

Clinical Validation of an Offline Android Certainty Factor System for Health-Worker-Assisted Pharyngitis Screening

Emy Susanti^{1)*}, Robby Cokro Buwono²⁾, Dison Librado³⁾, Faiz Rifki Syuhada⁴⁾

^{1,2,3,4)}Informatics Study Program, Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia

¹⁾emy.susanti@email.com, ²⁾robby.cokro@email.com, ³⁾dison.librado@email.com, ⁴⁾faiz.rifki@email.com

Submitted : Sep 25, 2026 | Accepted : Oct 15, 2026 | Published : Oct 18, 2026

Abstract: Certainty Factor (CF) expert systems for disease screening are widely developed in Indonesia, yet they are rarely validated prospectively against specialist diagnosis. A previously published offline Android expert system for acute and chronic pharyngitis reported 93.33% agreement with a single general practitioner on 30 retrospective cases that included no patients with other causes of sore throat. This study prospectively validates that system against otolaryngologist diagnosis, following the STARD 2015 guideline. Consecutive patients with sore throat entered their symptoms into the application with the help of a trained health worker before being examined by an otolaryngologist blinded to the application output. A prespecified threshold ($\tau = 0.50$) allowed outputs to be classified as acute pharyngitis, chronic pharyngitis, or other. Performance was evaluated in two predefined schemes: a three-class scheme (3×3 confusion matrix, one-versus-rest per-class measures, Cohen's kappa, ROC analysis) and a binary scheme (pharyngitis versus other) used to quantify false positives and to select a data-driven threshold, which was validated by repeated stratified 10-fold cross-validation. Expert-weight robustness was examined by Monte Carlo perturbation, and usability with the System Usability Scale (SUS). Among 110 participants, accuracy was 71.8% (95% CI 62.8–79.4) and kappa 0.55. Sensitivity was 89.1% for acute and 88.9% for chronic pharyngitis, but only 21.4% for the “other” class: 22 of 28 patients with other diagnoses were falsely labelled as pharyngitis, accounting for 22 of the 31 errors. Requiring at least one class-specific symptom (a refined rule defined post hoc) reduced false positives from 22 to 13 at the prespecified threshold (specificity 53.6%, sensitivity 95.1%). A data-driven threshold of 0.85 gave 75.0% specificity in-sample but 68.2% under cross-validation (sensitivity 80.9%). The mean SUS score was 68.2 ($n = 60$; average usability). The system detects pharyngitis with high sensitivity but rarely rules out other conditions; it is therefore suitable only as a decision-support aid used with health workers, not for unassisted self-diagnosis, and automatic antibiotic recommendations have been removed from it.

Keywords: Certainty Factor; Clinical Validation; Diagnostic Accuracy; Expert System; Pharyngitis

INTRODUCTION

Pharyngitis is one of the most common reasons for visits to primary health care. Most cases are viral and self-limiting, whereas a minority are bacterial, mainly caused by group A *Streptococcus*, and require antibiotic therapy (Hamilton & McCrea, 2024; Lestari et al., 2022). Because the clinical features of these conditions overlap, clinicians rely on scores such as the Centor criteria and McIsaac's modification to stratify risk and reduce unnecessary antibiotic use (Hamilton & McCrea, 2024; Seeley et al., 2021). Diagnostic reasoning for sore throat is therefore inherently uncertain, and in Indonesia it is further constrained by the uneven distribution of physicians, particularly in remote areas.

Certainty Factor (CF) expert systems are a popular way to model this uncertainty. In Indonesia, the method has been applied to the early diagnosis of pharyngitis and laryngitis (Hasibuan et al., 2024), acute respiratory infections (Mandiri et al., 2024), cardiac arrhythmia (Hutasuhut & Nasyuha, 2021), and autism in children (Ginting et al., 2021). Evaluation in such studies, however, typically ends with black-box functional testing, worked examples, or agreement with a small set of cases assembled by the researchers. Performance in consecutive real

*Emy Susanti



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

patients, diagnosed independently by specialists, is seldom reported, even though this is the evidence that determines whether a system is safe to offer to the public. Systematic evaluations of digital symptom checkers likewise show that diagnostic accuracy is generally low and varies widely between tools (Wallace et al., 2022).

Our previous study developed an offline Android expert system that screens for acute and chronic pharyngitis using thirteen symptoms, a rule base with expert-elicited CF weights from a general practitioner, and six linguistic terms for user certainty (Susanti et al., 2026). Compared with the diagnosis of a general practitioner, the system agreed in 28 of 30 cases (93.33%), with class-level sensitivities of 94.12% for acute and 92.31% for chronic pharyngitis. That study explicitly identified several limitations: the sample was small, evaluation was retrospective, the knowledge and the reference diagnosis came from a single practitioner, the reference was not laboratory-confirmed, and patients with other causes of sore throat were not included. These limitations matter. With $n = 30$, the Wilson 95% confidence interval for the reported agreement spans 78.7% to 98.2%, and because the system always outputs the higher-scoring of the two pharyngitis classes, it has no mechanism to recognise a patient who has neither.

This study addresses these gaps through a prospective diagnostic accuracy study that compares the application with otolaryngologist diagnosis in consecutive patients presenting with sore throat. Specifically, it aims to: (1) estimate accuracy, sensitivity, specificity, predictive values, and chance-corrected agreement (Cohen's kappa) against a specialist reference standard in a population that includes non-pharyngitis cases; (2) introduce a confidence threshold that allows the system to return an "other" output, quantify the resulting false positives, and validate any data-driven threshold by cross-validation; (3) quantify how strongly the results depend on the expert-elicited weights; and (4) evaluate usability and offline technical performance. Because symptom entry was assisted by trained health workers, the application is evaluated as a decision-support tool used together with health workers rather than as an unassisted self-screening tool. The contributions of this work are specialist-referenced, prospective evidence on the clinical performance of the system, a replicable validation protocol for CF-based screening tools, and the identification of symptoms that lay users cannot assess without physical examination.

LITERATURE REVIEW

The Certainty Factor model was introduced in the MYCIN system as a way to represent inexact reasoning in medical diagnosis (Chaki, 2023). A CF value expresses the difference between a *measure of belief* and a *measure of disbelief* in a hypothesis, and evidence from several findings is combined incrementally. The method is computationally light and translates naturally from expert knowledge, which makes it well suited to offline mobile applications. Its weaknesses are equally well known: a CF is not a probability in the statistical sense, and the final output depends heavily on the weights assigned by the expert. The certainty values produced by a CF system therefore need to be calibrated and validated against clinical data rather than only checked for arithmetic correctness.

Various reasoning methods have been compared for respiratory disease expert systems, including forward chaining (Hasanah et al., 2022), machine-learning classifiers such as K-Nearest Neighbor and Naive Bayes (Jamil et al., 2024), and CF versus Bayes' theorem (Ginting et al., 2021). In most of these studies, the comparator is another method or a manual calculation rather than an independently established clinical diagnosis. The reported accuracies therefore reflect the internal consistency of the implementation more than diagnostic performance. Our earlier work moved one step further by reporting a class-level confusion matrix against a general practitioner (Susanti et al., 2026), but it remained retrospective and restricted to confirmed pharyngitis cases.

In clinical research, diagnostic tools are evaluated within the framework of diagnostic accuracy studies. The STARD 2015 guideline and its recent extension for artificial intelligence, STARD-AI, specify the essential reporting items, including a participant flow diagram, a clearly defined reference standard, blinding, and confidence intervals for every accuracy estimate (Sounderajah et al., 2025). Accuracy is also known to be inflated when samples contain only clear-cut cases, a phenomenon termed spectrum bias (Reitsma et al., 2023); a validation set that excludes patients without the target conditions is an extreme form of this bias. For multi-class classification, per-class measures such as precision, recall, and F1 should be reported together with a chance-corrected agreement statistic such as Cohen's kappa (Chicco et al., 2021; Opitz, 2024). This study applies that framework to a CF-based expert system so that its results can be compared directly with the literature on clinical diagnostic tools.

METHOD

This was a prospective cross-sectional diagnostic accuracy study reported according to STARD 2015 and, where applicable, its STARD-AI extension (Sounderajah et al., 2025). The index test was the Android pharyngitis expert system, and the reference standard was the diagnosis made by an otolaryngologist. All participants, or the parents or guardians of participants younger than 18 years, gave written informed consent. The participant flow, including the numbers assessed, excluded, and analysed and the cross-tabulation of index test results against the reference standard, is shown in Figure 1.

*Emy Susanti



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

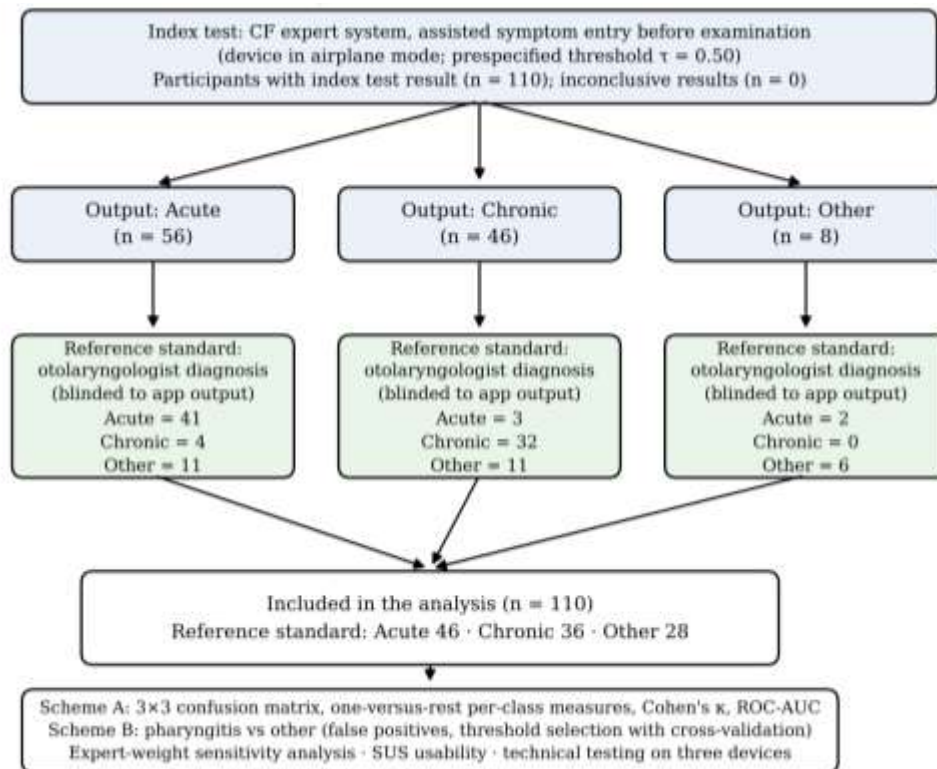


Figure 1. STARD 2015 Participant Flow Diagram

The knowledge base comprises two disease classes, acute pharyngitis (P1) and chronic pharyngitis (P2), and thirteen symptoms (G01–G13). The expert CF weights, elicited from a general practitioner in the previous study, are reproduced in Table 1 because they are the object of the sensitivity analysis. Users rate their certainty for each symptom using the six linguistic terms in Table 2. These weights were elicited from a single general practitioner and have not yet been validated by an independent panel of specialists. The Monte Carlo analysis described below therefore assesses the numerical robustness of the output to plausible disagreement around these values; it does not establish their clinical validity.

Table 1. Expert CF Weights in the Knowledge Base (Single General Practitioner)

Code	Symptom	CF Expert P1 (Acute)	CF Expert P2 (Chronic)
G01	Anorexia	0.4	0.3
G02	Fever	0.8	0.4
G03	Palate deeper than normal	–	0.8
G04	Dysphagia	0.7	0.6
G05	Pharyngeal edema or swelling	0.9	–
G06	Itchy tongue	–	0.7
G07	Nausea	0.5	–
G08	Vomiting	0.5	–
G09	Purplish discoloration of the pharyngeal wall	–	0.9
G10	Frequently itchy base of the throat	–	0.7
G11	Headache	0.6	–
G12	Frequent sneezing	–	0.6
G13	Recurrent sore throat	–	0.8

Table 2. User Certainty Scale (Susanti et al., 2026)

*Emy Susanti

This is anCreative Commons License This work is licensed under a Creative
Commons Attribution-NonCommercial 4.0 International License.

Linguistic Term	CF User
No / not present	0.0
Maybe	0.2
A little sure	0.4
Quite sure	0.6
Sure	0.8
Very sure	1.0

The CF of each symptom is the product of the expert weight and the user certainty, and symptom-level values are combined sequentially for each disease using equations (1) and (2).

$$CF_i = CF_{expert}(i) \times CF_{user}(i) \quad (1)$$

$$CF_{comb} = CF_{old} + CF_i \times (1 - CF_{old}) \quad (2)$$

In the previous version, the application always returned the disease with the higher CF. Because patients with sore throat in routine care may have other conditions, such as tonsillitis, laryngitis, or nonspecific upper respiratory tract infection, this study adds a threshold-based decision rule, given in equation (3). The threshold was prespecified at $\tau = 0.50$ before data collection, and all primary accuracy estimates are reported at this value; because it was fixed in advance, these primary estimates are not affected by threshold overfitting. Any data-driven threshold was treated as exploratory and validated separately by cross-validation, as described below.

$$Class = \operatorname{argmax}(CF_{P1}, CF_{P2}) \text{ if } \max(CF_{P1}, CF_{P2}) \geq \tau; \text{ otherwise "Other"} \quad (3)$$

Refined “Other” rule. The combination rule in equation (2) never decreases as symptoms are added and can reach a high CF from only a few non-specific symptoms, so a threshold on the maximum CF alone may not separate pharyngitis from other conditions. A refined rule (R2) was therefore defined post hoc, after inspection of the error pattern, as a secondary analysis: a pharyngitis class is returned only if $\max(CF_{P1}, CF_{P2}) \geq \tau$ and the participant reports, with a user CF of at least 0.4, at least one symptom that is specific to the winning class (i.e., not shared by both classes). Frequent sneezing (G12) is not counted as specific because it is common in allergic rhinitis and other upper respiratory conditions. Otherwise, the output is “Other”. Because R2 was defined after the errors were known, it is reported as exploratory. It was evaluated at the prespecified $\tau = 0.50$, without any threshold tuning, and in combination with a cross-validated threshold.

Diagnoses were recorded in three categories: acute pharyngitis, chronic pharyngitis, or other. The otolaryngologist was unaware of the application output when making the diagnosis, and the staff member assisting with symptom entry was unaware of the reference diagnosis.

Participants were recruited consecutively. Inclusion criteria were age of at least 12 years, presentation with sore throat or difficulty swallowing, and willingness to participate. Exclusion criteria were emergency conditions, communication impairments that prevented symptom entry, and antibiotic use within the previous seven days. The minimum sample size was calculated with the formula for estimating a single proportion in equation (4).

$$n = Z^2_{1-\alpha/2} \times p \times (1 - p) / d^2 \quad (4)$$

With $Z = 1.96$, an expected overall accuracy of $p = 0.85$, and a precision of $d = 0.07$, the minimum sample was 100 participants. The expected accuracy was set deliberately below the 93.33% reported previously because non-pharyngitis patients are now included, and the required precision is roughly twice as tight as the half-width of the previous confidence interval. The target was increased by 10% to 110 participants to allow for incomplete records.

The assistant read each question aloud and recorded the participant’s own certainty rating without offering interpretation and without examining the participant, so examination-dependent signs such as G03 and G09 were rated by the participants themselves. The device was kept in airplane mode so that all inference ran offline. The CF value for each class, the output class, and the inference time were recorded. The otolaryngologist’s diagnosis was recorded on a separate form, and the two data sets were merged only after data collection had been completed. Because symptom entry was assisted, the study evaluates the application as a tool used together with a health worker; it does not provide evidence on unassisted self-screening.

Two evaluation schemes were defined in advance and are reported separately. **Scheme A (three classes)** is the primary analysis. Application outputs and reference diagnoses were cross-tabulated in a 3×3 confusion matrix; overall accuracy and Cohen’s kappa were computed from the full matrix, and for each class k , TP_k , FP_k , FN_k , and TN_k were derived using a one-versus-rest approach, from which the per-class measures in equations (5) to (9) were calculated. In this scheme, the sensitivity of the “Other” class is the proportion of non-pharyngitis patients correctly returned as “Other”. Macro-averaged values are unweighted means of the three per-class values. **Scheme B (binary)** merges acute and chronic pharyngitis into a single positive class, “pharyngitis”, with “Other” as the

negative class. It was used only to quantify false positives and to select and validate the threshold. Its sensitivity is the proportion of pharyngitis patients returned as either pharyngitis class, and its specificity is numerically identical to the Scheme A sensitivity of the “Other” class.

$$Accuracy = \Sigma_k TP_k / N \quad (5)$$

$$Sensitivity_k = TP_k / (TP_k + FN_k); \quad Specificity_k = TN_k / (TN_k + FP_k) \quad (6)$$

$$PPV_k = TP_k / (TP_k + FP_k); \quad NPV_k = TN_k / (TN_k + FN_k) \quad (7)$$

$$F1_k = 2 \times PPV_k \times Sensitivity_k / (PPV_k + Sensitivity_k) \quad (8)$$

$$\kappa = (p_o - p_e) / (1 - p_e) \quad (9)$$

where p_o is the observed proportion of agreement and p_e is the proportion expected by chance (Chicco et al., 2021). Kappa was interpreted using the conventional benchmarks of slight, fair, moderate, substantial, and almost perfect agreement (Li et al., 2023). Ninety-five percent confidence intervals (CIs) for proportions were computed using the Wilson score method. The ability of each class CF to discriminate that class from all others was evaluated with ROC curves and the area under the curve (AUC) (Çorbacioğlu & Aksel, 2023; Nahm, 2022), with 95% CIs from 2,000 bootstrap resamples. In Scheme B, the threshold τ that best separates pharyngitis (acute or chronic) from other conditions was explored by maximising Youden’s index, $J = \text{sensitivity} + \text{specificity} - 1$ (Nahm, 2022).

Because a threshold selected and evaluated on the same data gives optimistic estimates (Arredondo Montero, 2025), the full-sample optimum was not used as a performance estimate. Threshold selection was instead validated with stratified 10-fold cross-validation repeated 100 times. In each repetition, τ was selected in the nine training folds by maximising Youden’s index over all maximum CF values observed in those folds (i.e., the ROC cut-offs) and was then applied, unchanged, to the held-out fold. Out-of-fold predictions were pooled to obtain cross-validated sensitivity, specificity, predictive values, and three-class accuracy, reported as the mean and the 2.5th–97.5th percentiles across repetitions, together with the distribution of the selected τ values. The refined rule R2 was evaluated with the same procedure.

To test how far the results depend on the subjective judgement of a single expert, every nonzero expert weight was perturbed by a uniformly distributed random amount of ± 0.1 and ± 0.2 , truncated to the range 0 to 1, over 1,000 iterations, following the Monte Carlo treatment of weight uncertainty used in multi-criteria decision analysis (Mazurek & Strzałka, 2022). In each iteration, output classes were recomputed from the same participant responses, and the resulting accuracy, macro-averaged F1, and proportion of changed outputs were recorded. A one-at-a-time analysis, shifting each weight by ± 0.2 in turn, identified the weights with the greatest influence on the output. Because the perturbation is centred on the original weights, this analysis tests the stability of the output, not whether the weights are clinically correct.

Usability was measured with the ten-item System Usability Scale (SUS) on a five-point Likert scale (Vlachogianni & Tselios, 2022), administered to a subsample of 60 participants immediately after the assisted session; the scores therefore reflect the experience of an assisted session. Scores were calculated with equation (10) and interpreted using the adjective rating scale and compared with the published benchmark for digital health apps (Hyzy et al., 2022). Technical performance was assessed on three devices with different Android versions and hardware classes, covering installation, first launch, operation in airplane mode, inference time, and installed application size; the results are reported in Table 9.

$$SUS = 2.5 \times [\Sigma(\text{odd} - \text{item score} - 1) + \Sigma(5 - \text{even} - \text{item score})] \quad (10)$$

Analyses were performed in Python 3 (pandas, NumPy, scikit-learn, and Matplotlib), and the confusion-matrix calculations were cross-checked in Microsoft Excel.

RESULT

During the study period, patients were screened, were excluded, and 110 were included in the analysis. According to the reference standard, 46 participants (41.8%) had acute pharyngitis, 36 (32.7%) had chronic pharyngitis, and 28 (25.5%) had other diagnoses. Participant characteristics are presented in Table 3.

Table 3. Participant Characteristics

Characteristic	Acute Pharyngitis	Chronic Pharyngitis	Other	Total
Participants, n (%)	46 (41.8)	36 (32.7)	28 (25.5)	110 (100)
Age, mean $\pm$ SD (years)	34.0 $\pm$ 16.0	38.8 $\pm$ 12.4	40.4 $\pm$ 15.2	37.2 $\pm$ 14.8
Female, n (%)	28 (60.9)	16 (44.4)	15 (53.6)	59 (53.6)
Symptoms selected, median	5	6	3	5

*Emy Susanti



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Diagnostic Accuracy (Scheme A, Three Classes). Table 4 cross-tabulates the application output at the prespecified threshold of $\tau = 0.50$ against the reference diagnosis, and Table 5 presents the per-class performance measures.

Table 4. Confusion Matrix of Application Output Against Otolaryngologist Diagnosis

Application Output \ Reference	Acute	Chronic	Other	Total
Acute	41	4	11	56
Chronic	3	32	11	46
Other	2	0	6	8
Total	46	36	28	110

Table 5. Per-Class Diagnostic Performance (95% CI)

Measure	Acute Pharyngitis	Chronic Pharyngitis	Other
Sensitivity (%)	89.1 (77.0–95.3)	88.9 (74.7–95.6)	21.4 (10.2–39.5)
Specificity (%)	76.6 (64.9–85.3)	81.1 (70.7–88.4)	97.6 (91.5–99.3)
PPV (%)	73.2 (60.4–83.0)	69.6 (55.2–80.9)	75.0 (40.9–92.9)
NPV (%)	90.7 (80.1–96.0)	93.8 (85.0–97.5)	78.4 (69.5–85.3)
F1-score	0.80	0.78	0.33
AUC	0.81 (0.72–0.89)	0.93 (0.88–0.97)	–

Overall, the application achieved an accuracy of 71.8% (95% CI 62.8–79.4), a macro-averaged F1 of 0.64, and a Cohen's kappa of 0.55 (95% CI 0.42–0.69), which corresponds to moderate agreement (Li et al., 2023). Sensitivity was high and almost identical for acute (89.1%) and chronic pharyngitis (88.9%), and chronic pharyngitis showed the best discrimination (AUC 0.93). The weakest class was “Other”: only 6 of the 28 patients with other diagnoses (21.4%) were returned as “Other”. The dominant direction of error was therefore over-calling pharyngitis; 22 of the 31 misclassifications were non-pharyngitis patients labelled as acute ($n = 11$) or chronic pharyngitis ($n = 11$), which lowered the PPV of both pharyngitis classes to about 70%. Confusion between the two pharyngitis classes was limited, with four chronic cases labelled as acute and three acute cases labelled as chronic, and two acute cases fell below the threshold and were returned as “Other”.

The ROC curves for the class CF values are shown in Figure 2. The CF for each pharyngitis class discriminated that class from all others with an AUC of 0.81 for acute and 0.93 for chronic pharyngitis.

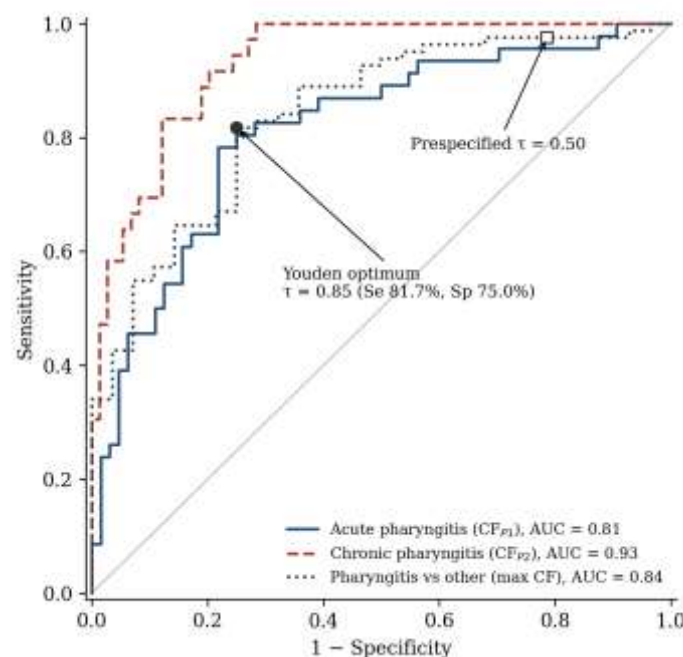


Figure 2. ROC Curves of CF Values for Acute and Chronic Pharyngitis

When acute and chronic pharyngitis were merged into a single positive class, the prespecified threshold of $\tau = 0.50$ gave a sensitivity of 97.6% (80 of 82) but a specificity of only 21.4% (6 of 28) (Table 6). In other words, 22 of the 28 patients without pharyngitis received a pharyngitis label (false-positive rate 78.6%, 95% CI 60.5–89.8), whereas only 2 of the 82 pharyngitis patients were missed. The Scheme B specificity equals the Scheme A sensitivity of the “Other” class because both are the proportion of non-pharyngitis patients returned as “Other”.

Table 6. Binary Evaluation of Pharyngitis Versus Other Conditions (Scheme B)

Decision rule	TP	FN	FP	TN	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, %	NPV, %
$\tau = 0.50$, prespecified (primary)	80	2	22	6	97.6 (91.5–99.3)	21.4 (10.2–39.5)	78.4	75.0
Refined rule R2, $\tau = 0.50$ (post hoc rule, no tuning)	78	4	13	15	95.1 (88.1–98.1)	53.6 (35.8–70.5)	85.7	78.9
$\tau = 0.85$, selected and evaluated in-sample (optimistic)	67	15	7	21	81.7 (72.0–88.6)	75.0 (56.6–87.3)	90.5	58.3
Youden-selected τ , cross-validated	66.3	15.7	8.9	19.1	80.9 (79.3–82.9)*	68.2 (64.3–75.0)*	88.2	55.0
R2 + Youden-selected τ , cross-validated	67.0	15.0	8.1	19.9	81.7 (80.5–84.1)*	71.2 (64.3–75.0)*	89.3	57.1

TP = pharyngitis patients labelled acute or chronic; FN = pharyngitis patients labelled “Other”; FP = patients with other diagnoses labelled acute or chronic; TN = patients with other diagnoses labelled “Other”. Cross-validated rows: stratified 10-fold cross-validation repeated 100 times; counts and percentages are means of pooled out-of-fold results. *Ranges are the 2.5th–97.5th percentiles across repetitions, not confidence intervals.

In the full sample, Youden’s index was maximised at a cut-off of 0.849 (reported as $\tau = 0.85$), which raised specificity to 75.0% and lowered sensitivity to 81.7%, increased the proportion of “Other” outputs from 7.3% to 32.7%, and gave a three-class accuracy of 75.5% (95% CI 66.6–82.5). Because these figures were obtained on the data used to select the threshold, they are optimistic. Under repeated 10-fold cross-validation, the threshold selected in the training folds ranged from 0.819 to 0.853 (median 0.849), showing that the location of the optimum is stable, but the pooled out-of-fold specificity fell to 68.2% at a sensitivity of 80.9%, with a three-class accuracy of 73.1%. The in-sample estimate therefore overstated the specificity of the tuned threshold by about seven percentage points.

The refined rule R2 improved recognition of the “Other” class without any threshold tuning. At the prespecified $\tau = 0.50$, it reduced false positives from 22 to 13, raising specificity from 21.4% to 53.6% (95% CI 35.8–70.5), while sensitivity fell only from 97.6% to 95.1% (two additional pharyngitis patients returned as “Other”). In the three-class scheme, accuracy rose from 71.8% to 79.1% (95% CI 70.6–85.6), macro-averaged F1 from 0.64 to 0.77, and kappa from 0.55 to 0.68. Combined with a cross-validated threshold, R2 achieved a specificity of 71.2% at a sensitivity of 81.7% (Table 6).

Robustness to Expert Weights. The results of the sensitivity analysis are presented in Table 7.

Table 7. Effect of Perturbing Expert CF Weights on Performance (1,000 Iterations)

Scenario	Accuracy, mean (P2.5–P97.5)	Macro F1, mean (P2.5–P97.5)	Outputs Changed (%)
Original weights	71.8	0.639	0
Perturbation ± 0.1	70.8 (68.2–72.7)	0.628 (0.599–0.647)	3.4
Perturbation ± 0.2	69.7 (64.5–73.6)	0.616 (0.573–0.655)	6.8

Performance remained stable under ± 0.1 perturbation: mean accuracy decreased by only 1.1 percentage points, and on average 3.4% of outputs changed class. Under ± 0.2 perturbation, the mean decrease was 2.2 percentage points with 6.8% of outputs changed, and the 95% range of accuracy (64.5–73.6%) still included the original estimate. In the one-at-a-time analysis, the most influential weights were those of the shared symptoms for acute pharyngitis. Increasing the weight of fever (G02, P1) by 0.2 changed 8 outputs and reduced accuracy to 66.4%, whereas decreasing it by 0.2 changed 6 outputs and slightly increased accuracy to 72.7%. Increasing the weights of anorexia (G01, P1) or dysphagia (G04, P1) changed 6 outputs each (accuracy 68.2%). Among the chronic weights, decreasing frequent sneezing (G12, P2) changed 5 outputs; every other single-weight shift changed at most 4 outputs.

From the 31 misclassified cases, 4 were chronic pharyngitis classified as acute. These four participants reported the shared symptoms strongly (mean user CF 0.65 for anorexia, 0.85 for fever, and 0.60 for dysphagia)

*Emy Susanti



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

while answering “no” or “maybe” for the examination-dependent signs (mean 0.05 for both G03 and G09), compared with 0.41, 0.48, and 0.36 for the shared symptoms and 0.14 and 0.11 for G03 and G09 among the 32 correctly classified chronic cases. In all four, the chronic CF was high (0.75–0.93) but slightly below the acute CF (0.79–0.98). Across all chronic cases, 69.4% rated G03 and 77.8% rated G09 as “no” or “maybe”. The largest source of error, however, was the 22 participants with other diagnoses who were classified as pharyngitis, 11 as acute and 11 as chronic. Those labelled as acute mainly reported fever (mean user CF 0.71) and dysphagia (0.40), whereas those labelled as chronic were driven mainly by frequent sneezing (0.64), a symptom weighted 0.6 for chronic pharyngitis. Although these participants selected a median of only three symptoms, the sequential combination rule still produced maximum CF values of 0.61–0.95 (median 0.80), well above $\tau = 0.50$. The remaining errors were three acute cases classified as chronic and two acute cases with low CF values (0.20 and 0.31) that fell below the threshold.

The SUS questionnaire was completed by 60 of the 110 participants. The mean SUS score was 68.2 (SD 12.0; median 68.8; range 45.0–92.5), and 50.0% of participants scored 68 or higher. This mean lies at the benchmark for digital health apps, whose pooled mean is approximately 68 (Hyzy et al., 2022), and falls between the “OK” and “good” adjective ratings, indicating average rather than good usability. Item-level results are shown in Table 8; the lowest-scoring item was item 9, confidence in using the system (mean contribution 2.57 of 4), and the highest-scoring items were item 7, expected ease of learning, and item 8, absence of cumbersomeness (both 2.82). Item scores were otherwise uniform (2.68–2.77), indicating no single usability problem. Because the questionnaire was completed after an assisted session, these scores may overestimate the usability of unassisted use.

Table 8. Item-Level System Usability Scale Results (n = 60)

No.	SUS item	Response, mean (SD) (1–5)	Score contribution, mean (0–4)
1	I think that I would like to use this system frequently.	3.72 (0.78)	2.72
2	I found the system unnecessarily complex.	2.28 (0.74)	2.72
3	I thought the system was easy to use.	3.75 (0.77)	2.75
4	I think that I would need the support of a technical person to be able to use this system.	2.30 (0.79)	2.70
5	I found the various functions in this system were well integrated.	3.68 (0.77)	2.68
6	I thought there was too much inconsistency in this system.	2.23 (0.85)	2.77
7	I would imagine that most people would learn to use this system very quickly.	3.82 (0.79)	2.82
8	I found the system very cumbersome to use.	2.18 (0.72)	2.82
9	I felt very confident using the system.	3.57 (0.79)	2.57
10	I needed to learn a lot of things before I could get going with this system.	2.27 (0.76)	2.73
SUS score (0–100)		–	68.2 (SD 12.0)

DISCUSSIONS

This study provides prospective, specialist-referenced evidence on the performance of the CF pharyngitis expert system in consecutive patients with sore throat. The agreement obtained ($\kappa = 0.55$) indicates moderate agreement beyond chance (Li et al., 2023). The overall accuracy of 71.8% is lower than the 93.33% agreement reported in the retrospective evaluation (Susanti et al., 2026). A lower figure was expected, but it must be interpreted with care: the earlier evaluation included only confirmed pharyngitis cases, so the system never faced a patient who had neither class, and its reference came from the same practitioner who supplied the knowledge base. The present estimate is therefore a more realistic indication of what users in routine care can expect.

The principal weakness of the system is its inability to rule out pharyngitis. Only 6 of the 28 patients with other diagnoses were returned as “Other” (sensitivity 21.4%), and these false positives accounted for 22 of the 31 errors. For a screening tool, this means that most patients without pharyngitis would be told that they have pharyngitis, which may cause unnecessary worry, inappropriate self-medication, and delay in identifying the actual cause of their symptoms. Two mechanisms contribute. First, the combination rule in equation (2) only increases as symptoms are added, so a few non-specific symptoms such as fever, dysphagia, or frequent sneezing are enough to push the CF above $\tau = 0.50$ (median maximum CF 0.80 among the false positives). Second, the knowledge base contains no evidence that argues against pharyngitis or points towards alternative diagnoses, so the system has no way to lower its belief. Requiring a class-specific symptom (rule R2) reduced false positives from 22 to 13 with

little loss of sensitivity, which confirms the first mechanism, but about half of the non-pharyngitis patients were still labelled as pharyngitis. A durable solution therefore requires extending the knowledge base with the most frequent alternative diagnoses, such as tonsillitis, laryngitis, and non-specific upper respiratory tract infection, and with disconfirming evidence in the form of measures of disbelief.

The most important finding concerns the structure of the knowledge base. Three symptoms, anorexia, fever, and dysphagia, are shared by both classes, and in each case the weight for acute pharyngitis is higher (0.4, 0.8, and 0.7 versus 0.3, 0.4, and 0.6). Meanwhile, the two most heavily weighted signs of chronic pharyngitis, a deeper-than-normal palate (G03, 0.8) and purplish discoloration of the pharyngeal wall (G09, 0.9), can only be judged by physical examination. When lay users cannot assess these signs, the chronic score is suppressed and chronic cases drift towards the acute class. The error analysis is consistent with this mechanism: the four chronic cases labelled as acute reported the shared symptoms strongly while rating G03 and G09 close to zero. The effect was smaller than anticipated, however, because most chronic patients also reported self-reportable chronic symptoms such as recurrent sore throat (G13), so chronic sensitivity remained 88.9%. Conversely, frequent sneezing (G12), a self-reportable but non-specific chronic symptom, pulled patients with other conditions towards the chronic class. The questionnaire should therefore separate self-reportable symptoms from examination-dependent signs. Together with the assisted design of this study, these findings mean that the application should be positioned as a decision-support tool used by or together with health workers in primary care, who can inspect the pharynx and rate examination-dependent signs, rather than as a self-diagnosis tool. Whether lay users can use it safely without assistance has not been tested.

The threshold-based decision rule also proved necessary. Without an “Other” category, every patient is labelled as acute or chronic pharyngitis, including patients with other conditions. In Scheme B, the full-sample optimum of $\tau = 0.85$ raised specificity from 21.4% to 75.0% while lowering sensitivity from 97.6% to 81.7%. At this threshold, three-class accuracy rose to 75.5% and macro-averaged F1 to 0.75, but 15 of the 82 pharyngitis patients (18.3%) would be returned as “Other”. These in-sample figures are optimistic because the threshold was selected on the same data used to evaluate it (Arredondo Montero, 2025). The cross-validated estimates in Table 6 (specificity 68.2%, sensitivity 80.9%) are the appropriate basis for choosing a threshold, and the prespecified $\tau = 0.50$ remains the basis for all primary results. For screening, the refined rule R2 at the prespecified threshold offers a better balance than threshold tuning alone: it removes 9 of the 22 false positives while keeping sensitivity at 95.1%, whereas the tuned threshold misses about one in five pharyngitis patients. Because R2 was defined post hoc, it should be confirmed, together with any final threshold, in an independent sample.

The sensitivity analysis showed that performance was stable under ± 0.1 perturbation of the expert weights. The output is therefore governed more by the symptom pattern than by the exact weight values. The weights with the greatest leverage were those of the shared symptoms fever (G02), anorexia (G01), and dysphagia (G04) for acute pharyngitis, which are the same weights implicated in confusion between the two pharyngitis classes. This is relevant to CF expert systems in general, which rarely report how much their outputs depend on the judgement of a single expert. Robustness, however, is not the same as validity. The perturbation was centred on the weights of a single general practitioner, so a stable output shows only that small disagreements around those values would not change the conclusions, not that the values themselves are clinically correct. Elicitation from several otolaryngologists, with formal assessment of inter-expert agreement such as the intraclass correlation coefficient or a Delphi procedure, is required before the weights can be regarded as clinically validated.

From a patient-safety perspective, the treatment module of the version evaluated recommended antibiotics for both classes of pharyngitis. This is a patient-safety issue, not merely an algorithmic limitation. Most acute pharyngitis is viral, and clinical guidance reserves antibiotics for suspected bacterial infection based on clinical scores or diagnostic testing (Hamilton & McCrea, 2024; Seeley et al., 2021; Spinks et al., 2021). A CF model based on self-reported symptoms cannot determine whether an infection is bacterial, and in this study 22 of the 28 patients without pharyngitis would have received a pharyngitis label and, with it, an antibiotic recommendation. The result screen now shows only the screening result, general supportive-care advice, warning signs that require prompt medical attention, and a recommendation to be examined by a health worker, who can apply the Centor/McIsaac criteria or diagnostic testing to decide whether antibiotics are indicated.

This study has several limitations. First, data were collected at a single health facility, so generalisation to other populations requires caution. Second, the reference standard was a specialist clinical diagnosis without laboratory confirmation for every patient, so misclassification in the reference cannot be fully excluded. Third, symptom entry was assisted by trained staff, so the results apply only to assisted use; performance under unassisted use by the public may be lower, as the self-triage accuracy of lay users and symptom assessment apps varies considerably (Kopka et al., 2025), and a separate study of unassisted use is required before any self-screening claim can be made. Fourth, the expert weights came from a single general practitioner, and the Monte Carlo analysis demonstrates numerical robustness but not clinical validity. Fifth, threshold selection was validated only internally by cross-validation, and an independent sample is still needed. Sixth, only 28 participants had other diagnoses, so the estimates for the “Other” class have wide confidence intervals (sensitivity 10.2–39.5%; PPV

40.9–92.9%), and the refined rule R2 was defined post hoc. Finally, usability was measured in a subsample of 60 participants after assisted sessions.

CONCLUSION

This study prospectively validated an offline Android CF expert system for pharyngitis screening against otolaryngologist diagnosis in 110 consecutive patients with sore throat. At the prespecified threshold, the application achieved an accuracy of 71.8% and a kappa of 0.55, with high sensitivity for acute (89.1%) and chronic pharyngitis (88.9%). Its main weakness was the “Other” class: only 21.4% of patients without pharyngitis were recognised as such, and these false positives accounted for 22 of the 31 errors. Confusion between acute and chronic pharyngitis, linked to examination-dependent signs, was a secondary source of error (7 of 31). Requiring a class-specific symptom reduced false positives from 22 to 13 at the prespecified threshold while keeping sensitivity at 95.1%, whereas the data-driven threshold of 0.85 was optimistic in-sample (specificity 75.0%) and reached 68.2% under cross-validation. Usability was average (mean SUS 68.2, $n = 60$). On this basis, the application is suitable only as a decision-support aid used with health workers, not for unassisted self-diagnosis, and it no longer recommends antibiotics. Future work should extend the knowledge base with alternative diagnoses and disconfirming evidence, separate self-reportable symptoms from clinical signs, incorporate the Centor/McIsaac criteria, elicit weights from multiple specialists with formal agreement testing, and conduct multicentre validation, including a separate study of unassisted use.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the **Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kemdiktisaintek)** for the support and funding provided through the **Beginner Lecturer Research Grant (Hibah Penelitian Dosen Pemula) 2026**. This research would not have been possible without the valuable assistance, encouragement, and financial support provided by this grant. The authors also extend their appreciation to all parties who contributed to the successful implementation of this research.

REFERENCES

- Arredondo Montero, J. (2025). A tailored fit that doesn't fit all: The problem of threshold overfitting in diagnostic studies. *Diagnosis*, 12(3). <https://doi.org/10.1515/dx-2025-0096>
- Chaki, J. (2023). Certainty factor and evidential reasoning to handle uncertainty in artificial intelligence. In *Handling uncertainty in artificial intelligence* (pp. 37–45). Springer. https://doi.org/10.1007/978-981-99-5333-2_4
- Chicco, D., Warrens, M. J., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more informative than Cohen's kappa and Brier score in binary classification assessment. *IEEE Access*, 9, 78368–78381. <https://doi.org/10.1109/ACCESS.2021.3084050>
- Çorbacioğlu, Ş. K., & Aksel, G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish Journal of Emergency Medicine*, 23(4), 195–198. https://doi.org/10.4103/tjem.tjem_182_23
- Ginting, R., Zarlis, M., & Rosnelly, R. (2021). Analisis perbandingan metode Certainty Factor dan Teorema Bayes untuk mendiagnosa penyakit autisme pada anak. *Jurnal Media Informatika Budidarma*, 5(2), 583–589. <https://doi.org/10.30865/mib.v5i2.2930>
- Hamilton, J. L., & McCrea, L. (2024). Streptococcal pharyngitis: Rapid evidence review. *American Family Physician*, 109(4), 343–349.
- Hasanah, R., Muhasshanah, & Helmi, F. (2022). Sistem pakar diagnosa penyakit saluran pernapasan menggunakan metode forward chaining. *Jurnal Advance Research Informatika*, 1(1), 33–50. <https://doi.org/10.24929/jars.v1i01.2450>
- Hasibuan, M. S., Triase, & Hutabarat, D. W. H. (2024). Sistem pakar diagnosis awal penyakit faringitis dan laringitis menggunakan metode forward chaining dan Certainty Factor. *Journal of Science and Social Research*, 7(3), 1137–1146. <https://doi.org/10.54314/jssr.v7i3.2144>
- Hutasuhut, M., & Nasyuha, A. H. (2021). Analisis aritmia (gangguan irama jantung) menerapkan metode Certainty Factor. *Jurnal Media Informatika Budidarma*, 5(4), 1386–1393. <https://doi.org/10.30865/mib.v5i4.3289>
- Hyzy, M., Bond, R., Mulvenna, M., Bai, L., Dix, A., Leigh, S., & Hunt, S. (2022). System Usability Scale benchmarking for digital health apps: Meta-analysis. *JMIR mHealth and uHealth*, 10(8), Article e37290. <https://doi.org/10.2196/37290>

- Jamil, M., Rozi, F., & Saputra, Y. F. (2024). Komparasi kinerja algoritma machine learning untuk deteksi penyakit infeksi saluran pernapasan. *Jurnal Rekayasa Teknologi Informasi (JURTI)*, 8(1), 84–92. <https://doi.org/10.30872/jurti.v8i1.16965>
- Kopka, M., von Kalckreuth, N., & Feufel, M. A. (2025). Accuracy of online symptom assessment applications, large language models, and laypeople for self-triage decisions. *npj Digital Medicine*, 8, Article 178. <https://doi.org/10.1038/s41746-025-01566-6>
- Lestari, D. L. P. A., Jayanti, N. P. S. D., Putra, T. W., Fridayanthi, P. U., Tjahyadi, I. G. K. D. P. P., Maharani, L. G. S., & Cahyawati, P. N. (2022). Diagnosis dan tatalaksana faringitis Streptococcus Group A. *WICAKSANA: Jurnal Lingkungan dan Pembangunan*, 6(2), 88–95. <https://doi.org/10.22225/wicaksana.6.2.2022.88-95>
- Li, M., Gao, Q., & Yu, T. (2023). Kappa statistic considerations in evaluating inter-rater reliability between two raters: Which, when and context matters. *BMC Cancer*, 23, Article 799. <https://doi.org/10.1186/s12885-023-11325-z>
- Mandiri, P. D., Hartanti, D., & Sari, A. A. (2024). Prototipe sistem pakar untuk diagnosis penyakit infeksi saluran pernapasan (ISPA) menggunakan metode Certainty Factor. *SKANIKA: Sistem Komputer dan Teknik Informatika*, 7(2), 180–191.
- Mazurek, J., & Strzałka, D. (2022). On the Monte Carlo weights in multiple criteria decision analysis. *PLOS ONE*, 17(10), Article e0268950. <https://doi.org/10.1371/journal.pone.0268950>
- Nahm, F. S. (2022). Receiver operating characteristic curve: Overview and practical use for clinicians. *Korean Journal of Anesthesiology*, 75(1), 25–36. <https://doi.org/10.4097/kja.21209>
- Opitz, J. (2024). A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics*, 12, 820–836. https://doi.org/10.1162/tacl_a_00675
- Reitsma, J. B., Rutjes, A. W. S., Whiting, P., Yang, B., Leeflang, M. M. G., Bossuyt, P. M., & Deeks, J. J. (2023). Assessing risk of bias and applicability. In J. J. Deeks, P. M. Bossuyt, M. M. G. Leeflang, & Y. Takwoingi (Eds.), *Cochrane handbook for systematic reviews of diagnostic test accuracy* (2nd ed.). Wiley. <https://doi.org/10.1002/9781119756194.ch8>
- Seeley, A., Fanshawe, T., Voysey, M., Hay, A., Moore, M., & Hayward, G. (2021). Diagnostic accuracy of Fever-PAIN and Centor criteria for bacterial throat infection in adults with sore throat: A secondary analysis of a randomised controlled trial. *BJGP Open*, 5(6), Article BJGPO.2021.0122. <https://doi.org/10.3399/BJGPO.2021.0122>
- Sunderajah, V., Guni, A., Liu, X., Collins, G. S., Karthikesalingam, A., Markar, S. R., Golub, R. M., Denniston, A. K., Shetty, S., Moher, D., Bossuyt, P. M., Darzi, A., & Ashrafian, H. (2025). The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature Medicine*, 31(10), 3283–3289. <https://doi.org/10.1038/s41591-025-03953-8>
- Spinks, A., Glasziou, P. P., & Del Mar, C. B. (2021). Antibiotics for treatment of sore throat in children and adults. *Cochrane Database of Systematic Reviews*, 2021(12), Article CD000023. <https://doi.org/10.1002/14651858.CD000023.pub5>
- Susanti, E., Buwono, R. C., Librado, D., & Syuhada, F. R. (2026). AI-based pharyngitis detection expert system using the Certainty Factor method. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 10(3), 1908–1920. <https://doi.org/10.33395/sinkron.v10i3.16463>
- Vlachogianni, P., & Tselios, N. (2022). Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review. *Journal of Research on Technology in Education*, 54(3), 392–409. <https://doi.org/10.1080/15391523.2020.1867938>
- Wallace, W., Chan, C., Chidambaram, S., Hanna, L., Iqbal, F. M., Acharya, A., Normahani, P., Ashrafian, H., Markar, S. R., Sunderajah, V., & Darzi, A. (2022). The diagnostic and triage accuracy of digital and online symptom checker tools: A systematic review. *npj Digital Medicine*, 5, Article 118. <https://doi.org/10.1038/s41746-022-00667-w>