Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine learning implementation for host-based infectious disease diagnostics

Samuel Harrington

samharrington0711@gmail.com

Los Alamos National Laboratory https://orcid.org/0009-0000-9534-9800

Research Article

Keywords: Machine learning, Diagnostics, Host-Based, Bioinformatics, Bacterial, Viral

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10907969/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

frontiers

Machine Learning Implementation for Host-Based Infectious Disease Diagnostics

Samuel Harrington [1], **Sarah Nace** [1], **Kaitlyn Martinez** [2,*] **and Jessica Kubicek-Sutherland** [1]

[1] *Chemistry Division, Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America*
[2] *Analytics, Intelligence and Technology Division, Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America*

Correspondence*:
Kaitlyn Martinez
kaitlynm@lanl.gov

ABSTRACT

Host-based diagnostics coupled with machine learning may support faster discrimination of infectious states and reduce unnecessary antimicrobial treatment. We reviewed recent studies that applied machine-learning methods to clinical patient data for pathogen detection, infection classification, or closely related host-response diagnosis. Eligible infection targets included bacterial and viral infections, bacterial-versus-viral classification, and syndrome-based presentations such as sepsis. Malaria was included when evaluated within a mixed febrile cohort. Searches were conducted of four literature databases spanning January 2020 through August 2025. Of 144 retrieved records, 51 passed an initial host-diagnostics screen and 21 articles met the final eligibility criteria. The included articles covered structured clinical variables, gene or molecular-expression measurements, medical imaging, physiological signals, and microscopy. For the quantitative model-class comparison, we extracted 54 evaluations belonging to four machine-learning classes: random forest (23), neural network (17), support vector machine (9), and logistic regression (5). Less frequently evaluated approaches, including gradient boosting, decision trees, naive Bayes, and k-nearest neighbors, were summarized descriptively. Model performance could not be compared directly because input modalities, cohort composition, preprocessing, validation strategies, hyperparameter selection, and reported endpoints varied substantially. Accuracy and area under the receiver operating characteristic curve (AUC) were the dominant performance measures. Among the 54 focal-class evaluations, 14 reported accuracy alone, 20 reported AUC alone, and 20 reported both. Only 16 evaluations reported additional measures such as sensitivity, specificity, F1 score, or Matthews correlation coefficient. Imaging studies generally favored neural-network approaches, whereas molecular-expression studies frequently used random forests. However, the evidence did not support a universally optimal model.

Keywords: machine learning, host-based diagnostics, infectious disease, clinical diagnostics, random forest, neural network

1 INTRODUCTION

Early identification of an infectious disease etiology is central to appropriate clinical treatment. When laboratory tests are unavailable, delayed, or inconclusive, clinicians may begin treatment from symptom presentation and standard-of-care assessments. This uncertainty contributes to avoidable antibiotic use. An analysis of United States ambulatory-care visits estimated that approximately 30% of outpatient antibiotic prescriptions were unnecessary (Fleming-Dutra et al., 2016). Inappropriate antimicrobial exposure can promote the selection and spread of resistant organisms (CDC, 2024).

Host-based diagnostics evaluate the patient's biological response rather than directly testing for a predetermined pathogen. Relevant measurements can include blood-cell counts, inflammatory proteins, cytokines, transcriptional signatures, metabolites, microbiome features, imaging findings, and physiological signals. Machine learning is a family of computational methods that identifies patterns in labeled or unlabeled data and uses those patterns for prediction or classification (IBM, n.d.). Combining host-response measurements with machine learning may therefore provide a broad diagnostic aid when the causative pathogen is unknown.

Current clinical evaluation can include complete blood counts, C-reactive protein measurements, procalcitonin, bacterial cultures, and pathogen-specific polymerase chain reaction assays (Cleveland Clinic, n.d.a; Largman-Chalamish et al., 2022; Paudel et al., 2020; MedlinePlus, n.d.; Cleveland Clinic, n.d.b). Each approach provides useful information, but no single test resolves all presentations. Culture can be slow, targeted PCR requires a suspected organism, and nonspecific inflammatory biomarkers may overlap across bacterial, viral, and noninfectious conditions. A host-response model could complement these methods by estimating an infection category or clinically relevant state while more definitive testing proceeds.

The model-data relationship is especially important in biological applications. Transcriptomic measurements, structured clinical variables, microscopy images, computed tomography scans, and electrocardiographic signals differ in dimension, noise structure, missingness, and interpretability. Since these measurements may also vary with age, sex, comorbidities, treatment exposure, and immune status, these characteristics may confound model development or contribute to performance shifts across patient populations. Consequently, an algorithm that performs well for one modality may not be appropriate for another. This review examined how recent host-based diagnostic studies paired data modalities with machine-learning methods, how model performance was evaluated, and which reporting practices limit reproducibility and cross-study comparison. A conceptual flow chart is shown in Figure 1.

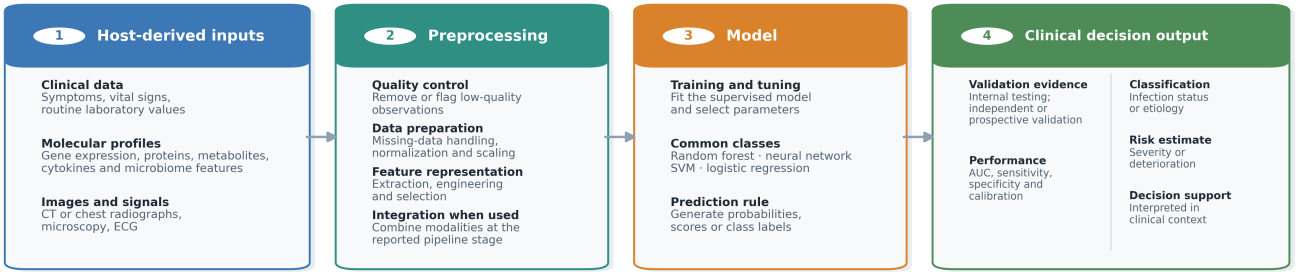


Figure 1. Conceptual architecture of a host-based machine-learning diagnostic pipeline. Host-derived inputs undergo quality control, data preparation, and feature extraction. The resulting features are used to train models such as random forests, neural networks, support vector machines, and logistic regression. Model performance is assessed through internal testing and independent or prospective validation. The resulting predictions may support infection or etiologic classification, estimation of disease severity or deterioration risk, and clinical decision-making.

We focused the quantitative comparison on the four most frequently evaluated supervised model classes found in our search: neural networks, logistic regression, random forest, and support vector machines. Less frequently evaluated methods, including gradient boosting, naive Bayes, decision trees, and k-nearest neighbors, were retained in the study-level and modality-specific synthesis but were not included in the four-class quantitative comparison. The objective was not to identify one universally superior algorithm, but to describe where particular approaches were used, evaluate the consistency of performance reporting, and identify practical requirements for clinically interpretable and reproducible diagnostic models.

2 MATERIALS AND METHODS

2.1 Literature search and eligibility criteria

The literature search covered January 2020 through August 2025 and used four sources: Nature, PLOS, PubMed, and Elsevier. Each search retained a common conceptual core comprising host-based diagnosis, bacterial or viral infection, infection classification, and machine learning. Search syntax, field restrictions, subject-area filters, and article-type filters were adapted to the indexing structure and search interface of each source. During pilot searches, queries returning fewer than approximately 20 records were broadened, whereas queries returning more than approximately 100 records were refined with additional diagnostic or host-response terms. Subject-area and article-type restrictions were applied to PLOS and Elsevier because broader searches produced large numbers of records unrelated to clinical host-response diagnosis. Equivalent restrictions were not required for Nature and PubMed to obtain manageable result sets. These adaptations were intended to maximize retrieval of potentially relevant articles while limiting excessive off-topic results and screening burden, rather than to impose different eligibility standards across sources. The same inclusion and exclusion criteria were applied to all retrieved records. Complete source-specific search strings and filters are provided in Table 1. The absence of a prospectively registered protocol and the use of source-adapted search strategies are acknowledged as limitations of the review.

Table 1. Search sources and concepts used for article retrieval. Search strings were adapted to the indexing and result volume of each source.

Source	Search concepts and filters
PLOS	Viral infection; bacterial infection; host response; host-based diagnostics; infection classification; rapid diagnostics; clinical decision support. Subject areas included diagnostic medicine, infectious diseases, and medicine and health sciences. Research and review articles were considered.
Nature	Host-based diagnosis; bacteria; virus; infection; machine learning; deep learning; human.
PubMed	Host-based diagnosis; bacteria; virus; infection; machine learning.
Elsevier	Viral infection; bacterial infection; host response; host-based diagnostics; infection classification; rapid diagnostics; clinical decision support. Research and review articles in immunology and microbiology were considered.

Eligible studies used clinical patient samples or patient-derived data and applied machine-learning methods to pathogen detection, bacterial-versus-viral classification, infection-state classification, or a closely related host-response diagnostic objective. Studies focused solely on in vitro systems, non-host-based measurements without a clinical host component, conventional statistical analysis without machine learning, or prediction of infection risk without diagnostic classification were excluded. Eligible targets included bacterial and viral infections, bacterial-versus-viral discrimination, and syndrome-based presentations such as sepsis. The search was designed primarily to identify host-based diagnostics for bacterial and viral infections, including bacterial-versus-viral discrimination and syndrome-based presentations such as sepsis. Malaria was represented only within a mixed febrile cohort. Fungal infections, parasitic infections other than malaria, healthcare-associated infections, and opportunistic infections were not systematically targeted. The conclusions should not be generalized to these categories. One tuberculous-meningitis cohort involving people with HIV provided limited evidence relevant to opportunistic infection.

Across the included articles, evaluation was limited by inconsistent reporting of data preprocessing, feature selection, hyperparameter optimization, validation procedures, and class-specific performance. Accuracy and area under the receiver operating characteristic curve were the most frequently reported measures, whereas sensitivity, specificity, precision, F1 score, Matthews correlation coefficient, calibration, and other clinically informative measures were reported less consistently. Differences in patient populations, diagnostic outcomes, data modalities, and validation designs further limited direct comparisons among models. Consequently, model-performance values were interpreted within the context of each individual study rather than treated as directly comparable estimates of diagnostic effectiveness. These limitations were incorporated into the narrative synthesis and should be considered when interpreting the performance of individual machine-learning approaches.

2.2 Study screening and selection

The search returned 144 records: 21 from Nature, 43 from PLOS, 27 from PubMed, and 53 from Elsevier. The first screen retained 51 records with a host-based diagnostic focus and excluded 93 records that were off-topic or not host-based. The second screen excluded 30 records without a suitable machine-learning diagnostic model, leaving 21 included articles. The complete workflow is shown in Figure 2.

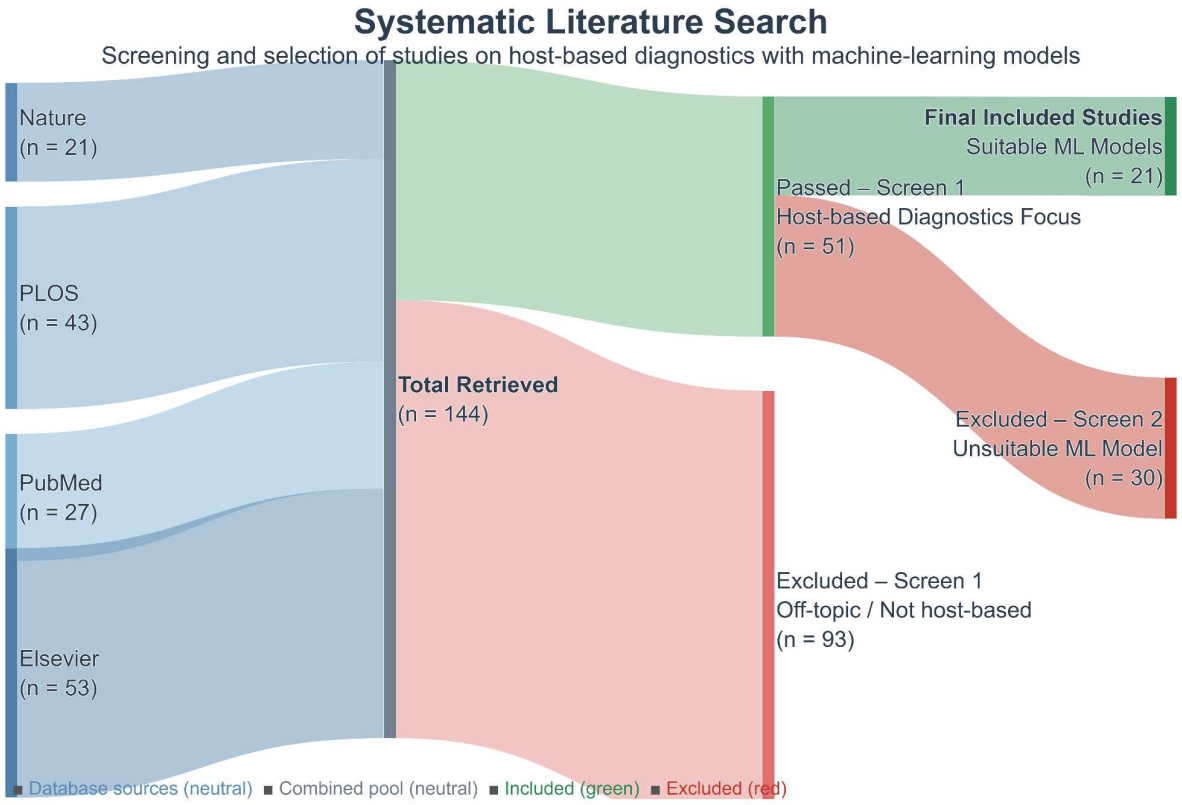


Figure 2. Literature search and study-selection workflow. Four sources yielded 144 records. Screen 1 retained 51 records focused on host-based diagnostics and excluded 93 records that were off-topic or not host-based. Screen 2 retained 21 articles containing suitable machine-learning diagnostic models and excluded 30 records with unsuitable model objectives or implementations.

2.3 Data extraction and synthesis

For each included article, we extracted the primary diagnostic objective, clinical population, input modality, machine-learning model class, reported validation approach, and performance metrics. Data modalities were grouped into structured clinical data, gene or molecular-expression data, and imaging or physiological-signal data. For quantitative model-class counting, evaluations were restricted to four machine-learning classes: random forest, neural network, support vector machine, and logistic regression. Multiple implementations of these classes within the same article were counted separately, resulting in 54 focal-class evaluations across 21 articles. Less frequently evaluated methods were extracted and described in the study-level and modality-specific synthesis but were excluded from the quantitative comparisons.

Since datasets, outcomes, validation procedures, and performance measures were heterogeneous, the evidence was synthesized descriptively rather than combined in a quantitative meta-analysis. Accuracy and area under the receiver operating characteristic curve were coded as the principal performance categories. Sensitivity, specificity, precision, F1 score, Matthews correlation coefficient, and related measures were coded as additional metrics. Where an article compared several models, the authors' identified optimal model was recorded when the selection criterion was sufficiently described.

3 RESULTS

3.1 Included studies and diagnostic data inputs

The included studies represented a broad range of clinical contexts, including bacterial-versus-viral fever, dengue and malaria, COVID-19, influenza, tuberculosis, sepsis, and mixed infectious-disease cohorts. Structured clinical variables were used for fever and infection classification by Tran et al. and Vu et al., whereas Mor et al. evaluated a three-protein host-response score for bacterial-versus-viral classification. Imaging and signal-based analyses included computed tomography, electrocardiographic records, microscopy, and medical imagery (Topff et al., 2023; Nguyen et al., 2022; Petkidis et al., 2024; Aboshosha, 2025). Molecular and host-response inputs included transcriptomics, metagenomics, cytokines, metabolites, immune biomarkers, and microbiome measurements (Ng et al., 2021; Bai et al., 2025; Vargas et al., 2023; Myall et al., 2021; Li et al., 2023; Ramachandran et al., 2022; Ekpenyong et al., 2021; Meserve et al., 2025; Hogan et al., 2021; Malic et al., 2025; Elgedawy et al., 2024; Martinez et al., 2023; Su et al., 2024; Fabbrini et al., 2024).

Many articles combined data types or compared several model classes. Within the four focal classes, 54 model evaluations were counted: 23 random-forest models, 17 neural-network models, 9 support-vector-machine models, and 5 logistic-regression models (Table 2). These totals represent evaluated model instances rather than mutually exclusive articles, because one article could contribute more than one model. The interpretability, data requirements, scalability, strengths, limitations, and descriptive performance ranges of the four focal classes are compared in Table 3. Less frequently evaluated approaches and detailed study characteristics are summarized in Table 4. The study-level pairing of input modalities with primary and comparator model families is shown in Figure 3. Patient-level covariates were handled inconsistently across the included studies. Across the nine demographic and clinical domains extracted for this review, 49.2% of study-domain entries were not reported, while 23.8% were reported only descriptively. Age was included as a diagnostic-model input in one study, whereas no study clearly incorporated sex, comorbidities, ongoing pharmacological treatment, or immune status as predictive features. When considered, these characteristics were used mainly for eligibility, matching, adjustment, stratification, or limited subgroup analyses. No study reported fairness or interaction analyses. Reported performance may therefore reflect cohort composition or unmeasured confounding, limiting generalization across demographically and clinically distinct patient populations.

Table 2. Evaluations of the four focal machine-learning model classes across the included articles. Counts represent evaluated model instances, and individual articles could contribute to multiple categories. Less frequently evaluated methods were excluded from the 54-evaluation denominator and are summarized in Table 4.

Model class	Number of evaluations	Percentage of 54 focal-class evaluations
Random forest	23	42.6%
Neural network	17	31.5%
Support vector machine	9	16.7%
Logistic regression	5	9.3%

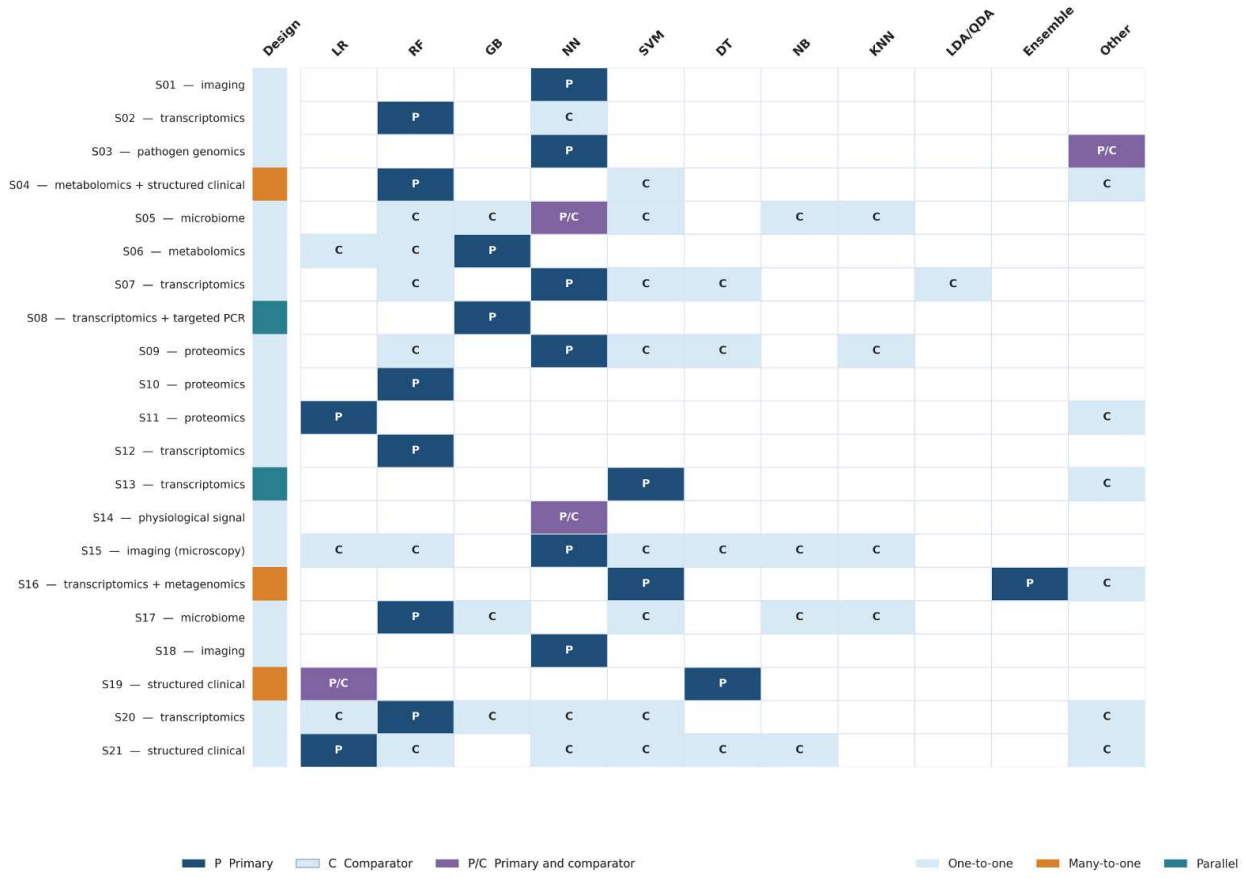


Figure 3. Study-level mapping of diagnostic input modalities to machine-learning model families across the 21 included studies. Rows represent individual studies and their input modalities, while columns represent the evaluated model families. Within the matrix, P denotes a primary model, C denotes a comparator model, and P/C indicates that a model family was evaluated in both roles. The color band on the left identifies the input-model integration design as one-to-one, many-to-one, or parallel. The mapping demonstrates that model selection varied by input modality and that several studies evaluated multiple model families or combined multiple data types.

3.2 Molecular and sequencing-derived data

Numeric biological and clinical inputs represent each patient or specimen as a combination of measured features. Structured inputs include demographics, symptoms, physical-examination findings, vital signs, and routine laboratory values, whereas molecular inputs include transcript counts, proteins, cytokines, metabolites, and microbiome abundance. Although sequencing remains more resource-intensive than targeted assays, streamlined library preparation, host-DNA depletion, real-time long-read nanopore sequencing, and faster informatics pipelines have reduced response times to several hours and enabled same-day pathogen identification (Gu W., 2021; Cheng H., 2022; Alcolea-Medina A., 2024). Targeted PCR and immunoassay panels retain a smaller, predefined feature space and may therefore be easier to implement in rapid clinical workflows.

Structured clinical variables were evaluated using multivariable logistic regression and classification and regression trees for scrub typhus versus dengue (Tran et al., 2022). Boosted logistic regression, random forest, support vector machine, multilayer perceptron, naive Bayes, and decision tree models were compared for dengue and malaria classification (Vu et al., 2023). Elgedawy et al. combined routine blood

measurements with targeted metabolites using random forest as the primary classifier and support vector machine and partial least-squares methods as comparators. Host-transcriptomic studies used random forests for infection-state and tuberculosis classification (Bai et al., 2025; Myall et al., 2021; Vargas et al., 2023), a neural network for multiclass infection classification (Li et al., 2023), extreme gradient boosting for a sepsis signature transferred to targeted PCR platforms (Malic et al., 2025), and support vector machines for COVID-19 and tuberculous meningitis (Ng et al., 2021; Ramachandran et al., 2022). Protein, cytokine, and metabolite panels were modeled using random forest (Meserve et al., 2025), logistic regression (Mor et al., 2023), neural networks with conventional comparators (Martinez et al., 2023), and LightGBM (Hogan et al., 2021). Microbiome profiles were evaluated using a convolutional neural network in Fabbrini et al. and a random forest in Su et al. Ekpenyong et al. applied a neural network to SARS-CoV-2 genomic features, although some of the data represented pathogens rather than host-response measurements.

Preprocessing practices varied substantially. Feature-selection methods were reported in 16 of the 21 studies and included differential-expression or significance filtering, LASSO and other penalized regression methods, random-forest importance, Boruta package, information gain, SHAP ranking, backward elimination, and genetic algorithms. Normalization was modality-specific and included median normalization and log transformation for metabolomics, housekeeping-gene normalization for targeted PCR, within-sample gene-pair encodings for transcriptomics, trimmed mean of M-values followed by log-counts per million for RNA sequencing, and centered log-ratio transformation for microbiome data. Myall et al. used the ComBat package for batch correction, whereas Li et al. and Vargas et al. used within-sample gene relationships intended to reduce platform dependence. Hogan et al. assessed technical variation using pooled quality-control samples and blank subtraction, while Ramachandran et al. distributed outcome groups across sequencing batches. Formal batch-effect assessment was absent or unreported in 13 studies, and normalization or scaling was insufficiently described by several studies, including Aboshosha, Martinez et al., Ng et al., Tran et al., and Vu et al. Feature selection or importance ranking was performed on the complete dataset or outside the validation loop in six studies (Bai et al., 2025; Elgedawy et al., 2024; Fabbrini et al., 2024; Meserve et al., 2025; Myall et al., 2021; Tran et al., 2022), creating a risk of optimistic performance estimates. Eight studies used independent external cohorts, three provided partial cross-platform or non-patient validation, and the remainder relied on internal evaluation or did not report external validation. These inconsistencies limit reproducibility and confidence in clinical generalization.

Metagenomic next-generation sequencing (mNGS) can also provide direct pathogen identification rather than serving only as a source of microbiome features. The study by Ramachandran et al. was the only included study to integrate mNGS with host-response data. The investigators trained support vector machines using cerebrospinal-fluid host transcripts and combined their predictions with threshold-based cerebrospinal-fluid mNGS. Host reads were removed, no-template controls were processed, and organism-specific sequence-identity and abundance thresholds were applied relative to control samples and cohort background. Candidate pathogens could also be confirmed through targeted PCR, clinical microbiology, or repeat mNGS. When the two modalities disagreed, a qualifying mNGS pathogen call overrode the host classifier. Thus, the host response provided partial evidence of infection, whereas discrimination of true pathogens depended on the mNGS control and threshold pipeline rather than being learned by the classifier. Concordance between the two modalities was only 50–61%, and the override rule was not optimized, illustrating the need for validated contamination controls and prespecified procedures for resolving discordant results.

3.3 Structured clinical data and model performance

Only three studies incorporated structured clinical variables or routine laboratory measurements as model inputs, including one study that combined these variables with metabolomic data. Across these studies, logistic regression was the most frequently evaluated model class, with five pipelines, followed by decision trees with three pipelines, random forests and support vector machines with two each, and neural-network and naive Bayes models with one each. Within the same test cohort, Vu et al. reported similar dengue-classification accuracy across the six evaluated algorithms (0.91–0.92). Five models had 0% sensitivity, whereas boosted logistic regression achieved a sensitivity of only 5.5%, demonstrating that accuracy obscured poor case detection. Malaria-classification accuracy in the same study ranged from 0.53 to 0.69. Tran et al. reported logistic-regression AUCs ranging from 0.831 to 0.993 across different clinical and laboratory feature sets, while Elgedawy et al. reported random-forest AUCs of 0.88–1.00 across severity contrasts using combined metabolomic and routine laboratory inputs. All of these comparisons relied on internal evaluation. These findings cannot establish whether structured clinical models outperform transcriptomic or imaging models because the studies differed in disease target, outcome definition, class balance, and validation design. Instead, the evidence indicates a modality-dependent pattern that structured clinical studies favored comparatively interpretable models, transcriptomic studies frequently used random forests or support vector machines, and imaging and physiological-signal studies primarily used neural networks.

3.4 Imaging and physiological-signal data

Visual and signal-based inputs included computed tomography, electrocardiographic traces, and microscopy images. Images are represented computationally as arrays of pixel intensities, while physiological signals are represented as ordered measurements over time. Neural networks were frequently selected for these applications because they can learn spatial or temporal features directly from high-dimensional inputs. Their performance, however, depends on the amount and diversity of training data, annotation quality, acquisition protocols, and safeguards against institution-specific artifacts.

Imaging-based diagnostics are also constrained by clinical availability. A model can be useful only when the relevant organ system is imaged or the signal is collected. Microscopy-based methods may require culture or sample preparation, which can delay results. These methods may therefore serve as specialized diagnostic aids rather than universal host-based screening tools.

3.5 Multimodal inputs and model integration

The included studies should be distinguished between those that compared multiple algorithms using the same input modality and those that incorporated more than one data or assay type into a diagnostic pipeline. Most studies used a single modality. Three studies used multiple data or assay categories, although their approaches represented fdifferent approaches rather than a common multimodal design.

Elgedawy et al. provided the clearest example of blending multiple data inputs. The investigators measured targeted metabolites and 17 routine blood parameters in the same participants and compared models using metabolites alone with models receiving both types of features. The combined models produced a single prediction from a unified feature matrix rather than separate predictions for each modality. For discrimination between severe and critical COVID-19, the random-forest model retained an accuracy of 0.92 after the blood parameters were added, while the reported AUC increased from 0.88 for the metabolite-only model to 0.97 for the combined model. An SVM was also trained using selected metabolites and the blood parameters. Since the inputs were combined before model fitting, the study did

not report agreement between independent metabolomic and clinical predictions or specify a procedure for resolving discordant outputs.

Ramachandran et al. used late, decision-level fusion to combine host and pathogen information for tuberculous meningitis. A host gene-expression classifier and cerebrospinal-fluid mNGS were evaluated separately before their outputs were combined. When the host classifier predicted tuberculous meningitis but mNGS identified a non-tuberculous pathogen, the metagenomic result overruled the host-classifier output. The investigators explicitly evaluated agreement between the two components, reporting concordance of 50% (11/22) against the uniform case definition for definite and probable tuberculous meningitis. Concordance increased to 61% (11/18) after probable cases in which mNGS detected a non-tuberculous pathogen were excluded. The host classifier alone produced an AUC of 0.74 in the blinded test cohort, compared with 0.87 for the combined assay.

Malic et al. represented cross-platform translation rather than fusion of independent clinical modalities. The investigators derived a six-gene expression signature from RNA-sequencing data, selected XGBoost from 18 candidate algorithms using repeated cross-validation and implemented the same signature using RT-qPCR, digital droplet PCR, and the PREDICT microfluidic platform. These platforms measured the same underlying gene-expression signature and generated the same type of deterioration-risk output. Cross-validation was used for model selection and performance estimation. It did not combine predictions from separate modalities. Performance was reported across the transcriptomic datasets and targeted assay implementations, but the study did not generate concurrent modality-specific predictions requiring a cross-modal agreement statistic or discordance rule.

The reference standard was applied at the patient level within each study, although the target outcome differed among studies. Elgedawy et al. used PCR-confirmed infection and WHO severity categories, Malic et al. used subsequent worsening of the SOFA score or intensive-care admission, and Ramachandran et al. used the consensus uniform case definition for tuberculous meningitis. Use of a shared reference label allowed each component to be evaluated against the same outcome, but it did not by itself demonstrate consistency between modality-specific predictions.

No included study simultaneously combined structured clinical data, host gene expression, and imaging within one pipeline. No study reported a learned weighted or stacked ensemble of independent modality-specific predictions. Cross-validation was used to estimate performance or select models rather than to reconcile disagreement between modalities. Future multimodal diagnostic studies should specify the stage at which data are combined, report the performance of each modality separately and in combination, quantify agreement between modality-specific outputs, and define how discordant predictions are converted into a final clinical result.

3.6 Characteristics of the four selected machine learning classes

The four focal comparative categories were supervised learning methods that require labeled training data and an independent or resampled validation process (Kotwal et al., 2022). Logistic regression estimates the probability of a binary or multinomial outcome and provides coefficients that are comparatively easy to interpret. Its assumptions and linear decision structure can be limiting when relationships are strongly nonlinear, but regularization can make it useful for high-dimensional biomarker panels.

Support vector machines construct decision boundaries that maximize separation between classes. Kernel functions can extend the approach to nonlinear data, and multiclass problems can be addressed through

multiple decision boundaries. Performance depends on feature scaling, kernel selection, tuning parameters, and the handling of imbalanced classes.

Random forests aggregate predictions from multiple decision trees and can model nonlinear relationships and interactions with limited distributional assumptions. They were the most frequently evaluated model family in the reviewed articles. Random-forest feature-importance measures can assist biomarker prioritization, although importance estimates can be biased by correlated predictors, feature scale, or the selected importance method (Speiser et al., 2019).

Neural networks consist of interconnected layers that learn complex representations. Convolutional architectures are particularly suited to images and structured signals, while multilayer perceptrons can model clinical or molecular features. Their flexibility can improve predictive performance, but extensive tuning, larger training sets, calibration assessment, and interpretability analyses are often needed before clinical deployment.

3.7 Performance reporting

Reported performance varied substantially, and the available evidence did not identify a universally optimal model class. Differences in input modality, sample size, class prevalence, preprocessing, feature selection, hyperparameters, train-test splitting, cross-validation, and external validation limited direct comparisons. Model architecture and evaluation procedures were not always described at a level sufficient for reproduction.

Accuracy and AUC were the dominant metrics across the 54 evaluations of the four focal model classes. Accuracy alone was reported for 14 evaluations (26%), AUC alone for 20 (37%), and both for 20 (37%). Sixteen evaluations (30%) reported one or more additional metrics, including sensitivity, specificity, F1 score, or Matthews correlation coefficient. These additional metrics formed a cross-cutting subset rather than a separate mutually exclusive group (Figure 4).

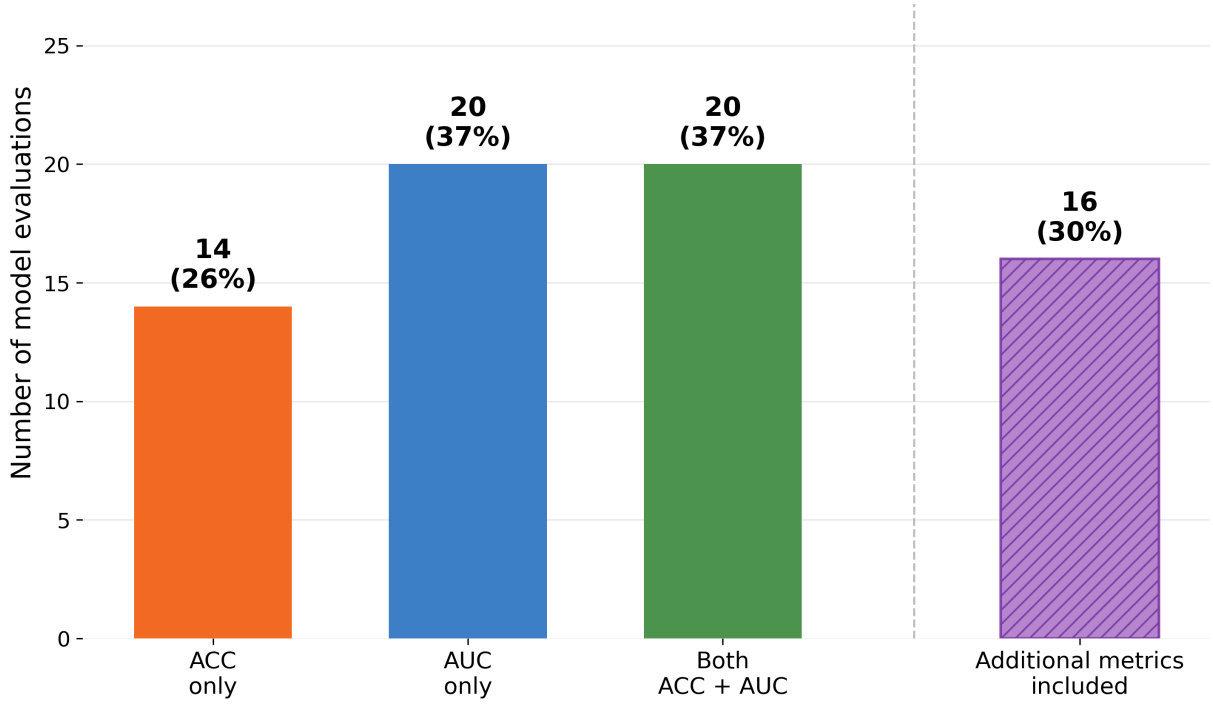


Figure 4. Performance metrics reported across the 54 evaluations of the four focal model classes. Accuracy (ACC) alone was reported for 14 model evaluations (26%), area under the receiver operating characteristic curve (AUC) alone for 20 (37%), and both ACC and AUC for 20 (37%). Sixteen evaluations (30%) also reported supplementary measures, including sensitivity, specificity, F1 score, Matthews correlation coefficient, or related metrics. The supplementary-metric group is a cross-cutting subset distributed across the three primary categories.

Accuracy is the proportion of all observations classified correctly, but it can be misleading when classes are imbalanced or prevalence differs across cohorts. A model that predicts the majority class can achieve high accuracy while failing to identify the clinically important minority class. AUC summarizes ranking performance across classification thresholds and is generally less dependent on one selected operating point, but it does not by itself communicate sensitivity, specificity, predictive values, calibration, or clinical utility. Cross-study interpretation is therefore strongest when AUC is reported alongside class-specific measures and confidence intervals.

4 DISCUSSION

4.1 Experimental considerations

Model comparisons must account for the properties of the underlying data. Transcriptomic, targeted molecular, clinical, and imaging inputs differ in measurement error, preprocessing requirements, turnaround time, cost, and clinical feasibility, limiting direct comparisons across modalities. The four focal model classes differed in their suitability for specific data types and clinical applications (Table 3). Random forests were applied broadly to molecular and structured data, whereas neural networks predominated in imaging and physiological-signal analyses. Support vector machines were suited to high-dimensional, small-sample datasets, while logistic regression offered the greatest interpretability for structured clinical variables and compact biomarker panels. Because validation designs and diagnostic targets varied, the reported performance ranges do not establish the superiority of any single model class.

Broad discovery platforms are valuable for identifying candidate biomarkers, whereas targeted assays may be more practical for rapid diagnostics. A useful translational strategy is therefore established. We must discover robust features in high-dimensional data, validate them across independent cohorts and platforms, and then reduce the feature set to a reproducible panel that can be measured within a clinically relevant timeframe. Feature stability across populations is as important as model accuracy within the development cohort.

Study design also affects the apparent value of a model. Random train-test splits may allow samples from the same site, processing batch, or patient population to appear in both sets, inflating performance. External validation by geography, institution, age group, disease prevalence, and acquisition platform provides a stronger test of generalization. Prospective evaluation is ultimately required to determine whether predictions improve decisions and outcomes when integrated into clinical workflows.

4.2 Computational and reporting considerations

Transparent reporting should include the complete preprocessing workflow, missing-data handling, feature selection, model architecture, hyperparameter search, class weighting, data-partition strategy, and performance uncertainty. When multiple models are tested, authors should report the prespecified criterion used to select the preferred model and provide results for relevant comparators. Reporting only the best result obscures the stability of the conclusion and increases the risk of optimistic selection.

Performance reporting should not rely on accuracy alone. At minimum, classification studies should report AUC or an appropriate multiclass analogue, sensitivity, specificity, precision or positive predictive value, negative predictive value when clinically meaningful, and confidence intervals. Confusion matrices and calibration assessments are also valuable. Decision-curve analysis or another clinical-utility framework can help determine whether a model improves decisions at realistic thresholds rather than merely separating groups in a retrospective dataset.

Interpretability is especially important when models are intended to influence medical decisions. Feature-importance analyses can identify host-response markers that drive predictions and can support the development of smaller assay panels. These analyses should be evaluated for stability across resampling, cohorts, and correlated feature sets. Mechanistic plausibility can strengthen confidence, but biological interpretation should not be inferred solely from a predictive importance score.

4.3 Clinical translation and standardization

A clinically useful host-based diagnostic system should define the intended population, specimen type, turnaround time, decision point, and consequence of false-positive and false-negative predictions. A model optimized to distinguish one pathogen from healthy controls may not perform adequately when confronted with other infections, coinfections, chronic inflammatory disease, immunosuppression, or different stages of illness. Disease-agnostic outputs, such as bacterial, viral, mixed, or noninfectious categories, may be more broadly useful than pathogen-specific classifiers when the presenting diagnosis is unknown.

Standardization does not require that every study use the same algorithm. Instead, it requires sufficient methodological and reporting consistency to determine whether a model is valid for its intended use. Common data dictionaries, reproducible preprocessing, prespecified evaluation plans, external validation, and standardized metric panels would substantially improve comparability. Publicly available code and de-identified data, where ethically and legally possible, would further support independent validation.

4.4 Limitations

This review was limited to four literature sources and publications from January 2020 through August 2025. Although all searches retained common concepts related to host-based diagnosis, bacterial or viral infection, and machine learning, source-specific syntax and filters may have altered retrieval sensitivity and favored particular databases or study types. Relevant studies may therefore have been missed, particularly those involving underrepresented pathogen classes or clinical settings. Differences in diagnostic objectives, populations, modalities, reference standards, and validation designs also precluded pooled performance estimates.

The protocol was not prospectively registered because the project began as an exploratory mapping exercise and was later developed into a structured evidence synthesis. Eligibility criteria, extraction fields, and analytical procedures were therefore not fixed in a public protocol before searching began, creating a risk of retrospective methodological decisions. Final search strings and filters are provided in Table 1, but a complete version-controlled history of pilot searches, query modifications, database exports, and screening decisions was not maintained. Consequently, the final searches can be repeated, but the exact iterative process that produced the original record set cannot be fully reconstructed, limiting replicability and preventing the effect of search heterogeneity on retrieval completeness from being quantified.

A formal critical-appraisal procedure and certainty-of-evidence framework were not utilized. We did not apply a formal risk-of-bias instrument such as PROBAST+AI for prediction-model studies or QUADAS-2 for diagnostic-accuracy studies, and we did not apply a certainty-of-evidence framework. To reduce the resulting risk of bias and selective reporting, we used standardized search fields. Missing information was not inferred, and performance results were interpreted within their original task, cohort, metric, and validation context. This process considered participant selection, input and outcome measurement, analytical methods, missing data, class imbalance, feature-selection leakage, training-test separation, external validation, selective metric reporting, calibration, and clinical applicability. Since the appraisal was retrospective and did not use prespecified exclusion or weighting rules, it should be interpreted as descriptive rather than as a formal risk-of-bias determination. The findings therefore represent a structured mapping of the available evidence rather than a definitive comparison of diagnostic effectiveness.

5 CONCLUSION

Machine learning applied to host-based diagnostic data has clear potential to support the detection and classification of bacterial and viral infections, but the current evidence is methodologically heterogeneous. No model class was uniformly superior across clinical variables, molecular measurements, images, and physiological signals. Model selection should be guided by the structure and practical constraints of the input data, and performance should be confirmed through independent validation.

The most immediate opportunity for the field is improved standardization. Consistent reporting of preprocessing, validation, hyperparameter selection, AUC, class-specific performance, calibration, and feature importance would make studies more reproducible and clinically interpretable. Future work should prioritize generalizable, disease-agnostic outputs and targeted biomarker panels that can be measured within clinically useful time frames. With these improvements, host-based machine-learning systems may become valuable decision-support tools rather than isolated retrospective demonstrations.

CONFLICT OF INTEREST STATEMENT

No conflicts of interest exist.

AUTHOR CONTRIBUTIONS

SH: Data Curation, Writing - original draft, Writing - review & editing, Methodology, Formal Analysis, Visualization SN: Data Curation, Writing - original draft. KM: Supervision, Conceptualization, Writing - review & editing JKS: Supervision, Funding acquisition

FUNDING

This work was supported by the U.S. Department of Energy through the Los Alamos National Laboratory. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001) and at the Lawrence Berkeley National Laboratory, operated by the University of California for the United States Department of Energy. Funding for this work was provided by the Joint Science and Technology Office, Defense Threat Reduction Agency (CB11015). The views expressed in this article are those of the authors and do not reflect the official policy or position of the U.S. Department of Defense or the U.S. Government.

SUPPLEMENTAL DATA

Supporting Information is uploaded separately.

DATA AVAILABILITY STATEMENT

The data analyzed in this study were obtained from previously published articles and are available through the original publications cited in this manuscript. No new primary datasets were generated as part of this review. Data extracted from the included studies and used for the descriptive synthesis are provided within the manuscript and its Supporting Information. Further inquiries regarding the data summarized in this review can be directed to the corresponding author.

REFERENCES

Aboshosha, A. (2025). AI based medical imagery diagnosis for COVID-19 disease examination and remedy. *Sci Rep* 15, 1607. doi:10.1038/s41598-024-84644-1

Alcolea-Medina A., L. B. S. T. C. A. A. G. N. T. W. S. G. S. D. R. B. P. R. C. H. M. S. N. M. W. . J. D. E., Christopher Alder (2024). Unified metagenomic method for rapid detection of microorganisms in clinical samples. *Communications Medicine* 4, 135. doi:10.1038/s43856-024-00554-3

Bai, F., Gong, Z., Cui, D., Zhang, X., Hong, W., Gao, Y., et al. (2025). Development of a host-signature-based machine learning model to diagnose bacterial and viral infections in febrile children. *Front Pediatr* 13, 1608812. doi:10.3389/fped.2025.1608812

[Dataset] CDC (2024). Antimicrobial resistance in health care: Causes and how it spreads. Antimicrobial Resistance Accessed 2026-02-12.

Cheng H., Q. Y. M. D. Z. Y. G. Z. J. Q. L. L. L. Y. Y. X., Yuhong Sun (2022). A rapid bacterial pathogen and antimicrobial resistance diagnosis workflow using oxford nanopore adaptive sequencing method. *Briefings in Bioinformatics* 23, bbac453. doi:10.1093/bib/bbac453

[Dataset] Cleveland Clinic (n.d.a). Complete blood count (CBC) test. Cleveland Clinic Accessed 2026-02-12.

[Dataset] Cleveland Clinic (n.d.b). PCR test: What it is, how it works & results. Accessed 2026-02-12.

Ekpenyong, M. E., Edoho, M. E., Inyang, U. G., Uzoka, F.-M., Ekaidem, I. S., Moses, A. E., et al. (2021). A hybrid computational framework for intelligent inter-continent SARS-CoV-2 sub-strains characterization and prediction. *Sci Rep* 11, 14558. doi:10.1038/s41598-021-93757-w

Elgedawy, G. A., Samir, M., Elabd, N. S., Elsaid, H. H., Enar, M., Salem, R. H., et al. (2024). Metabolic profiling during COVID-19 infection in humans: Identification of potential biomarkers for occurrence, severity and outcomes using machine learning. *PLoS One* 19, e0302977. doi:10.1371/journal.pone.0302977

Fabbrini, M., D'Amico, F., van der Gun, B. T. F., Barone, M., Conti, G., Roggiani, S., et al. (2024). The gut microbiota as an early predictor of COVID-19 severity. *mSphere* 9, e0018124. doi:10.1128/msphere.00181-24

Fleming-Dutra, K. E., Hersh, A. L., Shapiro, D. J., Bartoces, M., Enns, E. A., File, T. M., et al. (2016). Prevalence of Inappropriate Antibiotic Prescriptions Among US Ambulatory Care Visits, 2010-2011. *JAMA* 315, 1864–1873. doi:10.1001/jama.2016.4151

Gu W., M. L. Y. D. S. S. A. D. S. S. F. A. G. K. R. K. Z. H. S. G. Y. G. I. B. B. E. D. C. A. B. M. R. W. C. W. E. H. S. M. J. L. D. . C. Y. C., Xianding Deng (2021). Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nature Medicine* 27, 115–124. doi:10.1038/s41591-020-1105-z

Hogan, C. A., Rajpurkar, P., Sowrirajan, H., Phillips, N. A., Le, A. T., Wu, M., et al. (2021). Nasopharyngeal metabolomics and machine learning approach for the diagnosis of influenza. *EBioMedicine* 71, 103546. doi:10.1016/j.ebiom.2021.103546

[Dataset] IBM (n.d.). What is machine learning? | IBM. Accessed 2026-02-12.

Kotwal, S., Rani, P., Arif, T., Manhas, J., and Sharma, S. (2022). Automated bacterial classifications using machine learning based computational techniques: Architectures, challenges and open research issues. *Archives of Computational Methods in Engineering* 29, 2469–2490. doi:10.1007/s11831-021-09660-0

Largman-Chalamish, M., Wasserman, A., Silberman, A., Levinson, T., Ritter, O., Berliner, S., et al. (2022). Differentiating between bacterial and viral infections by estimated CRP velocity. *PLOS ONE* 17, e0277401. doi:10.1371/journal.pone.0277401

Li, Q., Zheng, X., Xie, J., Wang, R., Li, M., Wong, M.-H., et al. (2023). bvnGPS: a generalizable diagnostic model for acute bacterial and viral infection using integrative host transcriptomics and pretrained neural networks. *Bioinformatics* 39, btad109. doi:10.1093/bioinformatics/btad109

Malic, L., Zhang, P. G. Y., Plant, P. J., Clime, L., Nassif, C., Da Fonte, D., et al. (2025). A machine learning and centrifugal microfluidics platform for bedside prediction of sepsis. *Nat Commun* 16, 4442. doi:10.1038/s41467-025-59227-x

Martinez, G. S., Ostadgavahi, A. T., Al-Rafat, A. M., Garduno, A., Cusack, R., Bermejo-Martin, J. F., et al. (2023). Model-interpreted outcomes of artificial neural networks classifying immune biomarkers associated with severe infections in ICU. *Front Immunol* 14, 1137850. doi:10.3389/fimmu.2023.1137850

[Dataset] MedlinePlus (n.d.). Bacteria culture test: MedlinePlus medical test. Accessed 2026-02-12.

Meserve, K., Chapman, C. A., Xu, M., Zhou, H., Robison, H. M., Hilgart, H. R., et al. (2025). Multiplexed cytokine profiling identifies diagnostic signatures for latent tuberculosis and reactivation risk stratification. *PLoS One* 20, e0316648. doi:10.1371/journal.pone.0316648

Mor, M., Paz, M., Amir, L., Levy, I., Scheuerman, O., Livni, G., et al. (2023). Bacterial vs viral etiology of fever: A prospective study of a host score for supporting etiologic accuracy of emergency department physicians. *PLoS One* 18, e0281018. doi:10.1371/journal.pone.0281018

Myall, A. C., Perkins, S., Rushton, D., David, J., Spencer, P., Jones, A. R., et al. (2021). An OMICs-based meta-analysis to support infection state stratification. *Bioinformatics* 37, 2347–2355. doi:10.1093/bioinformatics/btab089

Ng, D. L., Granados, A. C., Santos, Y. A., Servellita, V., Goldgof, G. M., Meydan, C., et al. (2021). A diagnostic host response biosignature for COVID-19 from RNA profiling of nasal swabs and blood. *Sci Adv* 7, eabe5984. doi:10.1126/sciadv.abe5984

Nguyen, T., Pham, H. H., Le, K. H., Nguyen, A.-T., Thanh, T., and Do, C. (2022). Detecting COVID-19 from digitized ECG printouts using 1D convolutional neural networks. *PLoS One* 17, e0277081. doi:10.1371/journal.pone.0277081

Paudel, R., Dogra, P., Montgomery-Yates, A. A., and Coz Yataco, A. (2020). Procalcitonin: A promising tool or just another overhyped test? *International Journal of Medical Sciences* 17, 332–337. doi:10.7150/ijms.39367

Petkidis, A., Andriasyan, V., Murer, L., Volle, R., and Greber, U. F. (2024). A versatile automated pipeline for quantifying virus infectivity by label-free light microscopy and artificial intelligence. *Nat Commun* 15, 5112. doi:10.1038/s41467-024-49444-1

Ramachandran, P. S., Ramesh, A., Creswell, F. V., Wapniarski, A., Narendra, R., Quinn, C. M., et al. (2022). Integrating central nervous system metagenomics and host response for diagnosis of tuberculosis meningitis and its mimics. *Nat Commun* 13, 1675. doi:10.1038/s41467-022-29353-x

Speiser, J. L., Miller, M. E., Tooze, J., and Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications* 134, 93–101. doi:10.1016/j.eswa.2019.05.028

Su, Q., Lau, R. I., Liu, Q., Li, M. K. T., Yan Mak, J. W., Lu, W., et al. (2024). The gut microbiome associates with phenotypic manifestations of post-acute COVID-19 syndrome. *Cell Host Microbe* 32, 651–660.e4. doi:10.1016/j.chom.2024.04.005

Topff, L., Sánchez-García, J., López-González, R., Pastor, A. J., Visser, J. J., Huisman, M., et al. (2023). A deep learning-based application for COVID-19 diagnosis on CT: The Imaging COVID-19 AI initiative. *PLoS One* 18, e0285121. doi:10.1371/journal.pone.0285121

Tran, H. T. D., Schindler, C., Pham, T. T. T., Vien, M. Q., Do, H. M., Ngo, Q. T., et al. (2022). Simple clinical and laboratory predictors to improve empirical treatment strategies in areas of high scrub typhus and dengue endemicity, central Vietnam. *PLoS Negl Trop Dis* 16, e0010281. doi:10.1371/journal.pntd.0010281

Vargas, R., Abbott, L., Bower, D., Frahm, N., Shaffer, M., and Yu, W.-H. (2023). Gene signature discovery and systematic validation across diverse clinical cohorts for TB prognosis and response to treatment. *PLoS Comput Biol* 19, e1010770. doi:10.1371/journal.pcbi.1010770

Vu, D. M., Krystosik, A. R., Ndenga, B. A., Mutuku, F. M., Ripp, K., Liu, E., et al. (2023). Detection of acute dengue virus infection, with and without concurrent malaria infection, in a cohort of febrile children in kenya, 2014–2019, by clinicians or machine learning algorithms. *PLOS Global Public Health* 3, e0001950. doi:10.1371/journal.pgph.0001950

Table 3. Comparison of the four most frequently evaluated machine-learning model classes across the clinical contexts represented in the included studies. Less frequently evaluated methods are summarized in Table 4. Performance ranges are descriptive because clinical targets, input modalities, and validation methods differed across studies.

Model class	Uses in the included studies	Interpretability	Data needs and scalability	Strengths and limitations	Reported performance
Random Forest	Molecular-expression, microbiome, and structured clinical data for infection classification, disease severity, and tuberculosis risk.	Moderate. Feature importance and SHAP values can identify influential predictors, but correlated variables may bias their rankings.	Suitable for moderate sample sizes and high-dimensional inputs. Training has moderate computational cost and is readily parallelized.	**Strengths:** Captures nonlinear relationships, interactions, and mixed feature types. **Limitations:** May overfit small datasets, requires calibration, and does not extrapolate beyond the training distribution.	Internal AUCs ranged from 0.572 to 0.952; selected external-cohort AUCs ranged from approximately 0.84 to 0.93.
Neural Network	Imaging, ECG, microscopy, transcriptomic, and proteomic data for COVID-19, bacterial-versus-viral infection, and sepsis classification.	Low without post-hoc methods such as SHAP or Grad-CAM. Several studies provided no interpretability analysis.	Generally requires large labeled datasets and has the highest training burden. Transfer learning and data augmentation may reduce data requirements.	**Strengths:** Learns complex spatial, temporal, and nonlinear features directly from the input. **Limitations:** Susceptible to overfitting, dataset shift, limited transparency, and infrequent calibration assessment.	Internal AUCs ranged from approximately 0.875 to 1.00; external or cross-platform AUCs ranged from approximately 0.838 to 0.994.
Support Vector Machine	High-dimensional transcriptomic, proteomic, microbiome, and multimodal data for infection, sepsis, COVID-19, and tuberculosis classification.	Moderate for linear models because coefficients can be inspected; low for nonlinear kernels.	Effective with moderate sample sizes and more features than observations. Requires feature scaling, and nonlinear kernels become costly as sample size increases.	**Strengths:** Produces stable decision boundaries in high-dimensional, small-sample settings. **Limitations:** Sensitive to kernel choice, scaling, hyperparameter tuning, and class imbalance.	Evaluation-set AUCs ranged from approximately 0.435 to 0.991. Most evaluations used internal validation.
Logistic Regression	Structured clinical variables and compact biomarker panels for bacterial-versus-viral infection, influenza, scrub typhus, dengue, malaria, and tuberculosis.	High. Coefficients and odds ratios can be interpreted directly and translated into clinical scores.	Low computational burden and suitable for modest datasets with relatively few predictors. High-dimensional inputs require feature selection or regularization.	**Strengths:** Transparent, efficient, and readily calibrated. **Limitations:** Assumes an approximately linear relationship on the log-odds scale and may miss complex interactions.	Internal AUCs ranged from approximately 0.831 to 0.993; a prospectively validated bacterial-versus-viral model achieved an AUC of 0.96.

Abbreviations: AUC, area under the receiver operating characteristic curve; ECG, electrocardiography; SHAP, Shapley additive explanations. Performance ranges summarize heterogeneous studies and are not pooled estimates or head-to-head comparisons.

Table 4. Characteristics, input modalities, machine-learning methods, and validation designs of the 21 included studies. Model families include both primary and comparator approaches evaluated within each study.

Study	Clinical context or diagnostic target	Population and sample size	Input modality	Model families evaluated	Validation design
Aboshosha et al. (2025)	COVID-19 diagnosis, severity, and mortality	Public imaging repositories; sample size not reported	Chest radiography and CT	NN	No independent validation reported
Bai et al. (2025)	Bacterial versus viral febrile illness	Multicenter pediatric cohorts; 384 samples in the pediatric model and 1,042 in the generalized model	Transcriptomics	NN, RF	Partial external-cohort validation
Ekpenyong et al. (2021)	SARS-CoV-2 sub-strain classification	8,864 viral genomes from GISAID, representing 88 countries	Human transcriptomics and genomics	NN, other	No external validation
Elgedawy et al. (2024)	COVID-19 diagnosis, severity, and mortality	295 analyzed participants from a single Egyptian hospital	Metabolomics and structured clinical data; feature-level integration	RF, SVM, other	Internal validation only
Fabbrini et al. (2024)	COVID-19 severity prediction	315 participants from two European centers	Microbiome	GB, KNN, NB, NN, RF, SVM	External validation
Hogan et al. (2021)	Influenza diagnosis	332 participants: 236 in the discovery cohort and 96 in the validation cohort	Metabolomics	GB, LR, RF	Independent prospective validation
Li et al. (2023)	Bacterial versus viral acute infection	2,680 samples after filtering, including 2,533 discovery and 147 external-validation samples	Transcriptomics	DT, LDA/QDA, NN, RF, SVM	Independent external validation
Malic et al. (2025)	Sepsis and short-term clinical deterioration	514 discovery participants, with additional external, blinded, and prospective cohorts	Transcriptomics and targeted PCR	GB	External, blinded, and prospective validation
Martinez et al. (2023)	Sepsis and septic-shock severity	112 intensive-care patients from a single Irish center	Proteomics and immune biomarkers	DT, KNN, NN, RF, SVM	Internal validation only
Meserve et al. (2025)	Latent tuberculosis infection and risk stratification	144 patients providing 165 sample sets	Proteomic and cytokine measurements	RF	Internal validation only
Mor et al. (2023)	Bacterial versus viral febrile illness	465 pediatric emergency-department patients; 214 included in the primary analysis	Three-protein host-response score	LR, other	Prospective validation of a locked model
Myall et al. (2021)	Bacterial versus viral infection state	3,868 samples from 21 publicly available studies	Transcriptomics	RF	Cross-platform validation
Ng et al. (2021)	COVID-19 versus other acute respiratory illnesses	333 participants; 380 nasopharyngeal and 53 whole-blood samples	Transcriptomics	SVM, other	Internal hold-out validation

Continued on next page

Table 4 – continued from previous page

Study	Clinical context or diagnostic target	Population and sample size	Input modality	Model families evaluated	Validation design
Nguyen et al. (2022)	COVID-19 detection	1,937 ECG records from a public repository	Physiological signals	NN	Internal validation only
Petkidis et al. (2024)	Viral infectivity classification	Human spiked-saliva datasets	Microscopy	DT, KNN, LR, NB, NN, RF, SVM	Partial cross-platform and spiked-saliva validation
Ramachandran et al. (2022)	Tuberculous meningitis diagnosis	368 patients with cerebrospinal-fluid samples from Uganda	Transcriptomics and metagenomics; decision-level integration	Ensemble methods, SVM, other	Blinded internal test set
Su et al. (2024)	Post-acute COVID-19 syndrome	1,207 participants across Hong Kong cohorts	Microbiome	GB, KNN, NB, RF, SVM	Two-cohort validation
Topff et al. (2023)	COVID-19 diagnosis from CT	2,667 patients and 2,802 CT scans from seven European countries	CT imaging	NN	Institution-held-out external validation
Tran et al. (2022)	Scrub typhus versus dengue	608 patients from five Vietnamese hospitals	Structured clinical data	DT, LR	Internal validation only
Vargas et al. (2023)	Tuberculosis prognosis and treatment response	6,290 observations from 57 publicly available studies	Transcriptomics	GB, LR, NN, RF, SVM, other	Multicohort external validation
Vu et al. (2023)	Dengue and malaria diagnosis	7,543 enrolled participants; 6,208 complete cases from four Kenyan clinics	Structured clinical data	DT, LR, NB, NN, RF, SVM, other	Internal validation only

Abbreviations: CT, computed tomography; DT, decision tree; ECG, electrocardiography; GB, gradient boosting; KNN, k-nearest neighbors; LDA/QDA, linear or quadratic discriminant analysis; LR, logistic regression; NB, naive Bayes; NN, neural network; PCR, polymerase chain reaction; RF, random forest; SVM, support vector machine.