterion that the assignment into different sets had to respect was the maintenance of the balance between prevalence of candidemia (10%) and bacteremia (90%). The rationale for splitting the dataset was to assess the robustness of the predictive model among different configurations of training, validation, and test sets.

Model Performance Evaluation

The prediction task was evaluated by means of several performance measures, computed from the confusion matrices, and obtained varying the threshold that assigns the model result to the classes 0 or 1. The performance measures we employed were: sensitivity, specificity, TSS, PPV, wPPV, NPV, and area under the curve (AUC). For each split we checked if in the validation set a clinical condition of sensitivity ≥ specificity ≥ 0.60 was satisfied for a certain value of the threshold. If so, the threshold value was used in the corresponding test set, otherwise we employed the threshold that maximized the TSS in the validation set. The wPPV in all the analyses was the PPV computed in the weighted modality (i.e., the score was calculated as a weighted average of the precision scores of each class, where the weights depend on the number of samples of each class) in order to ensure that the precision of each class was fairly represented based on its frequency, preventing dominance of numerically larger classes in the overall performance metric. This provided stability for technical comparison between models. Indeed, even with slight changes in the prevalence of candidemia (e.g., in subgroup analysis, see below) large changes in classical PPV would have occurred (differently from NPV, which is more stable to slight variations in presence of an already low prevalence of candidemia). However, since wPPV does not allow for an intuitive clinical interpretation (it does not answer the question "what is the probability of a true positive if positive?"), also classical PPV was calculated and reported in text and tables alongside wPPV.

Model Performance Evaluation: Feature Selection

In order to identify the key features in predicting candidemia, we explored two advanced techniques for feature selection that build upon the widely recognized permutation feature importance (PFI) algorithm, and we used the SHapley Additive exPlanations (SHAP) algorithm for model explainability [24, 25].

PFI is a model-agnostic technique used to evaluate the importance of individual features

in a trained AI model and the method works by randomly shuffling the values of each feature and measuring the decrease in the model's performance. The greater the decrease in performance, the more important that feature is. Nevertheless, the PFI algorithm does not take into account the correlation between features. Therefore, in the present work, we considered two advanced PFI-type methods, which include the absolute correlation coefficients between features in the permutation process. The first is cross-validated permutation feature importance (CVPFI) [24]. CVPFI is an extension of PFI that accounts for the correlation between features and incorporates cross-validation to improve the robustness of the feature importance scores. Cross-validation splits the dataset into multiple folds, training and testing the model on different subsets of the data, which helps to mitigate overfitting and gives a more generalized estimate of feature importance. The second is correlation-driven permutation feature importance (CDPFI) [26]. CDPFI is a variant of CVPFI that accounts for the correlation between features. It adjusts for cases where features are highly correlated with each other, which can lead to misleading feature importance rankings if not properly accounted for.

For the two PFI-type approaches, we selected features with an importance score, based on the TSS, greater than a fixed threshold of 0.1. This means that a feature is considered relevant if it contributes to a change in the score of at least 0.1. These quantities were computed for all the splits and then aggregated to form a ranking of the features.

As introduced above, we also applied the SHAP algorithm to provide explanations for why the deep learning model made a specific prediction (positive or negative) for particular samples. SHAP is a machine learning explainability method based on Shapley values, which come from cooperative game theory and were originally designed to fairly distribute the payout (or value) among players in a game. In our context, each "player" is a feature, and the "payout" is the prediction made by the deep learning model. We also provided a ranking of the features based on the results from the SHAP algorithm, assigning an average position ranking to each feature derived from the SHAP analysis. While this did not allow to select features based on SHAP analysis, the average position ranking provided additional information to check whether features selected by means of CVPFI and CDPFI also had a high average position ranking by SHAP analysis. Eventually, we retained only those features that were selected by both CVPFI and CDPFI methods.

Subgroup Analyses in Patients with Available Serum BDG and PCT Values

The subgroup analyses were conducted only in the subgroup of episodes of candidemia and bacteremia for which serum BDG and PCT values were available ($n = 1165$), leveraging what was learned in the previous steps on the entire dataset of 12,483 episodes regarding the predictive ability of other nonspecific laboratory features. Of note, as above we started by considering 16 different splits of the entire dataset, which included all 12,483 candidemia and bacteremia episodes. However, in this case, the splitting strategy, in addition to maintaining the balance between candidemia and bacteremia (in terms of their relative prevalence), was designed to ensure that the balance was also preserved in the respective subgroups (employed in subgroup analyses) of episodes for which serum BDG and serum PCT values were available.

For subgroup analyses, we trained, validated, and tested the NN model in three different ways: (i) by considering only serum BDG and serum PCT; (ii) by considering the features selected by CVPFI and CDPFI plus serum BDG and serum PCT; (iii) by exploiting transfer learning, i.e., by including and freezing the weights of the dataset previously trained on the entire dataset (selected features-based model) in the architecture of the novel NN model trained on the subgroup of episodes with serum BDG and PCT values available. Overall, this latter model combined the pre-trained network with two additional features, serum BDG and serum PCT, and included two fully connected hidden layers with 24 and 12 neurons, respectively, to process the combined inputs from both the pre-trained model and the added features. The final output layer consisted of a single neuron with a sigmoid activation

function to generate a probability prediction. The novel model was trained in a similar way to the pre-trained one, using the same optimizer, the SOL function tailored to optimize the TSS, and the early-stopping strategy to prevent overfitting. While the weights of the pre-trained model were not updated, the new layers added after it were trained from scratch. Finally, to adapt the model to the new input data, the first layers of the new model were retrained, a process commonly referred to as fine-tuning, which is particularly useful when limited data are available.

RESULTS

The distribution of the causative agents of the 12,483 episodes of candidemia (10%) and bacteremia (90%) is shown in Table 1. The majority of candidemia episodes were caused by *Candida albicans* (524/1275; 41%), while the majority of bacteremia episodes were caused by *Escherichia coli* (2114/11,208; 19%).

The first step of our analysis was to assess the diagnostic performance for candidemia of the deep learning model trained after the split of the entire dataset of 12,483 episodes in training, validation, and test sets. In 16 out of the 16 validation sets there was a threshold value such that the clinical condition of sensitivity ≥ specificity ≥ 0.60 was satisfied. The mean ± standard deviation (SD) values of the performance metrics in the validation phase are shown in Supplementary Table S3, whereas the performance metrics of the NN model in the test set (for the different threshold identified in the validation phase that satisfied the clinical condition) are shown in Supplementary Table S4. As shown in Table 2, the mean performance metrics (± SD) of the deep learning model for the diagnosis of candidemia in the training set were sensitivity 0.80 (± 0.03), specificity 0.59 (± 0.06), PPV 0.18 (± 0.02), wPPV 0.88 (± 0.01); NPV 0.96 (± 0.00), and AUC 0.69 (± 0.02), while in the test set they were as follows: sensitivity 0.70 (± 0.05); specificity 0.58 (± 0.06); PPV 0.16 (± 0.01); wPPV 0.87 (± 0.00); NPV 0.95 (± 0.01); AUC 0.64 (± 0.01).

The second step of the analysis was to identify and select those features contributing mostly to the predictive ability of the trained NN model. In this regard, the features selected by means of CVPFI and CDPFI are reported in Supplementary Tables S5 and S6, respectively. More in detail, the sum of the weights on the 16 splits is reported in the second column of Supplementary Table S5 and S6, whereas the number of splits in which

Table 2 Mean performance metrics (± SD) for the early diagnosis of candidemia of the neural network model in training, validation and test sets from the full dataset (12,483 episodes and 29 features)

Sensitivity	Specificity	PPV	wPPV	NPV	TSS	AUC
Training set						
0.80 (± 0.03)	0.59 (± 0.06)	0.18 (± 0.02)	0.88 (± 0.01)	0.96 (± 0.00)	0.38 (± 0.05)	0.69 (± 0.02)
Validation set						
0.72 (± 0.05)	0.57 (± 0.06)	0.16 (± 0.01)	0.87 (± 0.01)	0.95 (± 0.01)	0.29 (± 0.04)	0.65 (± 0.02)
Test set						
0.70 (± 0.05)	0.58 (± 0.06)	0.16 (± 0.01)	0.87 (± 0.00)	0.95 (± 0.01)	0.29 (± 0.03)	0.64 (± 0.01)

AUC area under the curve, *SD* standard deviation, *NPV* negative predictive value, *PPV* positive predictive value, *TSS* true skill statistic, *wPPV* weighted PPV

a given feature was selected is reported in the third column. The average rankings on the 16 splits obtained by means of the SHAP method are reported in Supplementary Table S7, and three graphical examples showing the contribution of features to the prediction of candidemia by SHAP method in different splits are available as Supplementary Figure S1. We eventually retained seven features (platelet count, urea, uric acid, hematocrit, hemoglobin, red cell count, neutrophil cells count), i.e., those selected by both CVPFI and CDPFI (Table 3). The performance metrics obtained in the training, validation, and test sets from the entire cohort of the seven-feature model (including the seven features selected above) and the 12-feature model (including the 12 features previously selected in the phase 2 of the project) are reported in Table 4. As shown in the table, the mean performance metrics of the two feature-reduced models (seven features-based and 12 features-based), with the exception of a slight reduction in sensitivity, were very similar to those obtained exploiting all the 29 features included in the initial model.

The final step was to conduct a subgroup analysis including only the 1165 episodes of candidemia and bacteremia for which serum BDG and serum PCT values were available (in this subgroup the prevalence of candidemia was 15%, for details see Supplementary Table S8), leveraging what was learned in the previous steps on all 12,483 episodes with regard to the predictive ability of laboratory features other than BDG and PCT. The performance metrics obtained in the training, validation, and test sets of models (i), (ii), and (iii) are reported in Table 5. As shown in the table, the mean performance metrics of the two models trained by exploiting either feature selection or transfer learning did not show improvement compared with a model trained only on serum BDG and serum PCT values.

Table 3 Summary of feature ranking though the three different feature selection methods

Feature	CVPFI	CDPFI	SHAP
Platelet count	x	x	0.2 (± 0.3)
Urea	x	x	2.4 (± 1.4)
Uric acid	x	x	4.8 (± 3.3)
Hematocrit	x	x	6.6 (± 3.7)
Hemoglobin	x	x	8.2 (± 5.4)
Red cells count	x	x	12.9 (± 4.2)
Neutrophil cells count	x	x	13.1 (± 4.6)
Age	x		8.8 (± 2.9)
Albumin	x		9.3 (± 3.4)
Creatinine	x		10.0 (± 2.3)
PT	x		10.7 (± 4.4)
APTT	x		12.0 (± 3.7)
White cells count	x		13.4 (± 4.4)
Total proteins	x		14.9 (± 2.2)
INR	x		15.9 (± 1.5)
Previous *Candida* colonization	x		3.4 (± 1.1)
C-reactive protein			5.9 (± 1.9)
Eosinophil cells count			10.3 (± 4.1)

DISCUSSION

A deep learning model trained on nonspecific laboratory features showed some discriminatory ability to differentiate candidemia from

Table 3 continued

Feature	CVPFI	CDPFI	SHAP
Gamma-glutamyl transferase			14.5 (± 2.9)
Alkaline phosphatase			15.3 (± 2.7)
Glucose			15.3 (± 2.6)
Basophil cells count			15.7 (± 1.7)
Total bilirubin			16.2 (± 1.5)
Sex			16.2 (± 2.9)
Alanine aminotransferase			16.2 (± 1.4)
Lactate dehydrogenase			16.4 (± 2.3)
Monocyte cells count			16.4 (± 1.6)
Lymphocyte cells count			16.6 (± 1.3)
Aspartate aminotransferase			17.5 (± 1.3)

Retained features were those selected both by CVPFI and by CDPFI (first and second columns, respectively). Ranking of all the initial 29 features by SHAP, expressed as average ranking (± SD), is reported in the third column, showing a trend over higher ranking for features selected by CVPFI and CDPFI. *APTT* activated partial thromboplastin time, *CDPFI* correlation-driven permutation feature importance, *CVPFI* cross-validated permutation feature importance, *INR* international normalized ratio, *PT* prothrombin time, *SD* standard deviation, *SHAP* SHapley Additive exPlanations

bacteremia (sensitivity 0.70, specificity 0.58, PPV 0.16, wPPV 0.87, NPV 0.95, and AUC 0.64 in the test set), highlighting the ability of deep learning to exploit complex patterns within nonspecific laboratory data. However, when attempting to combine the learned predictive ability of nonspecific laboratory features with that of serum BDG and serum PCT (using either feature selection or transfer learning techniques), the diagnostic performance remained similar to that of serum BDG and serum PCT alone, showing no clear additive or synergistic effect.

Previous studies have attempted to predict candidemia through the use of machine learning techniques, although usually based on smaller datasets than our own and including a variable amount of clinical features besides laboratory values. In a retrospective, multicenter study of 433 patients (295 and 138 with candidemia and bacteremia, respectively), a random forest algorithm showed 84% sensitivity and 91% specificity for the differential diagnosis between candidemia and bacteremia [15]. In another retrospective, multicenter study of 7932 patients (of which 137 with candidemia), an XGBoost algorithm showed 84% sensitivity and 89% specificity for the diagnosis of candidemia vs. non-candidemia (either negative blood cultures or blood cultures positive for other pathogens) [17]. In a similar retrospective study of 501 patients with candidemia and 2000 patients with either negative blood cultures or bacteremia, a random forest algorithm showed 89% sensitivity and 90% specificity for diagnosing the former [16]. Sensitivity and specificity of 80% and 77%, respectively, were registered in an internal validation cohort of 100 patients when evaluating a random forest algorithm trained to differentiate cases of candidemia from controls with bacteremia (ratio 1:1) in a training set of 237 patients [14]. The same algorithm showed 78% sensitivity and 75% specificity in an external validation cohort of 77 patients [14]. Finally, a mean sensitivity and mean specificity of 72% and 80%, respectively, were registered when employing an XGBoost algorithm for the diagnosis of candidemia in a multicenter, retrospective study of 8002 patients with a candidemia prevalence of 0.62% [18]. For all these studies, we reported only sensitivity and specificity because they are less susceptible than other performance metrics (e.g., NPV and PPV) to changes in candidemia prevalence, which varied significantly across these studies due to their different design (low extrapolability of PPV and NPV to settings with

Table 4 Mean performance metrics (± SD) for the early diagnosis of candidemia of the neural network model in training, validation and test sets from the feature-reduced full datasets (12,483 episodes and seven or 12 features)[a]

Model	Sensitivity	Specificity	PPV	wPPV	NPV	TSS	AUC
Training set							
7 features	0.71 (± 0.06)	0.60 (± 0.04)	0.17 (± 0.01)	0.87 (± 0.01)	0.95 (± 0.01)	0.31 (± 0.05)	0.66 (± 0.02)
12 features	0.73 (± 0.03)	0.61 (± 0.05)	0.18 (± 0.02)	0.87 (± 0.01)	0.95 (± 0.01)	0.35 (± 0.05)	0.67 (± 0.03)
Validation set							
7 features	0.67 (± 0.05)	0.59 (± 0.05)	0.16 (± 0.01)	0.86 (± 0.01)	0.94 (± 0.01)	0.26 (± 0.04)	0.63 (± 0.02)
12 features	0.66 (± 0.06)	0.61 (± 0.05)	0.16 (± 0.01)	0.86 (± 0.01)	0.94 (± 0.01)	0.27 (± 0.05)	0.64 (± 0.02)
Test set							
7 features	0.62 (± 0.05)	0.59 (± 0.04)	0.15 (± 0.01)	0.85 (± 0.00)	0.93 (± 0.05)	0.22 (± 0.02)	0.61 (± 0.01)
12 features	0.62 (± 0.06)	0.61 (± 0.05)	0.15 (± 0.01)	0.85 (± 0.01)	0.93 (± 0.01)	0.22 (± 0.03)	0.61 (± 0.01)
29 features	0.70 (± 0.05)	0.58 (± 0.06)	0.16 (± 0.01)	0.87 (± 0.00)	0.95 (± 0.01)	0.29 (± 0.03)	0.64 (± 0.01)

AUC area under the curve, *SD* standard deviation, *NPV* negative predictive value, *PPV* positive predictive value, *TSS* true skill statistic, *wPPV* weighted PPV

[a]For descriptive comparison, the last line of the table shows the mean performance metrics in the test set for the original models trained on 29 features, already reported in Table 2

different prevalence of candidemia also affects the present study, especially for PPV, for the reasons reported in methods). We think focusing on sensitivity and specificity also helps to highlight more clearly an apparent lack of substantial improvement in performance when increasing sample size, contrary to what would be intuitively expected. In our opinion, this may be due—at least in part—to the fact that, while the number of laboratory features is generally similar, or even increasing, from small to large studies, the same is not true for clinical features such as granular medical and pharmacological history, acute phase conditions, and the presence and type of invasive devices, the number of which in most cases tends to decrease from small to large samples [14–18]. This consideration may rely on the fact that it is still very frequently unfeasible to extract automatically (time-consuming to collect manually) clinical features from unstructured data (e.g., daily notes in medical records), while accurate automated extraction of structured data such as laboratory values has already become possible in many research settings [27–30]. Overall, we think this may explain the lack of gain—and frequently the apparent decrease—of performance of models from small to large studies [13].

On the other hand, assessing how ML models perform on large datasets of laboratory values is an unprecedented opportunity to directly explore whether complex interactions between nonspecific laboratory values, difficult or even impossible to be recognized by clinicians, could improve the diagnostic performance of

Table 5 Mean performance metrics (± SD) for the early diagnosis of candidemia of the neural network model in training, validation and test sets from the subgroup of patients with available serum BDG and PCT values (1165 episodes)

Model	Sensitivity	Specificity	PPV	wPPV	NPV	TSS	AUC
Training set							
BDG plus PCT	0.66 (± 0.06)	0.70 (± 0.05)	0.29 (± 0.03)	0.82 (± 0.01)	0.92 (± 0.01)	0.36 (± 0.03)	0.68 (± 0.01)
BDG plus PCT plus 7 features	0.76 (± 0.07)	0.73 (± 0.07)	0.35 (± 0.05)	0.86 (± 0.01)	0.95 (± 0.01)	0.50 (± 0.07)	0.75 (± 0.04)
BDG plus PCT plus 7 features (TL)	0.68 (± 0.07)	0.68 (± 0.06)	0.28 (± 0.03)	0.83 (± 0.01)	0.92 (± 0.01)	0.36 (± 0.06)	0.68 (± 0.03)
Validation set							
BDG plus PCT	0.68 (± 0.10)	0.70 (± 0.06)	0.30 (± 0.05)	0.83 (± 0.03)	0.92 (± 0.02)	0.38 (± 0.10)	0.69 (± 0.06)
BDG plus PCT plus 7 features	0.66 (± 0.10)	0.73 (± 0.07)	0.31 (± 0.05)	0.83 (± 0.02)	0.92 (± 0.02)	0.38 (± 0.10)	0.69 (± 0.05)
BDG plus PCT plus 7 features (TL)	0.71 (± 0.12)	0.68 (± 0.05)	0.29 (± 0.04)	0.83 (± 0.03)	0.93 (± 0.03)	0.39 (± 0.11)	0.70 (± 0.05)
Test set							
BDG plus PCT	0.64 (± 0.07)	0.71 (± 0.05)	0.29 (± 0.02)	0.82 (± 0.01)	0.92 (± 0.01)	0.35 (± 0.05)	0.68 (± 0.03)
BDG plus PCT plus 7 features	0.59 (± 0.08)	0.72 (± 0.08)	0.28 (± 0.05)	0.81 (± 0.01)	0.91 ± 0.01	0.31 (± 0.06)	0.66 (± 0.03)
BDG plus PCT plus 7 features (TL)	0.65 (± 0.07)	0.68 (± 0.06)	0.27 (± 0.03)	0.82 (± 0.01)	0.92 ± 0.01	0.33 (± 0.05)	0.67 (± 0.02)

AUC area under the curve, *BDG* β-D-glucan, *SD* standard deviation, *NPV* negative predictive value, *PCT* procalcitonin, *PPV* positive predictive value, *TL* transfer learning, *TSS* true skill statistic, *wPPV* weighted PPV

more specific diagnostic markers (those usually employed by clinicians) in either an additive or a synergistic way. We think this is worth exploring also from an explainability perspective. Indeed, despite using black-box models (random forest and deep learning in phase 2 and phase 3 of the AUTO-CAND project, respectively) that are hampered by the inability to precisely recognize how a model arrived at a given prediction [31, 32], the fact that the input data was made up of only laboratory values means that the prediction (through model calculations) was based on their interactions only. This may reduce the risk of unrecognized biases related to complex interactions between more sophisticated clinical features. Although blindness to clinical features may also exert some confounding influence on the predictive ability of input data, the different nature of such biases means that models based only on laboratory markers could offer a complementary perspective to clinical features-based models (see below).

Notably, all of this is independent of the use of any possible explanation technique to identify which features eventually influenced predictions by black box models (in the present study, we employed the SHAP algorithm, which, albeit useful, like other explanation techniques is still far from perfect [33]). We also think it would be of little use for clinicians to know which were the single nonspecific markers that contributed to a given prediction, since we were mostly dealing with markers that are nonspecific *per definition*. Rather, it is more likely their complex interactions that contributed to predictions, in a way that is expected to be unexplainable or only partly explainable to clinicians *a priori*. In line with this view, our use of explanations was not strictly aimed to explain predictions. Rather, it was also aimed to further support the process of feature selection, i.e., the reduction of the number of required features without substantial losses in model performance, in line with any possible future usability of similar models in clinical practice (where only a few laboratory tests are usually and routinely performed).

In the phase 3 of the AUTO-CAND project, the aim of applying deep learning was to explore whether this technique could be able to exploit complex interactions between nonspecific laboratory features to improve the diagnostic performance for candidemia, either in general or in addition to more specific markers like serum BDG and serum PCT. In phase 2, another machine learning model (random forest) appeared to show a numerical trend in improving the diagnostic performance of serum BDG and PCT alone by exploiting additional nonspecific markers, although only minimally and not in a clinically significant way [13]. While also our deep learning model showed an higher than expected ability, for nonspecific markers, to discriminate between candidemia and bacteremia (sensitivity and specificity in the test set are very similar to those registered for the random forest model in phase 2, and higher than those of logistic regression models [13]), it did not improve the performance of a deep learning model based on serum BDG and PCT alone, thereby currently not supporting this strategy. Further investigation may allow to establish whether expanding the dataset, varying hyperparameters, and/or using trend data of markers variation in the days preceding candidemia or bacteremia could enhance the diagnostic performance of the model in a clinically meaningful way, capable of improving classical diagnostic approaches. In addition, as anticipated above, continuing to investigate laboratory-based only prediction of candidemia could allow to retain some kind of explanations of model predictions *a priori* (if only laboratory markers are fed into the model, predictions inherently arise from interactions between markers only). On the other hand, it is of great interest also to explore whether adding clinical features could improve model predictions, adapting to very large and automatically extracted datasets the conceptual principles of classical predictive models for candidemia based on clinical features [6–9, 34–36]. In this context, progress is being made in the automatic extraction of clinical variables from unstructured data in medical records, which could help to automatically extract and build also very large datasets of granular clinical features to support the development of more accurate machine learning-based predictive models of candidemia [37, 38]. Overall, we think these two approaches are not mutually exclusive. Indeed, continuing to explore the predictive ability of large laboratory markers-only models alongside large clinical models may prove synergistic in improving our understanding of possible biases, nuances, and potential advantages of applying machine learning techniques to the early diagnosis of candidemia.

The present study is not exempt from limitations. In addition to the intrinsic limitations related to the retrospective nature of the analysis, it should be acknowledged that there was also insufficient data to assess the possible impact of the evolution of the markers over time in the prediction of candidemia. However, the lack of this information reflects the absence of daily laboratory tests in clinical practice (as daily tests may not be justified clinically), and thus represents a limitation that all models based on laboratory data should take into account. A second notable limitation is that the laboratory-based nature of the study did not allow us to stratify the study population according to different baseline risks of candidemia development [7,

8]. Third, the lack of clinical features could technically be considered a limitation, although we think it also represents a strength from the point of view of explainability and bias detection as a complement to clinical features-based models, for the reasons detailed above. Fourth, while we internally validated our findings, external validation in other centers remains warranted. Finally, it is of note that we did not compare the performance metrics of the different models (a binomial regression accounting for repeated measures, considering the presence of splits, would have been appropriate) to provide p values and confidence intervals. This choice was based on the following: (i) the consideration that in a large sample even slight changes in metrics would possibly reach statistical significance, with consequent misleading interpretations of clinically irrelevant differences (either in favor or against a given model); (ii) the preference toward showing the presence of small SD, testifying to the consistency of metrics across splits, rather than confidence intervals (already expected to be small in large samples).

CONCLUSIONS

A deep learning model trained on nonspecific laboratory features showed some discriminatory ability to differentiate candidemia from bacteremia, highlighting the ability of deep learning to exploit complex patterns within nonspecific laboratory data. However, it did not show additive or synergistic effects over the prediction of candidemia vs. bacteremia based on serum BDG and serum PCT only. Further exploring the prediction of candidemia by laboratory markers through deep learning or other machine learning models remains interesting in terms of explainability and understanding of potential biases, not as an alternative but as a complement to the development of large-scale models including clinical features.

ACKNOWLEDGEMENTS

S.G. acknowledges the support of the "Hub Life Science—Digital Health (LSH-DH) PNC-E3-2022-23683267—Progetto DHEAL-COM—CUP: D33C2200198000", granted by the Italian Ministero della Salute within the framework of the Piano Nazionale Complementare to the "PNRR Ecosistema Innovativo della Salute—Codice univoco investimento: PNC-E.3".

Author Contributions. Conceptualization: Daniele Roberto Giacobbe, Sabrina Guastavino; methodology: Daniele Roberto Giacobbe, Sabrina Guastavino, Cristina Campi, Cristina Marelli, Marco Muccio, Alessio Signori, Michele Piana, Mauro Giacomini, Ylenia Murgia, Sara Mora, Nicola Rosso, Anna Razzetta; formal analysis and investigation: Daniele Roberto Giacobbe, Sabrina Guastavino, Cristina Marelli, Marco Muccio, Anna Razzetta; data collection: Chiara Russo, Giorgia Brucci, Alessandro Limongelli, Cristina Marelli, Sara Mora; writing—original draft preparation: Daniele Roberto Giacobbe, Sabrina Guastavino, Cristina Campi; writing—review and editing: Daniele Roberto Giacobbe, Sabrina Guastavino, Cristina Campi, Cristina Marelli, Marco Muccio, Alessio Signori, Michele Piana, Mauro Giacomini, Ylenia Murgia, Sara Mora, Nicola Rosso, Anna Razzetta, Chiara Russo, Giorgia Brucci, Alessandro Limongelli, Antonio Vena, Malgorzata Mikulska, Antonio Di Biagio, Anna Marchese, Matteo Bassetti; supervision: Daniele Roberto Giacobbe, Antonio Vena, Malgorzata Mikulska, Anna Marchese, nicola Rosso, Michele Piana, Mauro Giacomini, Cristina Campi, Matteo Bassetti. All authors have read and approved the final version of the manuscript.

Funding. The AUTO-CAND project was supported by Pfizer Global Medical Grants (GMG) for General Research [Project Tracking Number 69511763]. The funder had no role in the study design, data collection and analysis, decision to publish, and preparation of the manuscript. The journal's publication fee was covered by research funds of the main authors.

Data Availability. The data presented in this study will be available from the authors on reasonable request and provided all regulatory and privacy requirements are fulfilled.

Declarations

Conflicts of Interest. Matteo Bassetti and Malgorzata Mikulska are Editorial Board members of Infectious Diseases and Therapy, and Daniele Roberto Giacobbe is an Advisory Board member of Infectious Diseases and Therapy. Matteo Bassetti, Malgorzata Mikulska, and Daniele Roberto Giacobbe were not involved in the selection of peer reviewers for the manuscript nor any of the subsequent editorial decisions. Outside the submitted work, Matteo Bassetti has received funding for scientific advisory boards, travel, and speaker honoraria from Cidara, Gilead, Menarini, MSD, Mundipharma, Pfizer, and Shionogi. Outside the submitted work, Daniele Roberto Giacobbe reports investigator-initiated grants from Pfizer, Shionogi, BioMérieux, Menarini, Tillotts Pharma, and Gilead Italia, travel support from Pfizer, and speaker/advisor fees from Pfizer, Menarini, BioMérieux, Advanz Pharma, and Tillotts Pharma. The other authors have nothing to disclose.

Ethical Approval. The AUTO-CAND project was approved by the pertinent local ethics committee (Liguria Region Ethics Committee, registry number 71/2020). The requirement for informed consent for the present study was waived due to the retrospective nature of the analyses.

REFERENCES

1. Bassetti M, Righi E, Ansaldi F, Merelli M, Trucchi C, De Pascale G, et al. A multicenter study of septic shock due to candidemia: outcomes and predictors of mortality. Intensive Care Med. 2014;40(6):839–45. https://doi.org/10.1007/s00134-014-3310-z.

2. Bouza E, Munoz P. Epidemiology of candidemia in intensive care units. Int J Antimicrob Agents. 2008;32(Suppl 2):S87-91. https://doi.org/10.1016/S0924-8579(08)70006-2.

3. Wisplinghoff H, Bischoff T, Tallent SM, Seifert H, Wenzel RP, Edmond MB. Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study. Clin Infect Dis. 2004;39(3):309–17. https://doi.org/10.1086/421946.

4. Del Bono V, Delfino E, Furfaro E, Mikulska M, Nicco E, Bruzzi P, et al. Clinical performance of the (1,3)-beta-D-glucan assay in early diagnosis of nosocomial *Candida* bloodstream infections. Clin Vaccine Immunol. 2011;18(12):2113–7. https://doi.org/10.1128/CVI.05408-11.

5. Giacobbe DR, Mikulska M, Tumbarello M, Furfaro E, Spadaro M, Losito AR, et al. Combined use of serum (1,3)-beta-D-glucan and procalcitonin for the early differential diagnosis between candidaemia and bacteraemia in intensive care units. Crit Care. 2017;21(1):176. https://doi.org/10.1186/s13054-017-1763-5.

6. Guillamet CV, Vazquez R, Micek ST, Ursu O, Kollef M. Development and validation of a clinical prediction rule for candidemia in hospitalized patients with severe sepsis and septic shock. J Crit Care. 2015;30(4):715–20. https://doi.org/10.1016/j.jcrc.2015.03.010.

7. Leon C, Ruiz-Santana S, Saavedra P, Almirante B, Nolla-Salas J, Alvarez-Lerma F, et al. A bedside scoring system ("Candida score") for early antifungal treatment in nonneutropenic critically ill patients with *Candida* colonization. Crit Care Med. 2006;34(3):730–7. https://doi.org/10.1097/01.CCM.0000202208.37364.7D.

8. Ostrosky-Zeichner L, Sable C, Sobel J, Alexander BD, Donowitz G, Kan V, et al. Multicenter retrospective development and validation of a clinical prediction rule for nosocomial invasive candidiasis in the intensive care setting. Eur J Clin Microbiol Infect Dis. 2007;26(4):271–6. https://doi.org/10.1007/s10096-007-0270-z.

9. Paphitou NI, Ostrosky-Zeichner L, Rex JH. Rules for identifying patients at increased risk for candidal infections in the surgical intensive care unit: approach to developing practical criteria for systematic use in antifungal prophylaxis trials. Med Mycol. 2005;43(3):235–43. https://doi.org/10.1080/13693780410001731619.

10. Poissy J, Sendid B, Damiens S, Ichi Ishibashi K, Francois N, Kauv M, et al. Presence of *Candida* cell wall-derived polysaccharides in the sera of intensive care unit patients: relation with candidaemia and *Candida* colonisation. Crit Care. 2014;18(3):R135. https://doi.org/10.1186/cc13953.

11. Posteraro B, De Pascale G, Tumbarello M, Torelli R, Pennisi MA, Bello G, et al. Early diagnosis of candidemia in intensive care unit patients with sepsis: a prospective comparison of (1–>3)-beta-D-glucan assay, Candida score, and colonization index. Crit Care. 2011;15(5):R249. https://doi.org/10.1186/cc10507.

12. Giacobbe DR, Marelli C, Mora S, Cappello A, Signori A, Vena A, et al. Prediction of candidemia with machine learning techniques: state of the art. Future Microbiol. 2024;19(10):931–40. https://doi.org/10.2217/fmb-2023-0269.

13. Giacobbe DR, Marelli C, Mora S, Guastavino S, Russo C, Brucci G, et al. Early diagnosis of candidemia with explainable machine learning on automatically extracted laboratory and microbiological data: results of the AUTO-CAND project. Ann Med. 2023;55(2):2285454. https://doi.org/10.1080/07853890.2023.2285454.

14. Meng Q, Chen B, Xu Y, Zhang Q, Ding R, Ma Z, et al. A machine learning model for early candidemia prediction in the intensive care unit: Clinical application. PLoS ONE. 2024;19(9): e0309748. https://doi.org/10.1371/journal.pone.0309748.

15. Ripoli A, Sozio E, Sbrana F, Bertolino G, Pallotto C, Cardinali G, et al. Personalized machine learning approach to predict candidemia in medical wards. Infection. 2020;48(5):749–59. https://doi.org/10.1007/s15010-020-01488-3.

16. Yoo J, Kim SH, Hur S, Ha J, Huh K, Cha WC. Candidemia risk prediction (CanDETEC) model for patients with malignancy: model development and validation in a single-center retrospective study. JMIR Med Inform. 2021;9(7): e24651. https://doi.org/10.2196/24651.

17. Yuan S, Sun Y, Xiao X, Long Y, He H. Using machine learning algorithms to predict candidaemia in ICU patients with new-onset systemic inflammatory response syndrome. Front Med (Lausanne). 2021;8: 720926. https://doi.org/10.3389/fmed.2021.720926.

18. Yuan S, Xu S, Lu X, Chen X, Wang Y, Bao R, et al. A privacy-preserving platform-oriented medical healthcare and its application in identifying patients with candidemia. Sci Rep. 2024;14(1):15589. https://doi.org/10.1038/s41598-024-66596-8.

19. Giacobbe DR, Mora S, Signori A, Russo C, Brucci G, Campi C, et al. Validation of an automated system for the extraction of a wide dataset for clinical studies aimed at improving the early diagnosis of candidemia. Diagnostics (Basel). 2023;13(5):961. https://doi.org/10.3390/diagnostics13050961.

20. Mora S, Giacobbe DR, Russo C, Diana E, Signori A, Carmisciano L, et al. A wide database for future studies aimed at improving early recognition of candidemia. Stud Health Technol Inform. 2021;281:1081–2. https://doi.org/10.3233/SHTI210354.

21. Faisal S, Tutz G. Multiple imputation using nearest neighbor methods. Inf Sci. 2021;570:500–16. https://doi.org/10.1016/j.ins.2021.04.009.

22. Marchetti F, Guastavino S, Piana M, Campi C. Score-oriented loss (SOL) functions. Pattern Recogn. 2022;132: 108913. https://doi.org/10.1016/j.patcog.2022.108913.

23. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.

24. Kaneko H. Cross-validated permutation feature importance considering correlation between features. Anal Sci Adv. 2022;3(9–10):278–87. https://doi.org/10.1002/ansa.202200018.

25. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst. 2014;41(3):647–65. https://doi.org/10.1007/s10115-013-0679-x.

26. Guastavino S, Bahamazava K, Perracchione E, Camattari F, Audone G, Telloni D, et al. Forecasting geoffective events from solar wind data and evaluating the most predictive features through machine learning approaches. Astrophys J. 2024;971(1):94. https://doi.org/10.3847/1538-4357/ad5b57.

27. Brouwer L, Cunney R, Drew RJ. Predicting community-acquired bloodstream infection in infants using full blood count parameters and C-reactive protein; a machine learning study. Eur J Pediatr. 2024;183(7):2983–93. https://doi.org/10.1007/s00431-024-05441-6.

28. van den Berg MAM, Medina O, Loohuis IIP, van der Flier MM, Dudink JJ, Benders M, et al. Development and clinical impact assessment of a machine-learning model for early prediction of late-onset sepsis. Comput Biol Med. 2023;163: 107156. https://doi.org/10.1016/j.compbiomed.2023.107156.

29. Zhang F, Wang H, Liu L, Su T, Ji B. Machine learning model for the prediction of gram-positive and gram-negative bacterial bloodstream infection based on routine laboratory parameters. BMC Infect Dis. 2023;23(1):675. https://doi.org/10.1186/s12879-023-08602-4.

30. Zhang J, Liu W, Xiao W, Liu Y, Hua T, Yang M. Machine learning-derived blood culture classification with both predictive and prognostic values in the intensive care unit: a retrospective cohort study. Intensive Crit Care Nurs. 2024;80: 103549. https://doi.org/10.1016/j.iccn.2023.103549.

31. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745–50. https://doi.org/10.1016/S2589-7500(21)00208-9.

32. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15. https://doi.org/10.1038/s42256-019-0048-x.

33. Giacobbe DR, Bassetti M. The fading structural prominence of explanations in clinical studies. Int J Med Inf. 2025;197: 105835. https://doi.org/10.1016/j.ijmedinf.2025.105835.

34. Hermsen ED, Zapapas MK, Maiefski M, Rupp ME, Freifeld AG, Kalil AC. Validation and comparison of clinical prediction rules for invasive candidiasis in intensive care unit patients: a matched case-control study. Crit Care. 2011;15(4):R198. https://doi.org/10.1186/cc10366.

35. Michalopoulos AS, Geroulanos S, Mentzelopoulos SD. Determinants of candidemia and candidemia-related death in cardiothoracic ICU patients. Chest. 2003;124(6):2244–55. https://doi.org/10.1378/chest.124.6.2244.

36. Pittet D, Monod M, Suter PM, Frenk E, Auckenthaler R. *Candida* colonization and subsequent infections in critically ill surgical patients. Ann Surg. 1994;220(6):751–8. https://doi.org/10.1097/00000658-199412000-00008.

37. Guggilla V, Kang M, Bak MJ, Tran SD, Pawlowski A, Nannapaneni P, et al. Large language models outperform traditional structured data-based approaches in identifying immunosuppressed patients. medRxiv. 2025. https://doi.org/10.1101/2025.01.16.25320564.

38. Mora S, Giacobbe DR, Bartalucci C, Viglietti G, Mikulska M, Vena A, et al. Towards the automatic calculation of the EQUAL candida score: extraction of CVC-related information from EMRs of critically ill patients with candidemia in intensive care units. J Biomed Inf. 2024;156: 104667. https://doi.org/10.1016/j.jbi.2024.104667.