Original article

Outbreak, Surveillance, Investigation & Response (OSIR) Journal
ISSN 2651-1061 (Online)
https://he02.tci-thaijo.org/index.php/OSIR
Email: osireditor@osirjournal.net

March 2026
Volume 19, Issue 1
Article no. 279422

Machine Learning Application: Congenital Syphilis Classifier

Suppasit Srisaeng*, Chanissara Thuwakham

Office of Disease Prevention and Control 6 Chonburi, Department of Disease Control, Ministry of Public Health, Thailand

*Corresponding author, email address: mumiman@gmail.com

Received: 21 Dec 2025; Revised: 7 Feb 2026; Accepted: 23 Feb 2026
https://doi.org/10.59096/osir.v19i1.279422

Abstract

Objectives: Congenital syphilis (CS) is resurging in Thailand, which is challenging the nation's elimination status. Current surveillance relies on manual expert review of paper-based case investigation forms (CIFs), resulting in significant backlogs which hinder timely public health intervention. This study aimed to develop and evaluate an automated system using Large Language Models (LLMs) to digitize and classify CS cases.

Methods: We conducted a retrospective diagnostic study using 143 validated CIFs from Eastern Thailand during October 2024 to October 2025. The system utilized Google Gemini 2.5 Pro and Flash models for optical character recognition (OCR) and a rule-based algorithm for case classification. Cost was estimated by Google Gemini application programming interface (API) pricing. Performance was benchmarked against an expert committee's consensus.

Results: The system efficiently reduced processing time from months to under five minutes at a cost of approximately $0.006 per case. It achieved a data extraction accuracy of 94.2% with a Character Error Rate (CER) of 4.32% for complex handwriting. The overall classification accuracy was 80.0%, with sensitivity 83.3% and specificity of 75.0%. Automated "Red Flags" detection identified inadequate maternal treatment (69.7%) and late antenatal care (32.9%) as primary drivers for confirmed and probable cases.

Public Health Recommendations: The LLM and rule-based classifier offer a scalable "Paper-to-Digital Bridge" for resource-limited settings. While human oversight remains necessary for complex cases, the system functions as an effective high-volume triage tool, transforming retrospective surveillance into real-time actionable intelligence to support the elimination of mother-to-child transmission.

Keywords: artificial intelligence, congenital syphilis, surveillance, Thailand, Large Language Models, optical character recognition

Introduction

Congenital syphilis (CS) remains a critical public health challenge both globally and in Thailand, despite being preventable and curable. In 2022, the World Health Organization (WHO) estimated 8 million adults aged 15–49 were infected with syphilis worldwide.[1] Untreated maternal *Treponema pallidum* infection can lead to devastating outcomes, including stillbirth, prematurity, low birth weight, and multisystemic congenital infection.[2] Notably, CS is the second leading cause of stillbirth globally.[3] In 2016, Thailand achieved a milestone as the first nation in the Southeast-Asia region to be validated by the WHO for the elimination of mother-to-child transmission (EMTCT) of both HIV and syphilis, maintaining an incidence rate of fewer than 50 cases per 100,000 live births.[4] However, the national CS incidence had risen sharply to 247.7 cases per 100,000 live births by 2024, with the highest regional burden observed in the Eastern region (Health Region 6; HR6), where incidence reached of 414 cases per 100,000 live births.[5]

In Thailand, potential congenital syphilis cases are primarily detected through the national universal screening program at clinic or hospital with reverse

screening algorithm (treponemal then non-treponemal testing) for all pregnant women at their first antenatal care (ANC) visit, third trimester, and labor. For infants born to seropositive mothers, clinical diagnosis at the hospital is based on physical examination for signs of infection, comparison of neonatal and maternal non-treponemal titers, and specialized investigations such as long-bone radiography or cerebrospinal fluid analysis where indicated. First, a clinician diagnoses a suspected case with the International Classification of Diseases, 10th Edition (ICD-10) codes, then a hospital epidemiologist documents the comprehensive clinical and laboratory data onto a paper-based case investigation form (CIF). Subsequently, official CS classifications and verifications are performed by expert review committees (pediatricians and obstetricians) who manually evaluate these detailed paper-based forms at the regional and national levels.[6] This process needs careful consideration, which is slow, leading to a backlog of >60 cases in six months for HR6 area. Such delays impede timely public health action (e.g., reporting, contact tracing, and treatment follow-up), undermining efforts to control CS in high-incidence areas.

Advances in machine learning (ML), particularly within large language models (LLMs), offer an opportunity to streamline CS surveillance. Automating the CIF data extraction and classification process could drastically reduce working time, by leveraging optical character recognition (OCR) and the understandable rule-based algorithms. This approach aligns with innovative solutions to strengthen sexually transmitted infection (STI) surveillance and global health equity in resource-constrained settings.[7]

This study tried to address a critical operational gap by introducing a novel digital tool to aid in achieving CS elimination targets in Thailand. This study aimed to 1) describe the expert-classified CS newborn and maternal characteristics, timeliness of reporting, and the prevalence of ANC and newborn risk indicators (marked as Red Flags); 2) develop a ML based system for classifying CS cases; and 3) evaluate the performance and resource use aspect.

Methods

Study Design and Data Collection

This retrospective diagnostic test study used all the 143 validated CIFs in Thailand's HR6 from October 2024 to October 2025 from the Division of Epidemiology, Department of Disease Control (DDC), Ministry of Public Health. Public hospitals identified a suspected CS case when a newborn was delivered to a syphilis mother. The hospital epidemiologists conducted case investigation by collecting maternal ANC history and maternal/newborn laboratory and treatment information using the standardized paper-based CIF. The CIF was validated by the hospital's attending pediatricians and/or obstetricians, scanned or photographed and submitted to the DDC through the Congenital Syphilis Response System website, typically after neonatal evaluation was completed or after completion of a 10-day treatment course when indicated. The CIFs contain maternal, paternal, neonatal, and laboratory information used for case classification. The reference outcome classified as unlikely, probable, or confirmed CS—was established by consensus of the HR6 expert committee (at least 10 pediatric and obstetrician specialists from eight eastern provinces), as per national guidelines.[6]

This study described maternal age in median with quantile 1 and quantile 3, case distribution by province and calculated the case fatality rate as the proportion of stillbirths among all included cases. Reporting timeliness was defined as the interval in days between the newborn's date of birth and the date the case was reported/received for review, as recorded in the validated CIF. Timeliness was summarized using the median with interquartile range (IQR). This study also summarized the prevalence of ANC and newborn risk indicators (Red Flags) derived from validated CIF data and results were reported as counts and percentages of the total study population.

Study Definitions

Data fields

Titers: Laboratory ratios which represented antibody dilution (e.g., "1:8", "1:32") or qualitative results ("non-reactive (NR)", "reactive").

Dates: Date of birth, treatment, and laboratory testing, normalized from the Thai Buddhist Era (BE) to the Gregorian calendar (dd/mm/yyyy).

Checkbox: Selection fields which indicated the presence or absence of clinical signs, test performance, or status (e.g., "done/not done", "normal/abnormal").

Numeric: Continuous variables which included birth weight (grams), gestational age (weeks), and maternal age (years).

Free text: Unstructured handwritten notes which included diagnoses, ICD-10 codes, and clinical remarks.

Case classification

Specific classification: Cases were categorized using predefined alphanumeric codes in accordance with the investigation protocols of the Division of Epidemiology.[6] Confirmed cases were defined as those meeting any of the following criteria: code 1.1 (inadequate maternal treatment with infant clinical findings), code 1.2 (adequate maternal treatment with infant clinical findings), code 1.3 (syphilitic stillbirth), code 1.4 (positive infant serology at six months), and code 1.5.1 (inadequate maternal treatment with positive infant serology at six months). Probable cases were classified as code 1.5.2 (inadequate maternal treatment with an asymptomatic infant and negative or unknown baseline serology). Cases were considered unlikely (not a case) if classified as code 2 (adequate maternal treatment with an asymptomatic infant and normal serological findings).

Overall classification: For the purposes of summary performance analysis, individual codes were further grouped into three categories: confirmed (codes 1.1–1.5.1), probable (code 1.5.2), and unlikely (code 2).

Red Flags

This study defined Red Flags as automated alerts derived from extracted data to assist epidemiologists in monitoring high-risk scenarios. These included incomplete antenatal care (ANC), defined as fewer than four recorded ANC visits; late ANC initiation, defined as the first ANC visit occurring after 12 weeks of gestation; parental human immunodeficiency virus (HIV) infection, defined as documented HIV positivity in either the mother or father; and inadequate maternal treatment, defined as failure to receive a benzathine penicillin G regimen appropriate to the disease stage, with the final dose administered more than 30 days prior to delivery. Comparative analyses between the confirmed/probable CS group and the unlikely CS group were performed using the chi-square test or Fisher's exact test, as appropriate.

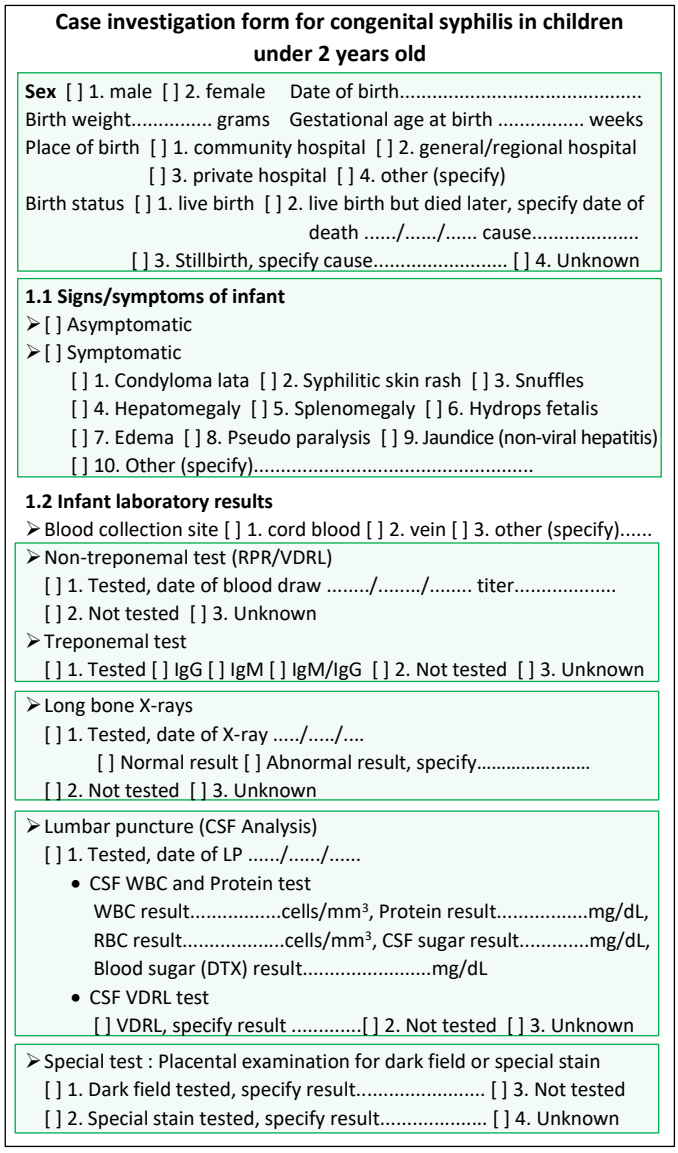
Application Development

Form version matching and alignment

To standardize the input images for processing, this study developed a computer vision pipeline using Python and open-source computer vision library. Due to the variability in scanning quality, input images undergo preprocessing including contrast limited adaptive histogram equalization (CLAHE), illumination correction, and unsharp masking to enhance features. This study utilized the Oriented FAST and Rotated BRIEF (ORB) feature detector to identify key points in the input forms. These features were matched against reference templates of the five CIF versions using k-nearest neighbors algorithm with a ratio test to filter false matches. A homography matrix was estimated using Random Sample Consensus to correct perspective and rotation distortions. Finally, an Enhanced Correlation Coefficient maximization algorithm was applied for pixel-level fine-tuning of the alignment.

Region of interest (ROI)

This study manually defined ROI bounding boxes to guide data extraction. These included five de-identification ROIs censoring personally identifiable information (Figure 1), which were redacted prior to OCR, and 15 data extraction ROIs which targeted key information such as maternal and newborn laboratory results (e.g., Rapid Plasma Reagin (RPR)/Venereal Disease Research Laboratory titers) and treatment history, newborn physical examination findings, X-ray results, and paternal laboratory and treatment history.



Case investigation form for congenital syphilis in children under 2 years old

Sex [ ] 1. male [ ] 2. female Date of birth.....
Birth weight..... grams Gestational age at birth ..... weeks
Place of birth [ ] 1. community hospital [ ] 2. general/regional hospital
[ ] 3. private hospital [ ] 4. other (specify)
Birth status [ ] 1. live birth [ ] 2. live birth but died later, specify date of death ...../...../..... cause.....
[ ] 3. Stillbirth, specify cause..... [ ] 4. Unknown

1.1 Signs/symptoms of infant
- [ ] Asymptomatic
- [ ] Symptomatic
  [ ] 1. Condyloma lata [ ] 2. Syphilitic skin rash [ ] 3. Snuffles
  [ ] 4. Hepatomegaly [ ] 5. Splenomegaly [ ] 6. Hydrops fetalis
  [ ] 7. Edema [ ] 8. Pseudo paralysis [ ] 9. Jaundice (non-viral hepatitis)
  [ ] 10. Other (specify).....

1.2 Infant laboratory results
- Blood collection site [ ] 1. cord blood [ ] 2. vein [ ] 3. other (specify).....
- Non-treponemal test (RPR/VDRL)
  [ ] 1. Tested, date of blood draw ...../...../..... titer.....
  [ ] 2. Not tested [ ] 3. Unknown
- Treponemal test
  [ ] 1. Tested [ ] IgG [ ] IgM [ ] IgM/IgG [ ] 2. Not tested [ ] 3. Unknown
- Long bone X-rays
  [ ] 1. Tested, date of X-ray ...../...../....
  [ ] Normal result [ ] Abnormal result, specify.....
  [ ] 2. Not tested [ ] 3. Unknown
- Lumbar puncture (CSF Analysis)
  [ ] 1. Tested, date of LP ...../...../.....
  - CSF WBC and Protein test
    WBC result.....cells/mm³, Protein result.....mg/dL,
    RBC result.....cells/mm³, CSF sugar result.....mg/dL,
    Blood sugar (DTX) result.....mg/dL
  - CSF VDRL test
    [ ] VDRL, specify result .....[ ] 2. Not tested [ ] 3. Unknown
- Special test : Placental examination for dark field or special stain
  [ ] 1. Dark field tested, specify result..... [ ] 3. Not tested
  [ ] 2. Special stain tested, specify result..... [ ] 4. Unknown

Figure 1. De-identification region of interest censoring case investigation form.

OCR process and LLM integration

Extracted ROI images were processed using Google's Gemini 2.5 Pro and Flash models via the Google Gemini application programming interface (API). To balance cost and accuracy, this study implemented a heuristic model selection strategy which included Gemini 2.5 Flash—used for high-clarity fields with standard layouts (e.g., demographic checkboxes), and Gemini 2.5 Pro—reserved for complex fields requiring high-acuity handwriting recognition, specifically maternal laboratory and treatment history (extracting dates and doses from unstructured notes) and newborn laboratory and X-ray results. To optimize API latency and token usage, images were padded to the nearest 768-pixel boundary to align with the model's native tile patching architecture.

Prompt engineering and data extraction

This study employed a prompting strategy with strict JavaScript Object Notation schema enforcement. The system prompts were designed to act as an "expert epidemiologist," with specific instructions to standardize titers by normalizing variations (e.g., "weakly reactive", "1:4", "RPR 1:8") into a standard enumerated list; normalize dates by converting various handwritten dates (e.g., 2024, 2566, 2567 B.E.) into the standard format (dd/mm/yyyy); apply checkbox logic to distinguish between checked, unchecked, and crossed-out (error) boxes, returning null for ambiguous entries rather than hallucinating values; and handle bilingual inputs by processing mixed Thai and English handwriting without translation while preserving the original clinical context.

Prompt improvement

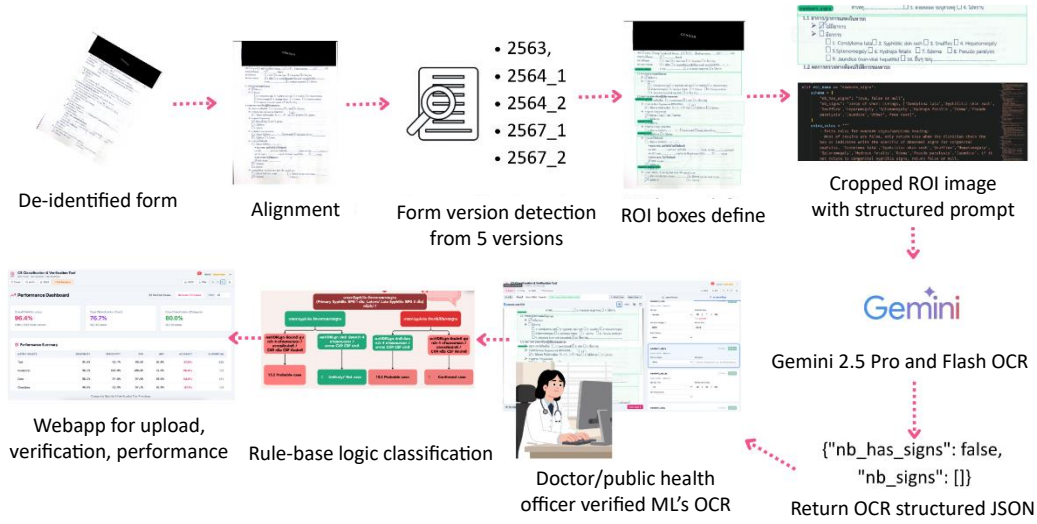
The dataset was randomly split into a training set (80%, n=113 forms, 5,598 data fields) and a testing set (20%, n=30 forms, 1,492 data fields). Fourteen prompt iterations were conducted to achieve satisfactory performance on the training set. Key improvements included adding negative constraints to prevent the model from reading printed boilerplate text and refining the logic for interpreting complicated information, e.g., maternal treatment adequacy.

Classification model

The structured JavaScript Object Notation output from the OCR module was fed into a simple rule-based classification algorithm based on national guideline, integrating the extracted maternal treatment adequacy, newborn titers (comparing more than four-fold differences from maternal titers), abnormal physical examination findings, and long-bone X-ray results to automatically classify cases.[6] Additionally, the system flagged the Red Flags signs to alert users.

Performance evaluation

The performance metrics used sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score, and accuracy on field data types, overall CS status, and specific CS code status. The 95% confidence interval (CI) was computed using Wilson Score Interval.[8] The text extraction quality was further evaluated using the character error rate (CER). CER is a standard metric for assessing OCR performance that measures the percentage of character-level mistakes, e.g., incorrect, missing, or extra characters, produced by the system when compared to the expert-validated ground truth.[9] The cost was calculated from Google Gemini API pricing.[10] This study summarized the application workflow in Figure 2.



ROI: region of interest. ML: machine learning. OCR: optical character recognition.

Figure 2. Application's workflow.

Results

Maternal age was recorded in 139 of the 143 validated CIFs, with a median age of 23 years (Q1=20, Q3=27). The majority of validated CIFs were from tertiary care facilities, including regional hospitals (>500 beds; 86, 60.1%) and large general hospitals (200–500 beds; 38, 26.6%). The remaining validated CIFs were provided by secondary care facilities (100–200 beds; 11, 7.7%), comprising smaller general hospitals and large community hospitals, and primary care facilities (30–100 beds; 8, 5.6%), which included medium-to-large community hospitals. The expert committee classified 21 cases (14.7%) as confirmed, 55 cases (38.5%) as probable, and 67 cases (46.8%) as unlikely CS. There were 55, 38, 22, 17, 6, 6, and 1 case(s) in Chon Buri, Samut Prakan, Chanthaburi, Rayong, Chachoengsao, Trat, and Sa Kaeo provinces, respectively. Three stillbirths were reported, resulting in a case fatality rate of 2.1%. The median report time was 38 days (Q1=19, Q3=56 days) after newborn birth. The maternal median age was 23 years old (Q1=19, Q3=26.25) among the confirmed and probable CS group, and 24 years old (Q1=20, Q3=27) among the unlikely CS group.

Red Flag indicators differed between the confirmed/probable CS group (n=76) and the unlikely CS group (n=67). Maternal treatment inadequate was more frequent in the confirmed/probable group than in the unlikely group (69.7% vs. 17.9%, *p*-value <0.001). No antenatal care documented (ANC=0) occurred only among confirmed/probable cases (21.1% vs. 0.0%, *p*-value <0.001). Among newborn outcomes, preterm birth was more common in the confirmed/probable group than in the unlikely group (28.9% vs. 4.5%, *p*-value <0.001). Other maternal indicators (late ANC >12 weeks, very late ANC >20 weeks, maternal RPR 4-fold increase/seroconversion) and paternal indicators did not differ significantly between groups (all *p*-value >0.05), although very low birth weight (<1,500 g) showed a borderline association (6.6% vs. 0.0%, *p*-value 0.061) (Table 1).

Table 1. Prevalence of automated Red Flag indicators from validated case investigation form.

Red Flag indicators	Confirmed/probable CS (n=76; count (%))	Unlikely CS (n=67; count (%))	*P*-value
Maternal characteristics			
Maternal treatment inadequate	53 (69.7%)	12 (17.9%)	<0.001*
Mother late ANC (>12 weeks)	10 (13.2%)	17 (25.4%)	0.086
Mother very late ANC (>20 weeks)	15 (19.7%)	6 (9.0%)	0.097
Maternal RPR 4-fold increase	15 (19.7%)	6 (9.0%)	0.097
No ANC documented	16 (21.1%)	0 (0.0%)	<0.001*
Inadequate ANC follow-up (<4 visits)	16 (21.1%)	6 (9.0%)	0.062
Mother HIV positive	2 (2.6%)	2 (3.0%)	1.000
Paternal characteristics			
Father not treated or incomplete	22 (28.9%)	22 (32.8%)	0.717
Father RPR high (>1:8)	7 (9.2%)	8 (11.9%)	0.786
Father RPR failed to decrease 4-fold	2 (2.6%)	4 (6.0%)	0.419
Father HIV positive	1 (1.3%)	0 (0.0%)	1.000
Newborn outcomes			
Preterm birth	22 (28.9%)	3 (4.5%)	<0.001*
Low birth weight (<2,500 g)	11 (14.5%)	5 (7.5%)	0.288
Very low birth weight (<1,500 g)	5 (6.6%)	0 (0.0%)	0.061
Extremely low birth weight (<1,000 g)	2 (2.6%)	0 (0.0%)	0.499

**P-value <0.05. CS: congenital syphilis. ANC: antenatal care. RPR: Rapid Plasma Reagin.*

For test dataset, the system had an overall field data extraction accuracy of 94.2% (1,405/1,492 fields correct) with data field type accuracy 84.7%–99.0%. The text type had a CER of 4.32%. The model achieved overall and specific classification accuracy of 80.0% (24/30 cases) and 80.0% (24/30 cases), respectively (Table 2).

Table 2. Performance metrics of the AI-based optical character recognition (OCR) and rule-based congenital syphilis case classification system, stratified by data type with 95% confidence interval.

Metric	Sensitivity	Specificity	PPV	NPV	F1-score	Accuracy	Total (n)
AI-based OCR field extraction							
Text	97.3% (95.0–98.6)	95.5% (92.4–97.3)	96.2% (93.5–97.7)	96.8% (94.1–98.3)	96.7% (95.1–97.9)	96.5% (94.7–97.6)	621
Numerical	96.3% (90.8–98.5)	87.7% (78.2–93.4)	92.0% (85.4–95.7)	94.1% (85.8–97.7)	94.1% (89.4–96.6)	92.8% (88.0–95.7)	180
Date	94.6% (90.3–97.0)	69.0% (60.1–76.7)	82.9% (77.2–87.4)	88.9% (80.7–93.9)	88.3% (84.2–91.5)	84.7% (80.2–88.3)	300
Checkbox	99.3% (96.0–99.9)	95.2% (84.2–98.7)	98.6% (94.9–99.6)	97.6% (87.4–99.6)	98.9% (96.1–99.7)	98.3% (95.2–99.4)	181
Titer	99.2% (95.4–99.9)	98.9% (94.0–99.8)	99.2% (95.4–99.9)	98.9% (94.0–99.8)	99.2% (96.6–99.7)	99.0% (96.6–99.7)	210
Overall fields	97.2% (95.9–98.1)	90.0% (87.3–92.1)	93.4% (91.6–94.8)	95.6% (93.6–97.0)	95.2% (94.0–96.2)	94.2% (92.9–95.3)	1,492
Rule-based classification							
Specific case classification	-	-	-	-	-	80.0%	30
Overall case classification	83.3% (60.8–94.2)	75.0% (46.8–91.1)	83.3% (60.8–94.2)	75.0% (46.8–91.1)	83.3% (66.4–92.7)	80.0% (62.7–90.5)	30

AI: artificial intelligence. PPV: positive predictive value. NPV: negative predictive value.

A confusion matrix analysis indicated that the model was most successful at correctly identifying probable cases (42/55) and unlikely cases (54/67) categories. Primary sources of classification error included misidentifying confirmed cases as either probable or unlikely cases, and misclassifying non-cases as probable cases (Table 3).

Table 3. Specific classification confusion matrix (n=143).

Human \ AI	1.1 Confirmed case	1.2 Confirmed case	1.3 Confirmed case	1.5.2 Probable case	2. Unlikely case	Unclassified
1.1 Confirmed case	**5**	**2**	**1**	-	**3**	-
1.2 Confirmed case	**2**	**3**	-	**1**	-	**1**
1.3 Confirmed case	-	-	**1**	**1**	**1**	-
1.5.2 Probable case	-	**1**	-	**42**	**12**	-
2. Unlikely case	**1**	**1**	-	**11**	**54**	-

A single validated CIF can be scanned, processed via OCR, and automatically classified in under five minutes. When evaluated in a batch setting, 143 cases were fully processed in approximately 12 hours. This compared with the current manual committee review process, which averages 3–4 months from initial case report to final determination. Furthermore, the system had an API processing cost of US$ 0.006 (0.20 Thai baht) per case.

Discussion

This study successfully developed and evaluated a novel automated system using ML, specifically Gemini 2.5, to digitize and classify CS investigation forms. Addressing the critical operational gap of surveillance backlogs in Thailand's HR6, the system demonstrated a data extraction accuracy of 94.2% and an overall classification accuracy of 80.0% compared to the expert committee. These results directly fulfilled the study's objective to create a resource-efficient tool for active surveillance. By reducing the processing time from months to minutes and the cost to $0.006 per case, the system provided a scalable mechanism to bridge the gap between national elimination targets and the reality of delayed manual reporting.

This study's Red Flags revealed that inadequate maternal treatment was the most significant prevalence of confirmed/probable CS cases (69.7% vs.

17.9%, *p*-value <0.001). This aligned with findings by two recent research studies of Thailand's tertiary hospitals, which reported that untreated or inadequately treated maternal syphilis is the primary determinant of adverse pregnancy outcomes, including stillbirth and neonatal infection.[2,11] Similarly, the United States Centers for Disease Control and Prevention (U.S. CDC) data indicated that despite high ANC coverage, inadequate treatment remains a leading missed opportunity, accounting for 31% of CS cases.[12] The absence of ANC was exclusively observed in the confirmed/probable group (21.1% vs. 0.0%, *p*-value <0.001) which underscored that no ANC equates to zero opportunity for screening and treatment, making vertical transmission almost inevitable in seropositive mothers. This finding also aligned with Thailand's research.[2,11]

Interestingly, late and very late ANC were not statistically significant different in two groups. This result was seemingly counter-intuitive result to the CDC research that late prenatal care accounted for 28.2% of CS cases.[12] However, it aligned with Thailand case-control research.[11] This result may be caused by a late ANC mother can still be treated effectively if diagnosed >30 days before delivery or the small study population lacked statistical power. Also, late ANC is a well-documented risk factor for CS in Thailand, which is associated with teenage pregnancy, low education, and no family planning.[14]

The results shown that the resurgence of CS due to lack of early ANC screening and treatment was similar to recent Thailand research and guideline.[2,14] 19.7% of mothers experienced a four-fold rise in titers which indicated a significant rate of treatment failure or reinfection.[15] These Red Flags serve as immediate, actionable signals that public health officers can use to intervene before delivery, shifting surveillance from retrospective counting to prospective prevention.

The disparity between the Gemini 2.5 high extraction accuracy (94.2%) and the rule-based model algorithm classification accuracy (80.0%) offers a critical insight into the nature of clinical decision-making. While the Gemini 2.5 excelled at deciphering complex handwriting in Thai, and achieved a CER of only 4.32% similar to recent Russian historical book OCR.[16,17] The rule-based algorithm struggled to fully replicate the clinical nuance required for final classification. The sensitivity of 83.3% suggested that while it effectively applied guidelines, it lacked the tacit knowledge experts use to adjudicate ambiguous cases, such as those with borderline adequate treatment histories. This interpretation supports a triage model where the artificial intelligence (AI) functions not as a replacement for expert committees, but as a high-volume filter that can autonomously validate unlikely cases (specificity 75.0%), thereby freeing human specialists to focus solely on complex probable and confirmed cases.

The economic implications of this study are profound for resource-constrained public health systems. With an API cost of US$ 0.006 (0.20 Thai baht) per case, the AI system is orders of magnitude cheaper than the prevailing cost of expert panel reviews in Thailand, which involve high-level specialist fees and significant opportunity costs. This extreme low operating cost renders the tool sustainable for nationwide scaling, even within limited public sector budgets.

This study utilized de-identification ROIs to redact personal information before OCR. For a national rollout, this redaction must be automated and fail-safe. Furthermore, for compliance with Thailand's Personal Data Protection Act (PDPA) the use of Gemini API's paid service guaranteed zero data retention for model training which is a prerequisite for PDPA.[18,19]

Across three recent OCRs in public health studies, the core aim was similar to ours—extracting structured data from scanned clinical documents—but the target documents, models, and evaluation metrics differed. In a Thai study using Tesseract OCR for information extraction from patient registration forms, the mean extraction accuracy was 74.62% for attributes and 68.46% for values.[20] In a study of COVID-19 assessment intake forms, an Optical Mark Recognition (OMR) with OCR pipeline combined with crowd validation achieved 70% average accuracy and 78% median accuracy compared with crowd-validated results, with a mean validation time of 157 seconds per document.[21] In contrast, this study had higher both field-level OCR accuracy (overall 94.2%) and case classification performance (80.0%), reflecting a different end-use: automated digitization plus surveillance classification aligned to the congenital syphilis workflow.

Significantly, this innovation arrived at a pivotal moment as Thailand strives to revive its WHO validation for the EMTCT of syphilis. The current resurgence and surveillance backlog threatens this ambition. By offering a "Paper-to-Digital Bridge," this study provides a pragmatic solution to modernize surveillance infrastructure without requiring expensive electronic medical record overhauls at every rural facility, securing the health of the next generation.

Limitations

The study utilized a small data (n=143) from a single industrialized region (HR6). This may not capture the full diversity of handwriting styles or healthcare barriers present in rural or border provinces, potentially affecting the model's accuracy if scaled nationally without retraining.

This study did not conduct comprehensive health-economic analysis. The cost estimate included only Gemini API charges and excluded personnel time, infrastructure, software development, maintenance, storage, security, and human verification costs. Costs of the existing manual workflow were not measured, so comparative economic evaluation was not performed.

Despite strict schema enforcement, generative AI carries an inherent risk of hallucination. The model might infer dates or values in low-quality scans, necessitating a human workflow to validate all classifications to prevent data contamination.

Public Health Recommendations

This study recommends piloting the system in HR 6 as a triage and decision-support tool integrated into the existing CS verification workflow. In early implementation, the model output should be used to prioritize expert review (e.g., expedited review for model-predicted confirmed/probable cases and cases with high-risk "Red Flags"), while maintaining human verification for final classification. In parallel, the automated "Red Flag" outputs can be connected to a notification channel (e.g., LINE/SMS) to support rapid follow-up by local teams. This shifts surveillance from retrospective counting to active intervention before delivery.

Early deployment should be accompanied by routine monitoring to support quality assurance and continuous quality improvement (CQI).[22] This study recommends tracking, at minimum, the following indicators monthly: (1) OCR performance (overall field accuracy and key field error rates for dates, titers, and treatment fields), (2) classification performance against expert decisions (sensitivity, specificity, PPV, NPV, F1-score, and accuracy for binary CS-positive vs. CS-negative and for exact classification), (3) operational performance (turnaround time from CIF receipt to triage output, proportion requiring manual correction, and volume processed), and (4) discordance review (systematic review of mismatched cases to classify errors as scan quality issues, extraction errors, missing CIF information, or classification rule/feature gaps). Also, a quarterly continuous monitoring and evaluation cycle where validated cases are added to the training dataset, to fine-tune the LLM prompts and recalibrate the classification rules. These real-time performance metrics indicators can be displayed in a real-time dashboard for program monitoring (Figure 4).

Although this study did not specifically analyze migrant status, this study proposes an equity-focused extension of the pilot because HR6 includes industrial zones with high population mobility and known barriers to ANC access. After initial implementation demonstrates stable performance, the system's Red Flag indicators (late/no ANC, inadequate treatment, reinfection signals) can be used to identify service gaps and inform targeted outreach activities. Any migrant-inclusive strategy (e.g., mobile screening or linkage-to-care initiatives in industrial settings) should be evaluated as a separate operational component with explicit data collection on nationality/mobility status and access barriers policy.

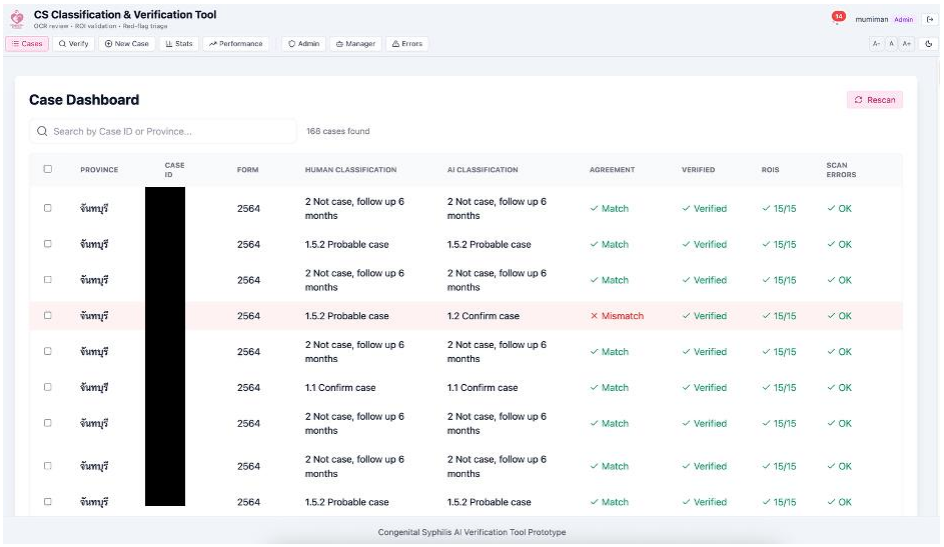


Figure 4. Machine Learning Application: Congenital Syphilis Classifier prototype dashboard.

Conclusion

This study has demonstrated that integrating a multimodal LLM into public health surveillance is a feasible, low-cost, and highly effective strategy to combat the resurgence of congenital syphilis. By automating the digitization of CIF, the system addressed the critical bottleneck of manual verification, reducing processing time from months to minutes. The findings successfully met the study objectives by characterizing the epidemic's drivers—inadequate treatment and late ANC—and providing a validated tool to detect them. While ML cannot replace clinical judgment, it serves as a powerful "Paper-to-Digital Bridge" and triage tool, visualizing invisible data to protect the health of the next generation and supporting Thailand's commitment to the elimination of mother-to-child transmission.

Acknowledgements

The authors wish to thank the Division of Epidemiology, Department of Disease Control, Ministry of Public Health, Thailand, for providing the CIFs and facilitating data access for this research.

Author Contributions

Suppasit Srisaeng: Conceptualization, methodology, software, validation, investigation, writing—original draft. **Chanissara Thuwakum:** Project administration, data curation, writing—review & editing.

Ethical Approval

This study was conducted as a retrospective analysis of de-identified secondary data obtained from the routine congenital syphilis surveillance system, managed by the Division of Epidemiology, Department of Disease Control. To safeguard privacy, all personally identifiable information was redacted from case investigation forms prior to machine learning processing, ensuring the data could not be linked back to individual subjects.

The research presented results in an aggregated format, posed no more than minimal risk to the subjects, and the waiver of consent did not adversely affect their rights, welfare, or clinical care. Consequently, a waiver of informed consent and an exemption from formal ethical review are applicable to this study.

Informed Consent

Patient consent was waived due to the retrospective nature of the study, which utilized de-identified secondary data from the national congenital syphilis response system. The research was conducted for public health surveillance improvement purposes under the authority of the DDC, and the analysis presented no more than minimal risk to the subjects. All data were processed in compliance with the Thailand PDPA.

Data Availability

Research data are not shared.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding Support

This research received no funding.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the authors used Google Gemini to enhance clarity, refine the structure of the discussion, and correct grammatical errors. The content produced by this tool was reviewed and edited by the authors, who accept full responsibility for the final text.

References

1. World Health Organization. Sexually transmitted infections (STIs) [Internet]. Geneva: World Health Organization; 2025 Sep 10 [cited 2025 Dec 5]. <https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis)>
2. Luengmettakul J, Apiwantana S, Jitrungruengnij N. Incidence of congenital syphilis and adverse pregnancy outcomes among syphilitic pregnant women according to the treatment adequacy in Tertiary Care Hospital, Thailand. Journal of the Medical Association of Thailand. 2025;108(1):9–16.
3. World Health Organization. Mother-to-child transmission of syphilis [Internet]. Geneva: World Health Organization; [cited 2025 Dec 5]. <https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/stis/prevention/mother-to-child-transmission-of-syphilis>
4. World Health Organization. Elimination of mother-to-child transmission of HIV, syphilis and hepatitis B [Internet]. Geneva: World Health Organization; [cited 2025 Dec 20]. <https://www.who.int/southeastasia/activities/elimination-of-mother-to-child-transmission-of-hiv-syphilis-and-hepatitis-b-virus>

5. Division of AIDS and STIs, Department of Disease Control (TH). STIs situation for executive [Internet]. Nonthaburi: Department of Disease Control; [cited 2025 Dec 5]. <https://hivhub.ddc.moph.go.th/executive/sti.php>

6. Division of Epidemiology, Department of Disease Control (TH). Guidelines for surveillance and investigation of congenital syphilis in children under 2 years old [Internet]. Nonthaburi: Department of Disease Control; 2021 Sep 1. 20 p. Thai.

7. Chen H, Zeng D, Qin Y, Fan Z, Ng Yu Ci F, Klonoff DC, et al. Large language models and global health equity: a roadmap for equitable adoption in LMICs. Lancet Reg Health West Pac. 2025;63:101707. doi:10.1016/j.lanwpc.2025.101707.

8. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927;22(158):209–12. doi:10.2307/2276774.

9. Abdoli F. Character error rate (CER): a friendly, no-nonsense guide [Internet]. Cologne (DE): WaterCrawl Blog; 2025 Aug 26 [cited 2026 Jan 24]. <https://watercrawl.dev/blog/Character-Error-Rate>

10. Google. Gemini Developer API pricing [Internet]. Mountain View (CA): Google LLC; [cited 2025 Dec 21]. <https://ai.google.dev/gemini-api/docs/pricing#gemini-2.5-pro>

11. Kulsirichawaroj P, Lumbiganon D. Incidence and associated factors of congenital syphilis at a tertiary care center in Thailand. Asian Biomed (Res Rev News). 2023 Feb;17(1):13–21. doi:10.2478/abm-2023-0039.

12. Kimball A, Torrone E, Miele K, Bachmann L, Thorpe P, Weinstock H, et al. Missed Opportunities for Prevention of Congenital Syphilis — United States, 2018. MMWR Morb Mortal Wkly Rep. 2020 Jun 5;69(22):661–5. doi:10.15585/mmwr.mm6922a1.

13. Soontornprakasit P, Mongkolchati A, Chompikul J. Factors associated with time to start antenatal care within 12 weeks gestational age among mothers in Mahasarakham province, Thailand. Journal of Public Health and Development. 2016 May;14(1):21–36.

14. Chayachinda C. Elimination of congenital syphilis in Thailand: what can bedone during antenatal period? Thai Journal of Obstetrics and Gynaecology. 2016;24(2):66–72. doi:10.14456/tjog.2016.16.

15. Treger RS, Menza TW, Truong TT, Lieberman JA. Advances in syphilis diagnostics to address the 21st-Century Epidemic. Clin Chem. 2025; 71(9):935–48. doi:10.1093/clinchem/hvaf072.

16. Nonesung S, Jaknamon T, Chaiophat S, Nitarach N, Wittayasakpan C, Sirichotedumrong W, et al. ThaiOCRBench: A Task-Diverse Benchmark for Vision-Language Understanding in Thai. arXiv:2511.04479 [Preprint]. 2025 [cited 2026 Jan 24]:[23 p.]. <https://arxiv.org/abs/2511.04479>

17. Levchenko M. Evaluating LLMs for Historical Document OCR: A Methodological Framework for Digital Humanities. arXiv: 2510.06743 [Preprint]. 2025 [cited 2026 Jan 24]:[12 p.]. <https://arxiv.org/abs/2510.06743>

18. Ministry of Digital Economy and Society (TH). Personal Data Protection Act, B.E. 2562 (2019) [Internet]. Bangkok: Ministry of Digital Economy and Society; [cited 2025 Dec 21]. 35 p. <https://www.mdes.go.th/law/detail/3577-Personal-Data-Protection-Act-B-E--2562--2019->

19. Google. Gemini API additional terms of service [Internet]. Mountain View (CA): Google LLC; 2025 [cited 2025 Dec 21]. <https://ai.google.dev/gemini-api/terms#paid-services>

20. Chumwatana T., Rattana-Amnuaychai W., Chauychu P. Patient information extraction using optical character. Journal of the Thai Medical Informatics Association [Internet]. 2022 Jun 8 [cited 2026 Jan 24];8(1):22–7. <https://he03.tci-thaijo.org/index.php/jtmi/article/view/198>

21. White-Dzuro CG, Schultz JD, Ye C, Coco JR, Myers JM, Shackelford C, et al. Extracting medical information from paper COVID-19 assessment forms. Appl Clin Inform. 2021; 12(01):170–8. doi:10.1055/s-0041-1723024.

22. Endalamaw A, Khatri RB, Mengistu TS, Erku D, Wolka E, Zewdie A, et al. A scoping review of continuous quality improvement in healthcare system: conceptualization, models and tools, barriers and facilitators, and impact. BMC Health Serv Res. 2024 Apr 19;24(1):487. doi:10.1186/s12913-024-10828-0.