



UNIVERSITY OF GENOA

MASTER'S PROGRAM IN BIOENGINEERING

Thesis submitted in partial fulfillment of the requirements for the title of
Master of Bioengineering

Deep Learning for the Early Detection of Candidemia: A Post-hoc Analysis Exploring Hyperparameter Optimization

Sanaz Rashedi

October 2026

Thesis advisor: Prof. Mauro Giacomini

Thesis co-advisor: Prof. Daniele Roberto Giacobbe

Contents

Contents.....	2
Abstract.....	3
Chapter 1 Introduction.....	5
1. Candidemia vs. Bacteremia.....	5
2. The AUTO-CAND Project.....	6
Chapter 2 State of the Art.....	10
1. Diagnostic Challenges and the Limitations of Blood Culture.....	10
2. Non-Culture Biomarkers: β -D-Glucan and Procalcitonin.....	10
3. Early Applications of Machine Learning to Candidemia Prediction.....	11
4. A Recent Random Forest Model in the Intensive Care Setting.....	12
5. The AUTO-CAND Project.....	13
6. Summary of the State of the Art and Positioning of the Present Thesis.....	14
Chapter 3 Materials and Methods.....	15
1. Study Design and Dataset.....	15
2. Data Preprocessing.....	16
3. Neural Network Architecture and Hyperparameters Optimization.....	17
4. Feature Selection and Feature Importance Analysis.....	20
Chapter 4 Results.....	25
1. Source Model Architectures.....	25
2. Source Model Performance.....	26
3. Feature Importance.....	28
4. Reduced Model Experiments.....	31
Chapter 5 Discussion.....	33
1. Principal Findings.....	33
2. Comparison with Previous AUTO-CAND Studies.....	36
3. Interpretation of CBPFI Feature Selection.....	37
4. Added Value of Selected Laboratory Features Beyond BDG and PCT.....	38
5. Interpretation of the Transfer Learning Results.....	39
6. Clinical Implications.....	40
7. Strengths and Limitations.....	41

8. Future Directions.....	42
Chapter 6 Conclusion.....	44
Bibliography.....	46

Abstract

Background: Distinguishing candidemia from bacteremia at an early stage is challenging because routinely available laboratory findings and commonly used biomarkers may not be sufficient when considered on their own. The AUTO-CAND project showed that routinely collected laboratory and microbiological data can support this differentiation. However, previous modelling choices were partly empirical, motivating a more systematic and reproducible modelling strategy.

Objective: The aim of this thesis was to develop and evaluate an optimized deep-learning pipeline for differentiating candidemia from bacteremia using routinely collected laboratory data. The study also investigated whether selected laboratory variables provide additional predictive information beyond beta-D-glucan (BDG) and procalcitonin (PCT), and whether transfer learning can be useful in the smaller cohort with complete BDG and PCT measurements.

Methods: The study used data automatically extracted within the AUTO-CAND project. To reduce the risk of data leakage, data were split at the patient level, and all preprocessing was fitted using the training set only. In the final pipeline, continuous variables were transformed using a QuantileTransformer with a normal output distribution rather than conventional Z-score standardization, because several laboratory variables showed skewed distributions and extreme values. Missing values were handled using training-set median imputation; K-nearest-neighbors imputation used in an earlier phase was re-evaluated but was not retained. The neural-network architecture was engineered progressively rather than fixed a priori: the number of hidden layers was first explored from one to three, with one hidden layer selected in 8 of 10 seeds and fixed for subsequent stages. The number of neurons was then explored from 32 to 256, with 32 neurons selected in 7 of 10 seeds and fixed. Bayesian optimization was subsequently used to tune dropout, L2 regularization, and learning rate. Model development was repeated across ten random seeds. Classification thresholds were selected on the validation set using a predefined clinical condition and the True Skill Statistic (TSS). Correlation-Based Permutation Feature Importance (CBPFI) was used to identify informative laboratory

variables. Seven highest-ranked laboratory features were then evaluated together with BDG and PCT in the complete-biomarker cohort. Finally, transfer learning was investigated by transferring the learned weights associated with the seven selected laboratory features from the larger source model to the smaller target cohort, followed by freezing, unfreezing, and fine-tuning.

Results: The optimized 29-feature source model achieved a mean test TSS of 0.279 ± 0.029 and a mean ROC-AUC of 0.695 ± 0.008 across ten seeds. In the complete-BDG/PCT cohort of 1,165 episodes, with approximately 15% of candidemia episodes, BDG and PCT alone achieved a mean TSS of 0.212 ± 0.029 and a mean ROC-AUC of 0.654 ± 0.023 . Adding the seven CBPFI-selected laboratory variables increased mean TSS to 0.257 ± 0.064 (+0.045) and mean ROC-AUC to 0.674 ± 0.046 (+0.020); the nine-feature model outperformed BDG+PCT alone on TSS in 9 of 10 seeds and on ROC-AUC in 8 of 10. Transfer learning achieved a mean TSS of 0.275 ± 0.073 , compared with 0.257 ± 0.064 for direct training of the same nine-feature configuration, while mean ROC-AUC was 0.677 ± 0.040 compared with 0.674 ± 0.046 . Transfer learning produced higher TSS in 7 of 10 seeds and higher ROC-AUC in 5 of 10. Its mean sensitivity increased from 0.573 to 0.710, whereas mean specificity decreased from 0.685 to 0.565, indicating a sensitivity-specificity trade-off rather than a clear overall improvement in discrimination.

Conclusion: A systematic approach to preprocessing and progressive neural-network architecture engineering produced a reproducible deep-learning pipeline with moderate discrimination between candidemia and bacteremia. Adding selected laboratory variables to BDG and PCT improved TSS and ROC-AUC, supporting their complementary predictive value. Transfer learning produced a modest improvement in mean TSS and higher sensitivity compared with direct training, but the gain was accompanied by lower specificity and only a small difference in ROC-AUC. Further evaluation in larger, independent cohorts is required before clinical application.

Chapter 1

Introduction

1. Candidemia vs. Bacteremia

Candidemia is a bloodstream infection caused by *Candida* species and is one of the most serious forms of invasive candidiasis. It occurs mainly in hospitalized patients, especially those who are critically ill, immunocompromised, or exposed to invasive medical procedures. Although antifungal treatment and supportive care have improved, candidemia is still associated with prolonged hospitalization, high healthcare costs, and considerable mortality. Its epidemiology is also changing, with non-albicans *Candida* species being increasingly reported in many clinical settings. These species may differ in their response to antifungal drugs, making early recognition and appropriate treatment particularly important [1].

An important difficulty is that candidemia does not usually produce specific clinical signs. Fever, hypotension, sepsis, and changes in routine laboratory values can also be observed in bacterial bloodstream infections. Therefore, it may be difficult to distinguish candidemia from bacteremia during the first hours of clinical assessment. Blood culture remains the standard method for confirming candidemia, but the result is not immediately available and culture-based diagnosis may miss some cases. Clinicians may consequently need to decide whether to start antifungal therapy before microbiological confirmation is obtained [1,2].

This decision is challenging because both delayed and unnecessary treatment may have negative consequences. A delay in appropriate antifungal therapy can worsen the outcome of a patient with candidemia. On the other hand, empirical antifungal treatment in patients without candidemia can cause avoidable adverse effects, increase costs, and lead to unnecessary antifungal use. Different tools have been studied to support earlier decision-making, including clinical scores, molecular tests, and serum biomarkers. Beta-D-glucan (BDG) is one of the most widely used non-culture markers, while procalcitonin

(PCT) has also been investigated in the differential diagnosis between fungal and bacterial infection. However, the performance of these biomarkers varies among patient groups and may be influenced by the *Candida* species involved [2,3].

2. The AUTO-CAND Project

The increasing availability of electronic laboratory data has created an opportunity to investigate machine learning as an additional diagnostic support. Machine learning models can evaluate several variables at the same time and may detect relationships that are difficult to identify through individual laboratory values alone. Routine laboratory tests are useful for this purpose because they are available for most hospitalized patients, are relatively inexpensive, and are usually obtained early during admission. For these reasons, recent studies have explored whether combinations of nonspecific laboratory findings can help distinguish candidemia from bacteremia [4,5].

The AUTO-CAND project was developed within this context. Its first phase focused on creating and validating an automated system for extracting a large number of microbiological and laboratory variables from the hospital laboratory information system. The validated system allowed the construction of a large dataset containing episodes of candidemia and bacteremia, which could then be used for predictive modelling [2]. In the following phase, different supervised machine learning methods were tested for the early differential diagnosis of the two conditions. Explainability techniques were also used to investigate which laboratory variables contributed most to the model predictions. The models identified informative patterns in routinely collected data, although their overall diagnostic performance remained moderate [3].

A later AUTO-CAND study evaluated a feed-forward neural network to determine whether deep learning could capture more complex relationships among laboratory variables. The model showed some ability to distinguish candidemia from bacteremia, but it did not clearly improve the diagnostic contribution of the more specific biomarkers BDG and PCT. Feature selection and transfer learning also did not produce a clear improvement. In addition, some modelling decisions, including the choice of network architecture and training hyperparameters, were based mainly on empirical testing. These

findings suggested that a more systematic optimization and validation procedure should be investigated [4].

The present thesis continues this line of research by developing a more structured deep learning pipeline for the early differential diagnosis of candidemia and bacteremia. The planned analysis includes patient-level data splitting, preprocessing based only on the training data, median imputation, quantile transformation, automated hyperparameter optimization, and repeated evaluation using different random seeds. The classification threshold will be selected using the true skill statistic. Model interpretation and feature importance will also be examined using SHAP and correlation-aware permutation feature importance. The main objective is to assess whether an optimized and reproducible neural network based on routinely available laboratory data can improve the early distinction between candidemia and bacteremia and provide useful information in addition to BDG and PCT.

Despite these encouraging findings, the previous studies also highlighted several methodological aspects that deserve further investigation. In particular, the overall performance of deep learning models depends not only on the available clinical data, but also on numerous design decisions made during model development. The architecture of the neural network, the selection of training hyperparameters, the adopted preprocessing strategy and the choice of feature selection methods may all substantially influence the final predictive performance. However, many of these methodological choices were based primarily on empirical testing rather than on a systematic optimization strategy [4].

Furthermore, predictive performance alone is not sufficient when developing artificial intelligence models intended for clinical practice. Medical decision-support systems should also demonstrate robustness, reproducibility and transparency. Consequently, identifying the most informative laboratory variables and understanding their contribution to the final prediction have become increasingly important aspects of modern machine learning research. Explainable artificial intelligence techniques and correlation-aware feature selection methods may therefore provide valuable insights while simultaneously reducing model complexity and improving model interpretability [5,6].

Another important consideration concerns the optimization process itself. Recent advances in automated hyperparameter optimization allow different network architectures and training configurations to be explored systematically instead of relying on manual trial-and-error procedures. Likewise, alternative preprocessing strategies, regularization

techniques and feature-importance approaches may further improve the stability, transparency and generalizability of predictive models, particularly when dealing with highly imbalanced biomedical datasets [6,7].

Building upon the results obtained during the previous phases of the AUTO-CAND project, the present thesis focuses on the systematic optimization of the existing deep learning framework rather than on the development of an entirely new predictive model. The objective is to investigate whether improvements in data preprocessing, neural network hyperparameter optimization, regularization strategies and feature selection can further enhance the early differentiation between candidemia and bacteremia using routinely collected laboratory data. By adopting a reproducible optimization pipeline, this work aims to improve both the predictive performance and the clinical applicability of artificial intelligence models for supporting early diagnostic decision-making [4]. The primary objective is to determine whether a more systematic optimization strategy can enhance the early discrimination between candidemia and bacteremia while maintaining the use of routinely available laboratory information [4].

To achieve this objective, several components of the predictive pipeline are investigated. The neural-network architecture was optimized progressively: the number of hidden layers and the number of neurons were first explored, after which the most consistently selected values were fixed for the subsequent tuning stages. Learning rate, regularization strength and dropout were then optimized automatically. Adam was used as the optimizer and binary cross-entropy as the loss function. In addition, preprocessing procedures were further refined through data normalization, missing-value imputation and patient-level data splitting to ensure a more reliable and reproducible training process [4].

Another important objective of this thesis is the identification of the laboratory variables that contribute most significantly to model performance. Correlation-aware permutation feature importance and SHAP are used to investigate the contribution of routinely collected laboratory parameters to model predictions. This analysis may provide further insight into the relevance of individual laboratory variables and contribute to the development of more interpretable predictive models suitable for future clinical implementation [5,6].

Finally, the optimized deep learning framework is evaluated using repeated experiments and appropriate performance metrics specifically selected for highly

imbalanced clinical datasets. Particular emphasis is placed on the True Skill Statistic (TSS), which combines sensitivity and specificity into a single measure, together with other complementary performance indicators in order to obtain a comprehensive assessment of model performance. Through this systematic optimization process, the present work aims to contribute to the development of more reliable artificial intelligence tools capable of supporting clinicians in the early diagnosis of candidemia and ultimately improving patient management [4].

The present study is expected to contribute to the ongoing development of artificial intelligence-based decision support systems for infectious diseases. By systematically optimizing the existing deep learning framework and integrating explainable artificial intelligence techniques, this work seeks to develop a more robust, reproducible and clinically applicable model for the early differentiation between candidemia and bacteremia. The proposed optimization strategy may also provide useful methodological insights for the development of predictive models based on routinely collected laboratory data in other clinical settings [5,6].

Chapter 2

State of the Art

1. Diagnostic Challenges and the Limitations of Blood Culture

Diagnosing candidemia quickly is still genuinely hard. Blood culture remains the reference standard, but it takes 48 to 72 hours to come back, and it misses a real share of true infections along the way—enough that a substantial fraction of invasive candidiasis cases never gets picked up by culture alone [5]. Part of the problem is that candidemia just doesn't look very different from bacteremia at the bedside: fever, sepsis, and nonspecific shifts in routine bloodwork show up in both. So clinicians are often forced to decide whether to start empirical antifungal treatment before any microbiological confirmation exists [1]. That uncertainty is exactly what has pushed research toward faster, non-culture-based approaches—molecular assays, serum biomarkers, and, more recently, machine learning models built on the clinical and laboratory data hospitals already collect [1,2].

2. Non-Culture Biomarkers: β -D-Glucan and Procalcitonin

Two non-culture markers have attracted the most attention: serum (1,3)- β -D-glucan (BDG) and procalcitonin (PCT). BDG picks up a component of the fungal cell wall and is probably the best-established non-culture marker for invasive fungal infection around, though how sensitive it is depends a fair bit on which *Candida* species is involved and where it's measured [6]. PCT works differently—it tracks more closely with bacterial systemic infection and tends to sit lower in candidemia than in bacteremia. On its own, though, it isn't reliable enough to tell the two conditions apart.

Giacobbe and colleagues put this to the test directly, looking at how BDG and PCT performed together in a multicentre cohort of 166 ICU patients across two Italian hospitals—73 with candidemia, 93 with bacteremia [7]. When both markers pointed the same way toward candidemia (BDG at or above 80 pg/ml, PCT below 2 ng/ml), the positive predictive value hit 96%, well above the 79% and 66% each marker managed alone. When both agreed on bacteremia instead, the negative predictive value for candidemia came out at 95%, not far off what BDG alone gave (93%). Interestingly, when the two markers disagreed, that barely moved the pre-test probability either way. The takeaway was that combining BDG and PCT genuinely adds diagnostic confidence beyond either marker by itself—and this combined-biomarker approach is exactly the benchmark this thesis measures the added value of routine laboratory variables against, later on.

3. Early Applications of Machine Learning to Candidemia Prediction

Alongside this biomarker work, several groups have explored machine learning (ML) models based on clinical and laboratory data for early candidemia prediction. Giacobbe and colleagues reviewed this still developing literature and found that most studies relied on ensemble methods, particularly random forest (RF), and combined laboratory measurements with manually collected clinical variables such as central venous catheter use, total parenteral nutrition, and recent antibiotic exposure [8].

Ripoli and colleagues, from Fondazione Toscana Gabriele Monasterio and collaborating Italian centres, trained an RF classifier on 295 candidemia episodes and 138 bacteremia controls from internal medicine wards, achieving 84% sensitivity and 91% specificity. Yoo and colleagues, from Samsung Medical Center and Sungkyunkwan University School of Medicine, developed the CanDETEC model for patients with malignancy, with RF achieving 89% sensitivity and 90% specificity [8].

Yuan and colleagues, from Peking Union Medical College Hospital and collaborating centres in China, compared XGBoost, support vector machine, RF, and ExtraTrees for candidemia prediction in ICU patients with new-onset systemic inflammatory response

syndrome. XGBoost achieved the highest discrimination (AUC 0.92), with 84% sensitivity and 89% specificity at the selected cutoff, further supporting the potential of ML for early candidemia prediction [9].

There's a pattern worth flagging here: the studies reporting the strongest numbers tend to be the ones with smaller cohorts and hand-collected clinical detail, while datasets that are larger but automatically extracted—laboratory and microbiological variables only—tend to land on more modest performance. The review's conclusion was that explainability, external validation, adequate sample size, and the ability to automatically pull clinical, not just laboratory, features out of unstructured records are still the main things standing between ML-based candidemia prediction and actual clinical use [8].

4. A Recent Random Forest Model in the Intensive Care Setting

Meng and colleagues published after that review came out, and their study is worth a closer look. They built and externally validated an RF-based model specifically for ICU patients—a 1:1 matched cohort of 167 candidemia and 167 bacteremia episodes from one hospital, plus an independent validation set of 77 episodes from a second hospital [10]. Eighteen clinical and laboratory features, narrowed down through recursive feature elimination, went into training and comparing RF, logistic regression, support vector machine, XGBoost, and a recurrent neural network. RF won out again, with an AUC of 0.87 internally and 0.83 externally. Using SHAP values, they found total bilirubin and age were the strongest predictors, and they went a step further by building a clinical nomogram so the model's output could actually be used at the bedside.

Two things stand out here. RF keeps winning—it's the algorithm that comes out on top study after study in this space. And the reliance on manually curated clinical variables hasn't gone away either: central venous catheter use, total parenteral nutrition, chemotherapy, comorbidities, all layered on top of the laboratory data. That works fine at a few hundred patients, but manual chart review doesn't scale to the size of dataset needed for automated extraction.

5. The AUTO-CAND Project

The AUTO-CAND project addressed an important limitation of earlier studies by automating the extraction of laboratory and microbiological information from the hospital laboratory information system rather than relying on manual chart review. In its first phase, the extraction system was validated against manually checked episodes with greater than 99% accuracy and was then applied to nine years of hospital records, generating 15,752 candidemia and bacteremia episodes [2]. The main limitation was that the automated pipeline was restricted to structured laboratory and microbiological fields and therefore could not capture detailed predictors contained in free-text clinical notes, such as central venous catheter use or total parenteral nutrition [2].

Using this automatically extracted dataset, subsequent AUTO-CAND studies evaluated explainable machine-learning and deep-learning approaches for early differentiation of candidemia and bacteremia. The machine-learning phase showed that RF performed better than penalized logistic regression under the study's predefined threshold strategy, while the later deep-learning phase used a feed-forward neural network and explainability methods to investigate informative laboratory variables [5,6]. The deep-learning model achieved moderate discrimination, and adding selected laboratory variables to BDG and PCT did not produce a statistically significant improvement over BDG and PCT alone [4]. The authors also noted that several modelling decisions had been selected empirically, leaving room for more systematic optimization and validation in subsequent work [4].

6. Summary of the State of the Art and Positioning of the Present Thesis

So where does that leave things? Put together, this body of work really splits into two complementary stories. On one side, the smaller studies that mix laboratory values with hand-curated clinical variables—Ripoli, Yuan, Yoo, and Meng—generally report strong predictive performance, particularly with RF-based classifiers [7,8,11]. These results are

promising, but they come with limited sample sizes, specific patient populations, and data-collection processes that may be difficult to scale. On the other side, AUTO-CAND has shown that automated extraction can generate much larger datasets, although models restricted mainly to laboratory and microbiological variables have shown more moderate discrimination. These studies therefore highlight a continuing trade-off between dataset scale, richness of clinical information, model performance, and reproducibility [4,5,6,9].

A few gaps in that picture are what really motivate this thesis. For one, in the most recent AUTO-CAND deep learning study, the key modelling decisions—network depth, width, training hyperparameters—came largely from manual, empirical experimentation rather than a systematic search, which leaves open whether a more rigorous approach could improve both performance and stability [4]. There's also the question of stability itself: feature importance and model performance in that study were only assessed across a fairly limited number of data partitions, so it isn't really clear how consistent the selected predictors and the resulting numbers would be under different patient-level splits. And then there's the transfer-learning question—tested there, but without a clear win, and structured tabular clinical data is still a comparatively underexplored area for transfer learning compared to how established it is in imaging [11].

This is exactly what the present thesis tries to address, by building a more systematic and reproducible optimization pipeline on top of the phase-three deep learning framework. That means patient-level data partitioning with preprocessing that never touches anything outside the training data, automated Bayesian hyperparameter search over both architecture and training parameters, a validation-based threshold rule built around TSS, and correlation-based permutation feature importance (CBPFI) evaluated independently across ten random seeds rather than one. It also pushes the transfer-learning experiment further—using a two-stage, layer-freezing strategy on a fixed set of the most informative laboratory variables—to get a more rigorous, repeatable answer to whether information from the larger source cohort can actually complement BDG and PCT in the smaller subgroup where both are available. The goal, ultimately, is to move AUTO-CAND from an exploratory proof of concept toward something more methodologically solid and ready for external validation.

Chapter 3

Materials and Methods

1. Study Design and Dataset

The present study was conducted as part of the AUTO-CAND project, which aims to explore data-driven methods for the early differential diagnosis of candidemia and bacteremia. The dataset analyzed in this work was obtained through an automated data extraction system previously developed and validated within the project. This system enables the retrieval of microbiological and laboratory data from the hospital laboratory information system and their organization into a structured dataset suitable for predictive modelling [2]. During the validation process, the system successfully extracted laboratory features associated with 15,752 episodes of candidemia and/or bacteremia recorded at IRCCS Ospedale Policlinico San Martino between January 1, 2011, and December 31, 2019. Candidemia accounted for approximately 10% of all episodes, including 2% characterized by mixed candidemia and bacteremia. The prediction task was formulated as a binary classification problem, with candidemia as the positive outcome and bacteremia as the negative outcome.

The dataset used in the present study was derived from the previous phase within the AUTO-CAND project. Prior to model development, the data underwent a series of cleaning and preprocessing procedures to ensure their suitability for subsequent analyses. In particular, episodes and features with more than 50% missing values were excluded. Following these procedures, the final dataset comprised 12,483 episodes and 29 features, with candidemia accounting for approximately 10% of the episodes.

The dataset was partitioned into training, validation, and test sets using the GroupShuffleSplit method, applied sequentially to separate the training and test sets and subsequently to obtain the validation set. Patient identifiers (ID) were used as grouping

variables to ensure that episodes belonging to the same patient were not distributed across different subsets, thereby preventing potential data leakage. The resulting partitions comprised 59.5% of the episodes for training, 10.5% for validation, and 30% for testing. Unlike the splitting procedure adopted in the previously published work, GroupShuffleSplit does not perform outcome stratification. Nevertheless, an examination of the class distributions across all 10 random seeds showed that the prevalence of candidemia remained between 8.5% and 10.5% in each partition. These findings indicate that the observed splits maintained comparable outcome distributions despite the absence of explicit stratification. However, this balance was obtained by chance rather than by design, as the splitting procedure does not guarantee the preservation of class proportions.

The complete modelling procedure was repeated for ten random seeds (1–10). The same seed controlled patient splitting and the stochastic components of model development, allowing the stability of the results to be assessed across different patient partitions and model initializations.

2. Data Preprocessing

In the previously published work, missing values in the dataset used during the third phase of the analysis were handled through an iterative imputation procedure based on chained equations and a nearest-neighbor regression approach. Specifically, the IterativeImputer was configured with a KNeighborsRegressor using five nearest neighbors. The imputation model was fitted exclusively on the training set and subsequently applied to the validation and test sets to prevent data leakage. Continuous variables were then standardized using z-score normalization, which consists of subtracting the mean and dividing by the standard deviation.

In the present work, alternative preprocessing strategies were evaluated to identify the most suitable approach for the neural network models. For feature scaling, a QuantileTransformer with a normal output distribution was adopted. The transformation was fitted on the training set and subsequently applied to the validation and test sets, while categorical variables were excluded from the scaling procedure. Missing values

were then imputed using a SimpleImputer based on the median values calculated from the training set. Following imputation, categorical variables were converted into binary values by applying a threshold of 0.5, with values strictly greater than this threshold assigned to 1 and the remaining values assigned to 0.

The selection of the preprocessing strategy was guided by empirical comparisons of model performance. In particular, iterative imputation based on nearest-neighbor regression was compared with median imputation, while quantile transformation was evaluated against conventional z-score standardization. The results favored the combination of median imputation and quantile transformation. The latter also yielded more satisfactory feature distributions for the majority of the variables, supporting its adoption in the final preprocessing pipeline.

3. Neural Network Architecture and Hyperparameter Optimization

In the third phase of the previously published research, the predictive model consisted of a feedforward neural network with three hidden layers containing 100, 50, and 25 neurons, respectively. Rectified Linear Unit (ReLU) activation functions were employed in the hidden layers, while a sigmoid activation function was used in the output layer to generate binary classification probabilities. Model training was performed using the Adam optimizer with a learning rate of 10^{-5} and the Smooth One-Layer (SOL) loss function, configured to optimize the True Skill Statistic (TSS) with a uniform distribution. Dropout was fixed at 0.3 across all hidden layers, and L2 regularization was applied with a coefficient of 0.01. Training was conducted with a batch size of 256, a maximum of 10,000 epochs, and data shuffling enabled. Early stopping was based on validation loss, with a patience of 1,000 epochs. The network architecture and training hyperparameters were established empirically through a trial-and-error approach.

In contrast, the neural network architecture adopted in the present work was determined through a progressive optimization process, in which the hyperparameter search space was refined according to the results obtained across multiple random seeds. Initially,

Bayesian Optimization was employed to explore architectures containing between one and five hidden layers, using ReLU activation functions in the hidden layers and a sigmoid activation function in the output layer. Unlike the previous approach, model training was based on Binary Cross-Entropy (BCE) loss, which also served as the optimization objective during hyperparameter tuning. For each random seed, Bayesian Optimization evaluated up to 50 trials, selecting the hyperparameter configuration that minimized validation loss.

The initial search space included hidden-layer widths ranging from 64 to 256 neurons, in increments of 32, dropout rates of 0.1, 0.2, 0.3, 0.4, and 0.5, L2 regularization coefficients of 10^{-5} , 10^{-4} , 10^{-3} , 10^{-2} , and 10^{-1} , and learning rates of 10^{-5} , 10^{-4} , 10^{-3} , and 10^{-2} . The results of this first optimization showed that four of the ten random seeds selected architectures with a single hidden layer, three selected two hidden layers, two selected three hidden layers, and only one selected four hidden layers. Since nine out of ten seeds favored architectures containing no more than three hidden layers, the search space was subsequently restricted to architectures with one to three hidden layers.

The second optimization further highlighted a preference for simpler architectures, with eight seeds selecting a single hidden layer and the remaining two selecting two hidden layers. Based on these findings, the final network configuration was restricted to one hidden layer. This progressive refinement allowed the architectural choice to be informed by the consistency of the optimization results across different data partitions, rather than being established a priori.

During hyperparameter optimization, each trial was allowed to train for a maximum of 10,000 epochs, with early stopping based on validation loss, a patience of 500 epochs, and a minimum improvement threshold of 10^{-4} . Once the optimal hyperparameters had been identified, a new neural network was initialized using the selected configuration and trained independently. This final training stage employed early stopping with a patience of 1,000 epochs and restoration of the weights associated with the lowest validation loss. Consequently, the model weights retained for subsequent analyses were obtained from this final training procedure rather than directly from the best-performing Bayesian

Optimization trial. To improve reproducibility, data shuffling was explicitly disabled during both hyperparameter optimization and final model training.

An additional methodological difference between the two approaches concerns the loss function used during training. Whereas the previously published model employed SOL to directly optimize a TSS-oriented objective, the present approach used BCE to guide both neural network training and hyperparameter selection.

In the present study, the optimal classification threshold was determined using the same methodology and clinical criteria adopted in the previous phase of the AUTO-CAND project. Specifically, threshold selection was performed on the validation set by maximizing the True Skill Statistic (TSS), subject to the clinical conditions that sensitivity must be greater than or equal to specificity and specificity must be at least 0.60. If no threshold satisfied these requirements, the threshold yielding the highest validation TSS was selected without constraints. The resulting threshold was then applied to the independent test set to evaluate model performance using sensitivity, specificity, TSS, precision, negative predictive value, F1-score, accuracy, and ROC-AUC. This procedure was repeated independently across all ten random seeds to assess average predictive performance and variability between data partitions.

To assess the influence of the training objective, additional experiments were conducted comparing SOL and BCE. The two loss functions yielded broadly comparable predictive performance, without substantial differences in the evaluated metrics. However, SOL-based hyperparameter optimization tended to favor more complex architectures, typically comprising three or more hidden layers, whereas BCE consistently selected shallower networks. These findings supported the adoption of BCE in the final modelling pipeline, as it achieved comparable predictive performance while favoring a less complex network architecture.

4. Feature Selection and Feature Importance Analysis

In the third phase of the AUTO-CAND project, feature importance was investigated using two complementary permutation-based approaches: Cross-Validation Permutation Feature Importance (CVPFI) and Correlation-Driven Permutation Feature Importance (CDPFI). The importance estimates obtained from both methods were aggregated across 16 data splits to assess the consistency of feature contributions. The final selection was determined by identifying the features considered relevant by both approaches, resulting in a subset of seven laboratory variables: platelet count, urea, uric acid, hematocrit, hemoglobin, red blood cell count, and neutrophil count.

The previous research evaluated the predictive performance of models trained on these seven selected features and compared them with models using the 12 variables identified during the second phase of the project. Additional analyses were performed on the subgroup of patients for whom both β -D-glucan (BDG) and procalcitonin (PCT) measurements were available. Within this subgroup, different modelling strategies were investigated, including the use of the two biomarkers alone, their combination with selected nonspecific laboratory features, and transfer learning.

In the present work, feature importance was assessed using Correlation-Based Permutation Feature Importance (CBPFI), with ten repetitions to obtain more stable importance estimates. Correlations between predictors were calculated using the training data, whereas feature importance was evaluated on the validation set by measuring changes in model performance following feature perturbation at the previously selected classification threshold. This procedure was applied directly to the trained neural network, without additional cross-validation or model retraining during the importance calculation.

An additional analysis based on SHapley Additive exPlanations (SHAP) was also implemented to provide a complementary interpretation of the model predictions. Specifically, the KernelExplainer algorithm was employed, using subsets of the test data generated through 30 stratified folds. Background datasets were summarized through k-means clustering to reduce the computational burden associated with the estimation of SHAP values.

To investigate the predictive contribution of the selected laboratory variables and their combination with specific biomarkers, a separate analysis was conducted on the subset of episodes for which both BDG and PCT measurements were available. This restriction resulted in a dataset comprising 1,165 episodes.

For each random seed, the reduced dataset was partitioned according to the patient assignments previously established during training of the source model. Rather than generating new splits, the original training, validation, and test allocations were preserved by matching patient identifiers. The consistency of these assignments was verified to ensure that no patient appeared in multiple partitions and that every patient in the biomarker subgroup could be assigned to an existing source partition. This approach ensured that all models evaluated within the reduced cohort were trained and tested using the same patient allocations.

Two alternative feature configurations were investigated: BDG and PCT alone, and the top k nonspecific laboratory features combined with BDG and PCT. With k set to seven, these configurations corresponded to models containing two and nine input features, respectively.

Both models were initialized independently, without transferring weights from the source network or performing additional architecture optimization. Instead, the architecture and training hyperparameters identified for the corresponding source model were retained for each random seed. Each network consisted of a single hidden layer with ReLU activation, followed by a sigmoid output layer. The number of hidden neurons, dropout rate, L2 regularization coefficient, and Adam learning rate were inherited from the seed-specific source configuration, while Binary Cross-Entropy (BCE) was used as the training loss.

Data preprocessing was performed independently for each feature configuration. Continuous variables were transformed using a QuantileTransformer with a normal output distribution, while missing values were handled through median imputation using SimpleImputer. Both preprocessing procedures were fitted exclusively on the training partition of the biomarker subgroup and subsequently applied to the corresponding

validation and test partitions. When categorical variables were included, their imputed values were converted into binary values using a threshold of 0.5.

Model training was performed with a maximum of 10,000 epochs, a batch size of 64, and data shuffling disabled to improve reproducibility. Early stopping monitored validation loss, with a patience of 1,000 epochs and a minimum improvement threshold of 10^{-4} . The weights associated with the lowest validation loss were restored at the end of training.

Transfer learning was also investigated to determine whether the information acquired by a neural network trained on the larger source cohort could improve predictions in the smaller subgroup with available BDG and PCT measurements.

In the previously published phase of the AUTO-CAND project, the transfer learning strategy involved a pretrained network based on selected nonspecific laboratory features, subsequently combined with BDG and PCT. Two additional fully connected layers containing 24 and 12 neurons were introduced, respectively. The pretrained weights were initially frozen while the newly added layers were trained using the SOL loss function. The procedure also included a subsequent fine-tuning stage involving the pretrained layers.

In the present work, a modified transfer learning architecture was developed to accommodate the final source network configuration, consisting of a single hidden layer. The proposed model included two separate input branches. The first received the top k selected nonspecific laboratory features, which were processed through a ReLU-activated hidden layer containing the same number of neurons as the corresponding source model and using the same L2 regularization coefficient. The weights associated with the selected input features were extracted from the first Dense layer of the pretrained source network, together with its complete bias vector, and used to initialize this branch.

The second input branch received BDG and PCT directly, without applying a pretrained transformation. The representations from the two branches were then concatenated and passed through a newly initialized hidden layer with ReLU activation, followed by a sigmoid output layer. The number of neurons in the new hidden layer was determined

according to the width of the source network: when the source layer contained fewer than 192 neurons, its width was divided by two using integer division; otherwise, it was divided by three.

Importantly, the transfer procedure reused only the relevant input weights and biases of the pretrained hidden layer, rather than transferring the entire source network. Since the final source architecture contained a single hidden layer, the transferred parameters represented the contribution of the selected predictors to the original hidden representation. Therefore, the transferred component did not originate from a network trained exclusively on the selected features, but from the larger model trained using the complete source feature set. The BDG and PCT inputs, the additional hidden layer, and the output layer were initialized independently and did not inherit pretrained weights.

To maintain consistency between the source and transferred representations, preprocessing parameters for the nonspecific laboratory features were reconstructed by fitting the QuantileTransformer and median imputer on the saved source training data. These transformations were subsequently applied to the selected predictors in the target cohort. Conversely, BDG and PCT were processed using a separate QuantileTransformer with a normal output distribution, fitted exclusively on the training partition of the biomarker subgroup.

Transfer learning was performed in two consecutive stages. Initially, the transferred hidden layer was frozen, allowing only the newly introduced layers to be trained. During this stage, the Adam optimizer was used with a learning rate of 10^{-4} . Subsequently, the pretrained layer was unfrozen, and the entire network was fine-tuned using a lower learning rate of 10^{-5} .

Both stages employed BCE loss, a maximum of 10,000 epochs, a batch size of 64, and validation-loss-based early stopping with a patience of 1,000 epochs and a minimum improvement threshold of 10^{-4} . The best-performing weights were restored after each stage, and data shuffling was disabled throughout training to improve reproducibility.

The reduced-feature and transfer learning models were evaluated using the same threshold selection methodology adopted for the source models and the previous phase of the AUTO-CAND project. Specifically, the optimal classification threshold was determined on the validation set by maximizing TSS, subject to the clinical conditions that sensitivity must be greater than or equal to specificity and specificity must be at least 0.60. If no candidate threshold satisfied these requirements, the threshold corresponding to the maximum validation TSS was selected without constraints.

Candidate thresholds included 0.05, 0.075, 0.10, and values ranging from 0.15 to 0.50 in increments of 0.05. Predicted probabilities strictly greater than the selected threshold were classified as positive. Model performance was assessed using sensitivity, specificity, TSS, positive predictive value (PPV), weighted PPV, negative predictive value (NPV), and area under the receiver operating characteristic curve (ROC-AUC).

The evaluation was repeated across ten random seeds, and the resulting performance metrics were summarized using their means and standard deviations. These statistics were used to describe variability across different patient partitions and should not be interpreted as confidence intervals derived from independent clinical validation cohorts.

Since the three reduced-feature models and the transfer learning model were evaluated on the same subgroup of 1,165 episodes while preserving the original patient allocations, their performance could be compared under consistent data partitioning conditions. Comparisons with the source model evaluated on the complete cohort were also performed to examine differences in predictive performance. However, these comparisons were considered descriptive rather than strictly direct, given the substantial difference in dataset size and patient populations used for the respective analyses.

Chapter 4

Results

1. Source Model Architectures

Phase 3 used a fixed three hidden layer network with 100, 50 and 25 neurons. Our final models all have one hidden layer; the number of layers was fixed before the final tuning, whereas width, dropout, regularization and learning rate remained seed specific. Table 1 reports every final tuning result.

Table 1: Model architectures across 10 seeds.

Seed	Layers	Width	Dropout	L2 Reg	Learning Rate	Tuning Validation Loss
1	1	160	0.1	1e-05	1e-05	0.287
2	1	224	0.1	1e-05	1e-04	0.30
3	1	64	0.5	1e-05	1e-02	0.280
4	1	64	0.5	1e-05	1e-02	0.282
5	1	64	0.3	1e-05	1e-05	0.313
6	1	64	0.5	1e-05	1e-02	0.273
7	1	64	0.5	1e-05	1e-02	0.292
8	1	224	0.1	1e-05	1e-02	0.296
9	1	64	0.5	1e-05	1e-02	0.308
10	1	64	0.4	1e-05	1e-03	0.286

Hidden widths ranged from 64 to 224: seven seeds used 64 units, one used 160, and two used 224. Dropout ranged from 0.1 to 0.5; 0.5 was chosen in five seeds. All seeds selected $L2 = 10^{-5}$. Learning rates were 10^{-5} in two seed, 10^{-4} in one and 10^{-2} in six. These results support architectural simplification relative to Phase 3, while showing that the preferred width and optimization settings vary across patient partitions.

Final fitting rebuilds and trains the selected architecture, so its validation loss can differ from the best tuning trial; these losses should not be interpreted as the same fit.

Table 2: Model training and validation loss across 10 seeds.

Seed	Training Loss	Validation Loss
1	0.254	0.292
2	0.192	0.303
3	0.213	0.286
4	0.205	0.283
5	0.289	0.318
6	0.204	0.276
7	0.198	0.296
8	0.055	0.297
9	0.210	0.311
10	0.227	0.288

2. Source Model Performance

The source comparison uses the deliberately shared starting cohort. Test sensitivity was similar (0.70 versus 0.68), while specificity increased from 0.58 to 0.63 and TSS from 0.29 to 0.31. AUC increased from 0.64 to 0.72, an absolute difference of 0.08. Training and validation AUC also increased, to 0.80 and 0.72. Test wPPV was similar (0.865 versus 0.87) and NPV remained the same (0.95). This indicates higher discrimination with a simpler network. All values below are mean $\pm$ SD across 10 seeds for our analysis and 16 splits for Phase 3.

Table 3: Source model performance metrics comparison with phase 3.

Metrics		Training	Validation	Test
Sensitivity	Phase 3	0.80 $\pm$ 0.03	0.72 $\pm$ 0.05	0.70 $\pm$ 0.05
	Current	0.81 $\pm$ 0.05	0.70 $\pm$ 0.06	0.68 $\pm$ 0.07
Specificity	Phase 3	0.59 $\pm$ 0.06	0.57 $\pm$ 0.06	0.58 $\pm$ 0.06
	Current	0.64 $\pm$ 0.08	0.63 $\pm$ 0.08	0.63 $\pm$ 0.08
TSS	Phase 3	0.38 $\pm$ 0.05	0.29 $\pm$ 0.04	0.29 $\pm$ 0.03
	Current	0.45 $\pm$ 0.05	0.33 $\pm$ 0.05	0.31 $\pm$ 0.03
wPPV	Phase 3	0.88 $\pm$ 0.01	0.87 $\pm$ 0.01	0.87 $\pm$ 0.00
	Current	0.89 $\pm$ 0.01	0.88 $\pm$ 0.01	0.86 $\pm$ 0.01
NPV	Phase 3	0.96 $\pm$ 0.00	0.95 $\pm$ 0.01	0.95 $\pm$ 0.01

	Current	0.97 ± 0.01	0.95 ± 0.01	0.94 ± 0.01
ROC-AUC	Phase 3	0.69 ± 0.02	0.65 ± 0.02	0.64 ± 0.01
	Current	0.80 ± 0.02	0.72 ± 0.03	0.72 ± 0.01

3. Feature Importance

In the third phase of the AUTO-CAND project, feature importance was assessed using two complementary permutation-based methods: Cross-Validation Permutation Feature Importance (CVPFI) and Correlation-Driven Permutation Feature Importance (CDPFI). Importance estimates were aggregated across 16 data splits, and the intersection of the features identified by both methods was used to determine the final subset. This procedure resulted in the selection of seven laboratory variables: platelet count, urea, uric acid, hematocrit, hemoglobin, red blood cell count, and neutrophil count. The predictive performance of models trained using these seven variables was compared with that of models incorporating the 12 features previously identified during the second phase of the project. Additional experiments were conducted on the subgroup of patients with available β -D-glucan (BDG) and procalcitonin (PCT) measurements, evaluating the use of biomarkers alone, their combination with selected nonspecific laboratory features, and transfer learning.

In the present work, feature importance was evaluated using Correlation-Based Permutation Feature Importance (CBPFI), with ten repetitions to improve the stability of the estimates. Correlations between predictors were calculated from the training data, whereas feature importance was determined by perturbing the validation data and measuring the resulting decrease in TSS at the selected classification threshold. This procedure was applied to the previously trained neural network without performing additional cross-validation or model retraining. A complementary analysis was conducted using SHapley Additive exPlanations (SHAP), implemented through KernelExplainer. SHAP values were estimated on test subsets generated using 30 stratified folds, with k-means clustering employed to summarize the background data and reduce computational complexity.

Feature importance was assessed for all 29 predictors across ten random seeds. The average CBPFI importance is expressed as the mean decrease in TSS and its standard deviation. The corresponding SHAP rankings were obtained by ordering features according to their average absolute SHAP values. These two measures capture different aspects of model behavior: CBPFI quantifies the reduction in predictive performance following feature perturbation, whereas SHAP describes the magnitude of individual feature contributions to model predictions. Consequently, their numerical values are not directly comparable. Furthermore, unlike the present analysis, which considers average absolute SHAP values, the third phase of the project reported mean SHAP ranking positions.

According to the average CBPFI results, neutrophil count and white cell count emerged as the most influential predictors. These were followed by previous Candida colonization, platelet count, hematocrit, urea, and red blood cell count. Five of these seven highest-ranked variables were also included in the feature subset selected during the third phase of the project. Uric acid and hemoglobin, which were among the previously selected features, occupied the ninth and eleventh positions, respectively, in the current CBPFI ranking.

The SHAP analysis revealed a partially different ordering of the predictors. Platelet count ranked first, followed by urea, previous Candida colonization, white cells count, alanine aminotransferase (ALT), C-reactive protein (PCR), and alkaline phosphatase (ALP). Four of the top seven features were also present in the feature ranking of the third phase. Conversely, white blood cell count and ALT occupied higher positions in the present ranking, while red blood cell count and hemoglobin were ranked eighteenth and twenty-first, respectively. Neutrophil count also exhibited a notable difference between the two importance methods, ranking first according to CBPFI but twentieth according to SHAP.

Table 4: Comparison of aggregated feature rankings derived from CBPFI and SHAP.

Feature	CBPFI rank	SHAP rank	Rank difference
NEUTROPHIL CELLS COUNT	1	20	19
WHITE CELLS COUNT	2	4	2
PREVIOUS CANDIDA	3	3	0

COLONIZATION			
PLATELET COUNT	4	1	3
HEMATOCRIT	5	10	5
UREA	6	2	4
RED CELLS COUNT	7	18	11
ALBUMIN	8	17	9
HEMOGLOBIN	9	21	12
CREATININE	10	8	2
URIC ACID	11	9	2
MONOCYTE CELLS COUNT	12	28	16
INR	13	11	2
ALP	14	7	7
LAD	15	12	3
APTT	16	13	3
TOTAL BILIRUBIN	17	16	1
PROTHROMBIN TIME	18	25	7
GGT	19	14	5
TOTAL PROTEINS	20	19	1
PCR	21	6	15
ALT	22	5	17
LYMPHOCYTE CELLS COUNT	23	27	4
BASOPHIL CELLS COUNT	24	22	2
AST	25	26	1
EOSINOPHIL CELLS COUNT	26	23	3
AGE	27	24	3
GLUCOSE	28	15	13
SEX	29	29	0

The aggregated CBPFI and SHAP rankings across 29 features showed a moderate positive correlation (Spearman's $\rho = 0.5335$, $p = 0.00288$). Four predictors were shared between the seven highest-ranked features of both methods: platelet count, urea, previous Candida colonization, and white blood cell count.

Despite this overall agreement, substantial differences were observed for individual predictors. Neutrophil count ranked first by CBPFI but twentieth by SHAP, representing the largest discrepancy of 19 positions. Monocyte count, hemoglobin and red blood cell count also ranked higher by CBPFI, with differences of 16, 12 and 11 positions,

respectively. Conversely, alanine aminotransferase and C-reactive protein ranked higher by SHAP, occupying fifth and sixth positions compared with twenty-second and twenty-first positions by CBPFI. Sex was ranked last by both methods.

Agreement varied across the 10 seeds, with Spearman’s ρ ranging from 0.2049 to 0.6350 and a mean $\pm$ SD of 0.4167 ± 0.1342 . Six seeds showed nominally significant correlations at $p < 0.05$, without adjustment for multiple comparisons. The mean absolute difference between feature ranks ranged from 5.59 to 8.41 positions across seeds. Thus, the methods showed moderate agreement in the aggregated analysis, alongside variation between seeds and marked differences in the ranking of specific predictors.

4. Reduced Model Experiments

The reduced-feature experiments were conducted on the subset of 1,165 episodes for which both BDG and PCT measurements were available. Three feature configurations were evaluated: the two biomarkers alone, the top seven laboratory features selected through CBPFI combined with BDG and PCT, and the seven selected laboratory features without the biomarkers. These configurations resulted in models with two, nine, and seven input features, respectively. For each random seed, the training, validation, and test partitions were inherited from the corresponding source model, preserving the original patient assignments and ensuring consistent data splits across all three configurations.

Biomarkers alone produced the highest mean test AUC among the three current subgroup models (0.71). Compared with Phase 3, test specificity increased from 0.71 to 0.76, while sensitivity decreased from 0.64 to 0.53. TSS decreased from 0.35 to 0.29 despite the higher AUC. Test wPPV and NPV were 0.810 and 0.90, compared with 0.82 and 0.92. This reflects a different operating tradeoff; a higher AUC alone does not establish better sensitivity at the chosen threshold.

Table 5: BDG+PCT model performance comparison with phase 3.

Metrics		Training	Validation	Test
Sensitivity	Phase 3	0.66 ± 0.06	0.68 ± 0.10	0.64 ± 0.07
	Current	0.53 ± 0.14	0.57 ± 0.15	0.53 ± 0.17

Specificity	Phase 3	0.70 ± 0.05	0.70 ± 0.06	0.71 ± 0.05
	Current	0.75 ± 0.13	0.78 ± 0.12	0.76 ± 0.13
TSS	Phase 3	0.36 ± 0.03	0.38 ± 0.10	0.35 ± 0.05
	Current	0.28 ± 0.04	0.36 ± 0.13	0.29 ± 0.07
wPPV	Phase 3	0.82 ± 0.01	0.83 ± 0.03	0.82 ± 0.01
	Current	0.81 ± 0.01	0.83 ± 0.05	0.81 ± 0.03
NPV	Phase 3	0.92 ± 0.01	0.92 ± 0.02	0.92 ± 0.01
	Current	0.90 ± 0.01	0.91 ± 0.03	0.90 ± 0.03
ROC-AUC	Phase 3	0.68 ± 0.01	0.69 ± 0.06	0.68 ± 0.03
	Current	0.72 ± 0.02	0.71 ± 0.07	0.71 ± 0.03

Adding the top seven predictors to the biomarkers yielded test sensitivity 0.64, specificity 0.64, TSS 0.28 and AUC 0.69. Compared with the corresponding Phase 3 model, sensitivity increased by 0.05 and AUC by 0.03, while specificity decreased by 0.08 and TSS by 0.03. Compared with current biomarkers alone, sensitivity increased by 0.11, but specificity decreased by 0.12; test TSS and AUC decreased by 0.01 and 0.02. Adding the selected features therefore did not improve mean test TSS or AUC.

Table 6: Top 7 features + BDG+PCT model performance comparison with phase 3.

Metrics		Training	Validation	Test
Sensitivity	Phase 3	0.76 ± 0.07	0.66 ± 0.10	0.59 ± 0.08
	Current	0.74 ± 0.13	0.68 ± 0.15	0.64 ± 0.13
Specificity	Phase 3	0.73 ± 0.07	0.73 ± 0.07	0.72 ± 0.08
	Current	0.68 ± 0.11	0.68 ± 0.11	0.64 ± 0.13
TSS	Phase 3	0.50 ± 0.07	0.38 ± 0.10	0.31 ± 0.06
	Current	0.42 ± 0.10	0.36 ± 0.12	0.28 ± 0.06
wPPV	Phase 3	0.86 ± 0.01	0.83 ± 0.02	0.81 ± 0.01
	Current	0.84 ± 0.02	0.83 ± 0.04	0.81 ± 0.02
NPV	Phase 3	0.95 ± 0.01	0.92 ± 0.02	0.91 ± 0.01
	Current	0.94 ± 0.02	0.92 ± 0.03	0.91 ± 0.02
ROC-AUC	Phase 3	0.75 ± 0.04	0.69 ± 0.05	0.66 ± 0.03
	Current	0.80 ± 0.07	0.73 ± 0.08	0.69 ± 0.04

Transfer learning had test sensitivity 0.63, specificity 0.65, TSS 0.27 and AUC 0.70. Relative to Phase 3 transfer, AUC improved by 0.03, but sensitivity, specificity and TSS were lower by 0.02, 0.03 and 0.06. The test AUC was below biomarkers alone (0.71), so

the transferred representation did not yield a higher mean test AUC in the current subgroup.

Table 8: Transfer Learning model with top 7 features + BDG+PCT performance comparison with phase 3.

Metrics		Training	Validation	Test
Sensitivity	Phase 3	0.68 ± 0.07	0.71 ± 0.12	0.65 ± 0.07
	Current	0.71 ± 0.14	0.69 ± 0.12	0.63 ± 0.16
Specificity	Phase 3	0.68 ± 0.06	0.68 ± 0.05	0.68 ± 0.06
	Current	0.69 ± 0.15	0.68 ± 0.12	0.65 ± 0.15
TSS	Phase 3	0.31 ± 0.05	0.39 ± 0.11	0.33 ± 0.05
	Current	0.36 ± 0.06	0.38 ± 0.14	0.27 ± 0.06
wPPV	Phase 3	0.87 ± 0.01	0.86 ± 0.01	0.85 ± 0.00
	Current	0.84 ± 0.02	0.83 ± 0.04	0.81 ± 0.02
NPV	Phase 3	0.92 ± 0.01	0.93 ± 0.03	0.92 ± 0.01
	Current	0.93 ± 0.02	0.92 ± 0.03	0.91 ± 0.02
ROC-AUC	Phase 3	0.68 ± 0.03	0.70 ± 0.05	0.67 ± 0.02
	Current	0.79 ± 0.05	0.71 ± 0.08	0.70 ± 0.03

Chapter 5

Discussion

1. Principal Findings

This study built on the deep learning framework developed earlier in the AUTO-CAND project, optimizing and extending it for the early differential diagnosis of candidemia and bacteremia. The findings fall into three areas: how reproducible and well-optimized the source neural network turned out to be, which laboratory variables proved informative once correlation-aware feature selection was applied, and whether those selected variables, together with transfer learning, could improve prediction in the smaller cohort with complete BDG and PCT measurements.

The optimized source model achieved moderate discrimination overall, and its ROC-AUC stayed fairly stable across the ten experimental seeds. Sensitivity and specificity shifted more between patient-level partitions, but the tight spread in ROC-AUC points to a model whose underlying ability to rank cases held up reasonably well across repeated runs - notable given how imbalanced the source dataset was, with candidemia making up only around 10% of episodes.

This lines up with earlier AUTO-CAND work. The first explainable machine-learning study found moderate performance for telling candidemia apart from bacteremia using routine laboratory and microbiological data, with its best random forest reaching a test TSS of 0.31 [3]. A later deep-learning study reached a similar conclusion - a neural network could pull discriminatory signal out of routine laboratory data, but performance stayed moderate, with an AUC of 0.64, sensitivity of 0.70, and specificity of 0.58 [4].

So systematic neural-network optimization doesn't turn this into a near-perfect discrimination task - that's not what these results show. What they do show is that a structured optimization pipeline can yield a reproducible model with stable performance,

and one that depends less on modelling choices picked by hand. The value of this work isn't only in predictive performance, then; it's also in setting up a more systematic framework for preprocessing, hyperparameter optimization, threshold selection, feature evaluation, and repeated seed-based validation.

A second finding was that routine laboratory variables added predictive information on top of BDG and PCT. Of all the configurations tested in the complete-biomarker cohort, the model built on the two biomarkers alone performed worst on average. Bringing in the seven CBPFI-selected laboratory variables improved both TSS and ROC-AUC in most seed-specific comparisons - supporting the idea that distinguishing candidemia from bacteremia can benefit from combining biomarkers with the broader patterns already sitting in routine laboratory data.

This fits with what's already known about BDG and PCT. BDG is a well-established non-culture marker for invasive fungal infection, though its sensitivity shifts depending on the *Candida* species and clinical setting [6]. PCT, for its part, tends to run lower in candidemia than in bacteremia, but on its own it isn't reliable enough to distinguish the two conditions. That's why combining BDG and PCT has been proposed as more informative than either marker alone [7]. What the present results add is that routinely available laboratory information may complement these biomarkers even further.

The transfer-learning results were more mixed. Mean TSS came out higher than direct training of the same nine-feature configuration, and TSS was higher in seven of the ten seeds. ROC-AUC, though, was nearly identical between the two approaches and didn't consistently favor transfer learning. The sensitivity gain also came with a drop in specificity, so it shouldn't be read on its own as proof that the model got better at spotting positive cases - the complete-BDG/PCT cohort simply had a much higher proportion of candidemia cases than the source cohort, and that difference in prevalence and patient mix has to be factored into how the sensitivity-specificity balance is interpreted.

2. Comparison with Previous AUTO-CAND Studies

Because this work builds directly on earlier phases of the AUTO-CAND project, it's worth comparing it closely with the machine-learning and deep-learning approaches already developed for the same clinical problem. In the first explainable machine-learning study, a random forest reached a test sensitivity of 0.74, specificity of 0.57, and TSS of 0.31 for distinguishing candidemia from bacteremia - evidence that routine laboratory and microbiological information does carry useful discriminatory signal, even if performance stayed moderate [3].

The follow-up deep-learning study asked whether a neural network could pick up more complex, nonlinear relationships among these same routine laboratory variables. It reached a test AUC of about 0.64, with sensitivity near 0.70 and specificity near 0.58 [4]. That study also tested whether information learned from nonspecific laboratory variables could boost prediction when combined with the more specific biomarkers BDG and PCT - but neither feature selection nor transfer learning gave a clear edge over the BDG+PCT model alone [4].

What sets the present study apart from that earlier work is mainly how systematically the model was optimized and validated. Instead of leaning on modelling choices picked empirically, the pipeline here progressively optimized network architecture and training parameters, kept preprocessing contained within each training split, repeated model development across ten patient-level seeds, and chose classification thresholds according to a predefined, validation-based clinical criterion - all with the aim of improving reproducibility and reducing how much any single data partition or modelling decision could sway the result.

This more structured strategy does appear to have produced a source model with stable discrimination across seeds. But the improvement is better understood as methodological refinement than as a dramatic jump in diagnostic accuracy - which fits the broader machine-learning literature on candidemia, where studies are still relatively scarce and

sample size, data quality, bias, generalizability, and clinical interpretability remain real barriers to routine use [8].

One notable point of divergence from the earlier deep-learning study showed up in the reduced-input experiments. Here, adding the seven seed-specific CBPFI-selected laboratory variables to BDG and PCT improved TSS and ROC-AUC in most seed-wise comparisons - unlike the earlier AUTO-CAND deep-learning analysis, where feature selection didn't clearly help BDG and PCT's diagnostic performance [4]. A few things could explain the difference: the revised feature-ranking procedure, seed-specific feature sets, the more systematic optimization pipeline, and comparing models in pairs within identical patient-level partitions. Even so, these findings are exploratory until confirmed and externally validated.

3. Interpretation of CBPFI Feature Selection

The CBPFI analysis turned up a fairly consistent group of routine laboratory variables across all ten seeds. Neutrophil count made the top seven in every single one; white blood cell count and previous *Candida* colonization within 30 days each made it in nine of ten. Urea and platelet count showed up often too, followed by uric acid and creatinine.

That these variables kept coming back up suggests the network wasn't leaning on any single laboratory marker, but on a broader mix of hematological, inflammatory, renal, and microbiological information. Clinically, that makes sense: candidemia and bacteremia often present with overlapping systemic signs, so no single routine value is likely to separate them cleanly. A multivariable pattern is simply more informative than any one nonspecific marker on its own.

Previous *Candida* colonization stands out in particular, since it combines routine microbiological information with a direct link to *Candida* exposure - its frequent selection fits the established clinical idea that prior colonization raises the odds of later invasive infection. Neutrophil count and white blood cell count are a different story: both are nonspecific markers of host inflammation, so their repeated importance doesn't mean

either one is individually diagnostic of candidemia. It more likely means they add value when the model reads them together with the other laboratory variables.

Renal parameters like urea and creatinine showing up may reflect differences in severity, comorbidity, organ dysfunction, or treatment exposure between the two patient groups. But since this study was built for prediction, not causal inference, the CBPFI rankings shouldn't be read as evidence that these variables mechanistically cause candidemia. All their importance really tells us is that perturbing them changed predictive performance - not that they're independent causal risk factors.

It's also worth noting that the top-seven subset itself varied across patient partitions, so feature importance wasn't completely fixed. At the same time, several variables kept reappearing across most seeds, pointing to a core group of predictors that contributed consistently. That mix of stability and variation is really the justification for evaluating feature importance across multiple seeds rather than trusting a single fitted model.

4. Added Value of Selected Laboratory Features Beyond BDG and PCT

Comparing the BDG+PCT model against the nine-feature model gets at one of the central questions of this study: can routine laboratory information add discriminatory value on top of two more specific biomarkers? By and large, yes - the nine-feature model outperformed BDG+PCT alone on TSS and ROC-AUC, with the improvement showing up in most seed-wise comparisons.

That matters because BDG and PCT capture biologically different, complementary information: BDG reflects fungal cell-wall components and is widely used as a non-culture marker of invasive fungal infection, while PCT tracks more closely with bacterial systemic infection. Reading them together has already been shown to improve differential diagnosis compared with either marker alone [7].

The present results push this idea a step further, suggesting that routine laboratory variables can complement the diagnostic information already in BDG and PCT. That's

practically useful, since the additional selected variables are generated during standard care anyway and don't call for any specialized testing.

That said, the improvement was moderate, not dramatic - the selected laboratory variables add incremental information rather than replacing the more specific biomarkers. So the takeaway is a complementary modelling strategy, where routine laboratory data gets integrated with biomarker information rather than treated as a competing approach.

One more thing worth keeping in mind: the complete-BDG/PCT cohort had a higher candidemia prevalence than the source dataset - about 10% in the source cohort versus roughly 18.7% overall in the complete-biomarker cohort. Since that kind of prevalence difference affects metrics like PPV and NPV, and can shift the sensitivity-specificity balance, paired comparisons between BDG+PCT and the nine-feature model within the same target cohort carry more weight than comparing either of them directly against the larger source model.

5. Interpretation of the Transfer Learning

Results

Transfer learning was tested to see whether what the network learned from the larger source cohort could be reused in the smaller cohort with complete BDG and PCT measurements. This mattered because requiring both biomarkers cut the number of available observations substantially, and training a network from scratch on that smaller target dataset risks limiting how much it can reliably learn.

The transfer-learning model edged out the directly trained nine-feature model on mean TSS, and had the higher TSS in seven of ten seeds. The gap in ROC-AUC between the two, though, was small and didn't consistently favor transfer learning. So it would be a stretch to conclude transfer learning was generally superior to direct training - what the results actually suggest is that reusing information from the larger source cohort helps under some patient-level partitions, but not reliably across all of them.

One thing that stood out was how the sensitivity-specificity balance shifted: transfer learning raised mean sensitivity but lowered mean specificity compared with direct

training. Higher sensitivity is generally valuable in a condition like candidemia, where missing a true case can have serious consequences - but this needs careful interpretation, since the target cohort had a higher proportion of candidemia cases than the source cohort, and differences in patient composition and prevalence likely contributed to the pattern.

More broadly, transfer learning is often proposed as a useful strategy whenever the target dataset is substantially smaller than the one used to build the original model. It has shown potential for non-image clinical data, though applications to structured tabular data are still less established than in medical imaging [11]. The present findings fit that pattern - encouraging, but preliminary. Confirming a consistent advantage for this diagnostic task would need validation in a larger, independent cohort.

6. Clinical Implications

Telling candidemia and bacteremia apart early on remains genuinely hard, since both conditions can present with similar signs and routine laboratory abnormalities. Delayed antifungal treatment has been linked to worse outcomes in candidemia patients, which is exactly where predictive models that draw on information already collected during routine assessment could play a role.

These findings suggest that routinely collected laboratory variables can complement BDG and PCT when estimating the probability of candidemia. Practically, that's an appealing prospect, since most of the selected variables are measured anyway during standard care - no extra diagnostic procedures needed.

None of the models tested here should be seen as a replacement for microbiological diagnosis or established clinical judgment, though. Their realistic role would be as decision-support tools - flagging patients who might need further work-up or closer evaluation for candidemia. That's in line with the broader literature, where machine-learning models for candidemia are considered promising but still in need of stronger external and prospective validation before routine clinical use [8].

The relatively high NPV seen in some configurations could matter for future clinical use, but predictive values depend heavily on disease prevalence, so the PPV and NPV reported here can't be assumed to hold in a different hospital or patient population. External validation would be essential before any of this reaches clinical use.

7. Strengths and Limitations

This study has several strengths worth highlighting. First, the modelling pipeline was evaluated across ten patient-level seeds rather than a single train-validation-test split, which made it possible to examine variability tied to patient allocation more directly.

Second, all preprocessing was fitted on the training data alone and then applied to the validation and test sets - combined with patient-level splitting, this reduced the risk of information leakage between model development and evaluation.

Third, network architecture and training hyperparameters were optimized systematically rather than by hand, and CBPFI allowed feature importance to be evaluated while accounting for correlation between predictors. Finally, BDG+PCT, the nine-feature model, and transfer learning were all compared using matched observations within each seed, which made for a more direct comparison between the three approaches.

There are important limitations too. The biggest is the small size of the complete-BDG/PCT cohort - the source dataset held more than 10,000 episodes, but complete BDG and PCT measurements existed for only a much smaller subset. That caps how much information is available for training and adds uncertainty to the evaluation of the reduced-input and transfer-learning models.

A second limitation is class imbalance: candidemia was a minority class in both the source and target datasets. Class weighting and validation-based threshold selection helped address this during model development, but the imbalance still needs to be kept in mind when interpreting the performance metrics.

Third, this analysis relied on retrospective data from the existing AUTO-CAND dataset, with no independent external validation. That matters especially for clinical machine-learning models, since performance can shift when applied to hospitals or populations with different disease prevalence, laboratory practices, or clinical characteristics - a point also raised in previous reviews of machine-learning approaches for candidemia [8].

Feature-selection stability is another limitation. Several variables were repeatedly selected across seeds, but the complete top-seven set was never identical from one partition to the next, which means feature ranking still depends partly on which patients happened to be sampled.

Finally, transfer learning wasn't uniformly better than direct training - mean TSS and sensitivity were higher, but ROC-AUC was about the same and specificity was lower. These results are best read as evidence of potential usefulness, not proof that transfer learning is the optimal approach for this dataset.

8. Future Directions

The first priority for future work is evaluating the final models in independent patient cohorts. External validation across different hospitals would show whether this performance holds up when disease prevalence, patient characteristics, and clinical practice differ from the development dataset.

Larger datasets with complete BDG and PCT measurements would help too. A bigger target cohort could clarify whether the gap between direct training and transfer learning mainly comes down to limited data availability, or reflects a genuine advantage of the transferred representations.

It would also be worth checking how stable the CBPFI-selected variables are in independent datasets. Features that stay important across institutions and populations would be stronger evidence of a reproducible predictive signature; large shifts in the

rankings, on the other hand, would suggest the model depends heavily on local population characteristics.

Alternative transfer-learning strategies are also worth exploring, compared against direct training using identical evaluation procedures. Since transfer learning with structured clinical data is still a developing area, carefully comparing different strategies could help pin down which components of a source model are actually worth reusing [11].

Finally, future studies should move past retrospective performance metrics and test clinical utility prospectively - looking at how predictions could fit into clinical workflows, whether they actually improve diagnostic decision-making, and whether they can support earlier, more appropriate antifungal management without driving an unacceptable rise in unnecessary treatment.

Chapter 6

Conclusion

This thesis investigated whether routinely collected laboratory data could be used to improve the differentiation between candidemia and bacteremia through a deep learning approach within the AUTO-CAND project. The work focused on making the modelling process more systematic and reproducible, while also examining the additional value of routine laboratory variables beyond BDG and PCT and the possible use of transfer learning in a smaller cohort.

The optimized source model showed a moderate level of discrimination and relatively stable performance across the ten patient-level seeds. It achieved a mean test TSS of 0.279 ± 0.029 and a mean ROC-AUC of 0.695 ± 0.008 . Although some variation was observed between the different patient partitions, the overall performance remained reasonably consistent.

The results also showed that routinely collected laboratory variables can add useful information to BDG and PCT. The model based on the seven laboratory variables selected using CBPFI performed better on average than BDG and PCT alone. This suggests that routine laboratory information can complement these biomarkers when differentiating candidemia from bacteremia. However, the improvement was moderate, so these laboratory variables should be considered as complementary information rather than a replacement for BDG and PCT.

Transfer learning was also investigated because the cohort with complete BDG and PCT measurements was much smaller than the source cohort. The transfer-learning model achieved a higher mean TSS than the directly trained nine-feature model and had a higher TSS in seven of the ten seeds. However, the difference in ROC-AUC was small and did not consistently favor transfer learning. The higher sensitivity was also accompanied by lower specificity. Therefore, the results suggest that transfer learning may be useful in

this setting, but they do not provide enough evidence to conclude that it is consistently better than direct training.

Overall, this study supports the potential of combining routinely collected laboratory information with BDG and PCT for the differentiation of candidemia and bacteremia. At the same time, the models should not be considered a replacement for microbiological diagnosis or clinical judgment. The relatively small complete-BDG/PCT cohort, class imbalance, retrospective nature of the data, variation in feature selection, and lack of independent external validation remain important limitations. Further evaluation in larger and independent patient cohorts, including external and prospective validation, is needed before these models can be considered for clinical use.

Bibliography

- [1] Mustafayev K, Kuşku MA, Akkoç-Mustafayev FN, et al. Early diagnosis of candidemia in the intensive care unit by clinical and molecular methods: a prospective observational study. *Infectious Diseases and Clinical Microbiology*. 2024;6(4):306–319. doi:10.36519/idcm.2024.443.
- [2] Giacobbe DR, Mora S, Signori A, 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*. 2023;13(5):961. doi:10.3390/diagnostics13050961.
- [3] Giacobbe DR, Marelli C, Mora S, et al. Early diagnosis of candidemia with explainable machine learning on automatically extracted laboratory and microbiological data: results of the AUTO-CAND project. *Annals of Medicine*. 2023;55(2):2285454. doi:10.1080/07853890.2023.2285454.
- [4] Giacobbe DR, Guastavino S, Razzetta A, et al. Deep learning for the early diagnosis of candidemia. *Infectious Diseases and Therapy*. 2025;14:1529–1545. doi:10.1007/s40121-025-01171-w.
- [5] Clancy CJ, Nguyen MH. Finding the ‘missing 50%’ of invasive candidiasis: how nonculture diagnostics will improve understanding of disease spectrum and transform patient care. *Clinical Infectious Diseases*. 2013;56(9):1284–1292. doi:10.1093/cid/cit006.
- [6] Ullah N, Muccio M, Magnasco L, et al. Species-specific sensitivity and levels of beta-D-glucan for the diagnosis of candidemia: a systematic review and meta-analysis. *Journal of Fungi*. 2025;11(2):149. doi:10.3390/jof11020149.
- [7] 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. doi:10.1186/s13054-017-1763-5.

- [8] 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):931-940. doi:10.2217/fmb-2023-0269.
- [9] 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. doi:10.3389/fmed.2021.720926.
- [10] 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. doi:10.1371/journal.pone.0309748.
- [11] Ebbeløj A, Thunbo MØ, Andersen OE, Glindtved MV, Hulman A. Transfer learning for non-image data in clinical research: a scoping review. *PLOS Digit Health*. 2022;1(2):e0000014. doi:10.1371/journal.pdig.0000014.