

Development and Multicenter Validation of an Ensemble Machine Learning Model for Symptom-Based Triage of Hepatopancreatobiliary Diseases

Rahul Rahul¹

rahulimsbhu@gmail.com

Sanjay Gandhi Post Graduate Institute of Medical Sciences

Rajan Saxena¹

Sanjay Gandhi Post Graduate Institute of Medical Sciences

Saumya Shukla²

Sanjay Gandhi Post Graduate Institute of Medical Sciences

Kush Satyen Parikh¹

Sanjay Gandhi Post Graduate Institute of Medical Sciences

Sourodeep Mandal³

King George's Medical University

Shaganti Rakesh⁴

Dr. Ram Manohar Lohia Institute of Medical Sciences

Ankit Shukla⁵

Duke-NUS Medical School

Nikhil Misra⁶

Indian Institute of Technology Kanpur

Article

Keywords: Hepatopancreatobiliary diseases, machine learning, symptom-based triage, Symptomatology Index, clinical decision support, primary care

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10927874/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Hepatopancreatobiliary (HPB) diseases are frequently diagnosed late because early symptoms are non-specific and standardized triage tools are limited. We developed and validated a symptom-based machine-learning clinical decision-support model for timely identification and referral of patients at risk of HPB disease. In this ambispective study, 2,296 patients were screened and 2,108 with complete data formed the development cohort; a separate, non-overlapping independent validation cohort of 781 patients (469 HPB-positive and 312 controls) was used for validation. The ensemble model achieved an AUC-ROC of 0.96 (95% CI 0.94–0.98), sensitivity of 76.1%, specificity of 93.9%, accuracy of 83.2%, positive predictive value of 94.9%, negative predictive value of 72.3%, and F1 score of 84.5%. Sensitivity was 14.1 percentage points higher than specialist clinical assessment (76.1% vs 62.0%), corresponding to an approximately 23% relative improvement. A bilingual mobile application enabled real-time symptom-based risk stratification. These findings support further prospective evaluation of the model as a low-cost triage aid, particularly in primary-care and resource-limited settings.

Introduction

Hepatopancreatobiliary diseases include benign and malignant disorders in which non-specific early symptoms can delay diagnosis and referral [1–6]. Definitive assessment often depends on imaging, endoscopic sampling, pathology, and biomarkers that may not be immediately available at first contact [7–11]. Machine-learning approaches can integrate routinely collected clinical variables for risk stratification [12–14]. Existing tools such as the Cancer of the Liver Italian Program (CLIP) score for hepatocellular carcinoma and the Genç recurrence score for pancreatic neuroendocrine tumors address prognosis or recurrence rather than symptom triage [15, 16]. Placido et al. reported an AUC of 0.88 for pancreatic-cancer prediction from longitudinal disease trajectories [17]. We therefore developed and validated an ensemble model for symptom-based HPB triage.

Results

Clinical profile of the patient cohort

The screened cohort comprised 2,296 individuals (1,134 [49.4%] male and 1,162 [50.6%] female; reported median age, 47 years). Baseline demographic and clinical characteristics of the screened cohort are summarized in Table 1.

Fifty-four individuals younger than 18 years were excluded according to the study eligibility criteria. Among the screened cohort, 892 (38.9%) were aged 18–40 years, 912 (39.7%) were aged 40–65 years, and 438 (19.1%) were older than 65 years. Tobacco or alcohol use was reported by 17.5% of screened participants and diabetes mellitus by 16.2%. Abdominal pain was the most common symptom (1,922/2,296; 83.7%); right upper quadrant, central, and generalized abdominal pain were reported in 37.3%, 27.6%, and 18.8% of the screened cohort, respectively. Other symptoms included vomiting

(33.5%), fever (28.9%), melena (6.9%), abdominal mass (5.5%), jaundice (4.7%), night sweats (1.7%), and overt gastrointestinal bleeding (0.3%).

Of the 2,108 participants with complete development-cohort data, 1,268 had confirmed HPB pathologies and 840 served as negative controls. Within the HPB-positive development cohort, 448 cases were enrolled prospectively and 820 retrospectively. Disease subtypes included 523 malignancies (41.2%), predominantly gallbladder cancer and periampullary carcinoma, 318 cases of gallstone disease (25.1%), 229 cases of common bile duct stones (18.1%), and 198 cases of chronic liver disease (15.6%). A separate, non-overlapping independent validation cohort comprised 781 patients, including 469 HPB-positive cases (188 malignancies, 102 gallstone disease cases, 89 common bile duct stone cases, and 90 chronic liver disease cases) and 312 controls. Hepatocellular carcinoma was assessed separately in 90 patients; 16 (17.8%) received curative treatment and 74 (82.2%) had metastatic or unresectable disease.

Screening and Study Cohort Characteristics

Table 1
Screening and clinical characteristics of the study cohorts

Gender distribution	Number (%)
Male	1,134 (49.4%)
Female	1,162 (50.6%)
Total	2,296 (100%)
Age groups	
< 18 years*	54 (2.4%)
18–40 years	892 (38.9%)
40–65 years	912 (39.7%)
> 65 years	438 (19.1%)
Total	2,296 (100%)
Personal history (n = 2,296)†	
Tobacco/alcohol use	402 (17.5%)
Diabetes mellitus	373 (16.2%)
Abdominal pain subtypes (n = 1,922)	
Right upper quadrant pain	856 (44.5%)
Central abdominal pain	634 (33.0%)
Generalized abdominal pain	432 (22.5%)
Total	1,922 (100%)
Other symptoms (n = 2,296)	
Vomiting	769 (33.5%)
Fever	663 (28.9%)
Melena	159 (6.9%)
Abdominal mass	127 (5.5%)
Jaundice	109 (4.7%)
Night sweats	40 (1.7%)
Overt GI bleeding	8 (0.3%)
Primary HPB pathologies cohort (n = 2,108)	
Prospective cases	448 (21.2%)

Gender distribution	Number (%)
Retrospective cases	820 (38·8%)
Negative controls	840 (40·0%)
Total	2,108 (100%)
Disease distribution (n = 1,268)	
Malignancies (GBC, PACA)	523 (41·2%)
Gallstone disease (GSD)	318 (25·1%)
CBD stones	229 (18·1%)
Chronic liver disease (CLD)	198 (15·6%)

Total	1,268 (100%)
Independent validation cohort (n = 781)	
HPB-positive cases	469 (60·0%)
Normal controls	312 (40·0%)
Disease distribution – HPB-positive cases (n = 469)	
Malignancies (GBC, PACA)	188 (40·1%)
Gallstone disease (GSD)	102 (21·7%)
CBD stones	89 (19·0%)
Chronic liver disease (CLD)	90 (19·2%)
Total	469 (100%)
Hepatocellular carcinoma (HCC) subset (n = 90)	
Curative treatment	16 (17·8%)
Metastatic/unresectable	74 (82·2%)
Total	90 (100%)

*Excluded from study per inclusion criteria.

†Patients may have more than one comorbidity; values are not mutually exclusive and do not sum to the cohort total. Percentages may not sum to 100% due to rounding.

CBD=common bile duct. CLD=chronic liver disease. GBC=gallbladder cancer. GI=gastrointestinal. GSD=gallstone disease. HCC=hepatocellular carcinoma. HPB=hepatopancreatobiliary. NA = not applicable. PACA=periampullary carcinoma.

Multivariate model to predict hepatopancreatobiliary diseases

In the development cohort, the ensemble model was trained and internally evaluated using stratified five-fold cross-validation with hyperparameter optimization as described in the Methods. Performance of the final model was then assessed in the independent validation cohort.

In the independent validation cohort ($n = 781$), comprising 469 patients with confirmed HPB disease and 312 controls, the ensemble model correctly classified 357 HPB-positive patients and 293 controls, with 112 false-negative and 19 false-positive classifications. Sensitivity was 76.1% (95% CI 72.0–79.9%), specificity 93.9% (90.7–96.3%), accuracy 83.2% (80.4–85.8%), PPV 94.9% (92.2–96.9%), NPV 72.3% (67.7–76.6%), and F1 score 84.5%. The false-negative rate was 23.9% (112/469) and the false-positive rate was 6.1% (19/312). The AUC-ROC was reported as 0.96 (95% CI 0.94–0.98). Validation performance metrics are shown in Fig. 1.

Disease-specific algorithm performance

Gradient Boosting achieved the best performance for CCA/GBC, with an accuracy of 0.88, sensitivity of 0.89, specificity of 0.88, and PPV of 0.89. For PACA, SVM achieved an accuracy of 0.87, sensitivity of 0.55, specificity of 0.94, and PPV of 0.87, indicating that approximately 45% of PACA cases were not identified using clinical features alone. For GSD, SVM achieved an accuracy of 0.88, sensitivity of 0.96, specificity of 0.69, and PPV of 0.70. AdaBoost showed the best performance for both CBD stones and CLD, achieving an accuracy of 0.86, sensitivity of 1.00, specificity of 0.75, and PPV of 0.75 for each condition, as shown in Fig. 2.

Performance comparison between the machine learning model and clinical assessment by surgical gastroenterologists

To evaluate clinical utility, the ensemble model was compared against surgical gastroenterologist (SGE) assessment in the 781-patient validation cohort (469 HPB, 312 controls), with ground truth from CT or pathological confirmation.

SGE clinical assessment had a reported sensitivity of 62.0% and PPV of 73.0%. The ML model achieved a sensitivity of 76.1% (95% CI 72.0–79.9%) and PPV of 94.9% (92.2–96.9%). This represents a 14.1-percentage-point absolute increase in sensitivity, corresponding to an approximately 22.8% relative improvement. PPV increased by 21.9 percentage points. The ML model correctly identified 357 of 469 HPB-positive patients, with a false-positive rate of 6.1% (19/312).

Sensitivity was 0.68 with SGE assessment compared with 0.89 with Gradient Boosting for GBC, representing a 21-percentage-point absolute increase in sensitivity (30.9% relative increase). For PACA, SGE assessment showed a sensitivity of 0.41 (specificity 0.78), whereas the SVM was reported to achieve a precision of 1.00 and specificity of 0.94, reducing the false-positive rate from 22% to 6%, an

absolute reduction of 16 percentage points. For benign conditions (GSD and CBD stones), SGE sensitivities were 0.91 and 0.95, respectively, and were comparable with the reported ML-model sensitivities. Receiver operating characteristic curves for the disease-specific classification models are shown in Fig. 3.

Impact on time-to-diagnosis and resectability

Separate disease-specific diagnostic-delay analyses were reported for GBC, PACA, and HCC. For GBC, mean diagnostic delay was 5.28 months among patients undergoing curative resection versus 7.54 months among patients with metastatic disease (mean difference, 2.26 months; $p < 0.001$; $n = 120$ curative and $n = 85$ metastatic). For PACA, mean diagnostic delay was 5.54 months in the curative group versus 7.27 months in the advanced-disease group (mean difference, 1.73 months; $p < 0.001$; $n = 110$ curative and $n = 75$ advanced). No significant difference in diagnostic delay was reported for HCC (5.08 vs 5.75 months; $p = 0.21$).

Within the reported overall malignancy cohort ($n = 250$), 77 patients (30.8% of the overall cohort) referred within four months and 53 patients (21.2% of the overall cohort) referred later underwent curative resection. When expressed using the overall cohort denominator, these proportions differ by 9.6 percentage points. Curative resection counts by referral timing are shown in Fig. 4.

Figure 4. Reported curative resection counts by referral timing in the overall HPB malignancy cohort. Bars show the numbers of patients undergoing curative resection among those referred within four months ($n = 77$) and those referred later ($n = 53$), as reported for the overall malignancy cohort ($n = 250$).

Formulating a clinical app for standardization of clinical assessment of HPB diseases

To facilitate implementation in primary healthcare centers (PHCs), the ensemble model was integrated into a clinical decision support application. Physicians enter patient symptoms, demographic information, and comorbidities, after which the application generates real-time risk scores and referral recommendations through four sequential steps: structured data entry, machine learning analysis, symptom-based risk stratification, and automated referral recommendations

(low, intermediate, or high risk). In the intended application workflow, symptom scores ≤ 5 were categorized as low risk, whereas scores > 5 prompted further diagnostic evaluation. The clinical workflow is shown in Fig. 5.

In the independent validation cohort, the model achieved a sensitivity of 76.1%, specificity of 93.9%, and accuracy of 83.2%. Separately, internal stratified cross-validation of the application was reported to yield an accuracy of 92.57%. Disease-specific cross-validated performance included sensitivity of 0.8871 (accuracy 86.36%) for CCA/GBC using Gradient Boosting, sensitivity of 0.8427 (accuracy 89.04%) for

PACA using SVM, accuracy of 87.90% for gallstone disease, and accuracy of 80.70% for common bile duct stones.

Discussion

The symptom-based ensemble ML model showed high specificity (93.9%) and moderate sensitivity (76.1%) in the independent validation cohort, with a reported AUC-ROC of 0.96. These findings support the potential of structured symptom-based triage to prioritize HPB evaluation, while also indicating that a substantial proportion of HPB-positive cases remained model-negative.

A major strength of this study is the inclusion of patients from multiple tertiary-care centers, improving the diversity of the study population and enhancing the generalizability of the findings.

In the independent validation cohort, model sensitivity was 76.1% compared with 62.0% for SGE assessment, a 14.1-percentage-point absolute increase corresponding to an approximately 22.8% relative improvement. The observed timing analysis suggests that shorter diagnostic intervals may be associated with greater opportunity for curative treatment; this observational association should not be interpreted as evidence that the ML model itself caused improved resectability.

This model differs from many AI approaches that rely on biochemical, imaging, genomic, or longitudinal electronic health-record data by focusing on structured symptomatology potentially available at first clinical contact [12–14]. Topol highlighted the potential of AI to augment clinical decision-making [13]. Placido et al. reported an AUC of 0.88 for prediction of pancreatic cancer from longitudinal disease trajectories [17], whereas Banerjee et al. used free-text clinical narratives to predict short-term survival in metastatic cancer, achieving an AUC of 0.89 [18]. The present study therefore explores a distinct use case: symptom-based HPB triage using routinely obtainable clinical features.

The bilingual application converts model outputs into real-time risk categories and referral recommendations. Such structured decision support may be relevant in low- and middle-income settings where access to specialist radiological and endoscopic diagnostics is variable, consistent with broader WHO priorities for responsible AI in health [19].

Disease-specific algorithms showed heterogeneous operating characteristics: CCA/GBC classification was more sensitive than PACA classification in the reported disease-specific analysis, whereas the models for CBD stones and CLD were reported to have high sensitivity. These differences should inform future threshold optimization and prospective validation rather than be interpreted as definitive rule-in or rule-out performance.

The higher reported sensitivity of the ML model relative to specialist assessment may reflect systematic integration of multiple clinical features, including persistent upper abdominal pain, weight loss, and new-onset diabetes, which are associated with pancreatic cancer risk [4–6]. However, the comparison was

performed within the same tertiary-care validation cohort and should be interpreted as evidence of potential adjunctive value rather than proof of clinical superiority.

The intended use of the application also differs from established disease-specific prognostic tools. The CLIP score was developed for prognosis in hepatocellular carcinoma [15], whereas the Genç score predicts recurrence after resection of grade 1–2 nonfunctional pancreatic neuroendocrine tumors [16]. In contrast, the present application targets symptom-based triage before definitive HPB diagnosis; prospective validation in primary-care populations is therefore required before routine implementation.

Several limitations should be considered. The development and validation data were obtained from tertiary referral centers, which may introduce spectrum and referral bias and limits direct generalization to primary-care populations. The ambispective design and complete-case exclusions may introduce selection bias. If model-negative patients did not undergo equivalent reference-standard verification, diagnostic accuracy estimates may be susceptible to partial verification bias. Calibration, decision-curve analysis, and prospective impact on referral or resection outcomes were not established in the current study. Future work should include prospective primary-care validation, prespecified threshold evaluation, calibration assessment, and transparent reporting of the complete Symptomatology Index.

Methods

Study design and setting

This ambispective multicenter cohort study developed and validated an ensemble ML model for symptom-based HPB disease triage. Data were collected from January 2, 2022 to March 1, 2025, integrating historical electronic medical records and prospective clinical data. The study was approved by the EXP – 51st Institutional Ethics Committee of Sanjay Gandhi Postgraduate Institute of Medical Sciences (SGPGIMS), Lucknow, India (IEC code: 2023–26-IP-EXP-51; approval reference: PGI/BE/220/2023). Written informed consent was obtained from prospectively enrolled participants; retrospective records were included under an IEC (EXP-51st Institutional Ethics Committee) approved waiver of informed consent ensuring compliance with ethical norms set by the ethics committee. The study methodologies adhered to applicable guidelines and regulations. The radiologists and researchers involved in the study only had access to non-identifiable patient data for analysis purposes. In contrast, all identifiable patient screening and diagnostic information remained securely stored on an internally protected server accessible only through credential-based authentication.

Clinical variables and Symptom Index

A structured Symptomatology Index was designed with weighted scores assigned to clinical features according to their predictive contribution to HPB pathology. Higher-weight features included persistent right upper quadrant pain, progressive weight loss, jaundice, and new-onset diabetes; lower-weight features included intermittent gastrointestinal symptoms and fatigue. A total score > 5 was classified as

model-positive, whereas a score ≤ 5 was classified as low risk/model-negative. The threshold was empirically determined in the development data.

The Symptomatology Index was interpreted using the clinical decision-support algorithm shown in Fig. 6. For model evaluation, scores > 5 were classified as model-positive and scores ≤ 5 as model-negative. Reference-standard disease status was used to define diagnostic outcomes: true positive, model-positive with HPB disease present; false positive, model-positive with HPB disease absent; false negative, model-negative with HPB disease present; and true negative, model-negative with HPB disease absent. In the intended clinical workflow, model-positive patients proceeded to confirmatory imaging, laboratory or molecular testing, endoscopic assessment, and/or biopsy as clinically indicated.

Study cohort

The screening cohort included 2,296 individuals with HPB-suggestive symptoms between January 2, 2022 and March 1, 2025. Participants were enrolled according to the study eligibility criteria. Of these, 2,108 had complete data and formed the development cohort (1,268 HPB-confirmed and 840 controls); 188 were excluded because of incomplete records, a low symptom score, or an indeterminate final diagnosis, as shown in Fig. 7. A separate, non-overlapping independent validation cohort included 781 patients (469 HPB-positive and 312 controls).

Eligible participants were adults (≥ 18 years) with a confirmed HPB diagnosis (CLD, GSD, CBD stones, CCA, GBC, PACA, PC, or HCC) or an appropriate disease-negative reference outcome, together with a verified symptom profile. Exclusion criteria were incomplete records, symptom score ≤ 5 for the development-cohort selection described above, indeterminate diagnosis, or concomitant extra-HPB malignancy. Patient-reported questionnaires were linked to electronic health records, and retrospective and prospective cases were combined in the development cohort.

Sample size considerations

The development sample comprised all eligible complete records available during the study period; no formal a priori model-development sample-size calculation was reported. The independent validation cohort included 469 HPB-positive patients and 312 controls. At the observed validation sensitivity of 76.1% and specificity of 93.9%, the exact 95% confidence intervals were 72.0–79.9% and 90.7–96.3%, respectively, providing an explicit measure of validation precision.

Data preprocessing and quality assurance

The development cohort ($n = 2,108$; 1,268 HPB-positive and 840 controls) and independent validation cohort ($n = 781$; 469 HPB-positive and 312 controls) underwent data preprocessing. Hash-based identifiers were used to identify and remove duplicate records; range violations and temporal inconsistencies were reviewed and corrected.

Continuous variables were standardized to z-scores and categorical variables were one-hot encoded. Feature engineering produced binary or ordinal variables and composite symptom clusters (for example, a hepatic triad of jaundice, pain, and fever). Diagnostic-delay variables were reserved for separate clinical-utility analyses and were not used as predictors in the diagnostic model. Quality assurance included dual review of 10% of records, with ambiguous cases adjudicated by consensus among three senior surgical gastroenterologists.

Machine learning model architecture

The ensemble method was a combination of Gradient Boosting, AdaBoost, and Support Vector Machines (SVM). Gradient Boosting combines nonlinear feature interactions by sequential weak learners; AdaBoost applies adaptive weighting to diagnostically ambiguous cases; SVMs utilize radial basis function kernels to model nonlinear decision boundaries. Imbalance (1,268 positive vs 840 controls) was handled by class-weighted learning without synthetic oversampling.

Each classifier was evaluated using stratified five-fold cross-validation, with approximately 80% of the development data used for model fitting and 20% held out in each outer fold. Hyperparameters were tuned within the training partition using nested grid search (learning rate and tree depth for Gradient Boosting, number of estimators for AdaBoost, and C and gamma for SVM). Final ensemble predictions were generated by soft voting, averaging class probabilities from the three base classifiers to balance sensitivity and specificity.

Performance metrics included sensitivity, specificity, accuracy, PPV, NPV, F1 score, and AUC-ROC. Generalizability was assessed in the independent multicenter validation cohort of 781 patients. Models were implemented in Python 3.9 using scikit-learn 1.0.2 on a high-performance computing cluster. Figure 7 summarizes data preprocessing, model training, hyperparameter optimization, ensemble integration, internal cross-validation, and independent validation.

Disease-specific models and subgroup analyses

We developed disease-specific classifiers for each major HPB subtype using stratified, condition-specific datasets to support direct algorithm comparisons in disease-specific contexts.

Clinical utility and time-to-diagnosis analysis

Clinical impact was explored using an overall malignancy cohort reported as n = 250 for the early-versus-delayed referral analysis. Diagnostic delay was defined as the interval between symptom onset and definitive diagnosis and was categorized as early (≤ 4 months) or delayed (> 4 months). The four-month cutoff was based on literature describing adverse outcomes associated with diagnostic or treatment delays beyond approximately 120 days [20, 21]. Separate disease-specific analyses were conducted for CCA/GBC, PACA/PC, and HCC.

Continuous diagnostic-delay variables were compared between independent timing or outcome groups using independent-samples t-tests when distributional assumptions were met; categorical outcomes such as resection status were compared using chi-square or Fisher exact tests as appropriate. Separate subgroup analyses were conducted for CCA/GBC, PACA, and HCC.

Statistical analysis

Diagnostic performance was assessed using AUC-ROC as the primary discrimination metric together with sensitivity, specificity, accuracy, PPV, NPV, and F1 score. Exact binomial 95% confidence intervals were calculated for proportion-based performance metrics in the development and validation cohorts separately.

Because ML and SGE assessments were obtained on the same validation participants, paired binary classification differences were assessed using McNemar's test. Time-to-diagnosis analyses used independent-samples t-tests with Cohen's d reported as an effect-size measure. Disease-specific analyses were considered exploratory. Analyses were performed using R 4.2.3 and Python 3.9. Statistical tests were two-sided with $\alpha = 0.05$; nominal disease-specific p values were interpreted cautiously because no multiplicity adjustment was applied.

Declarations

Use of AI-assisted tools in manuscript preparation

ChatGPT (OpenAI) was used during manuscript revision to assist with language editing, consistency checking, and verification of arithmetic derived from reported aggregate counts. Scientific interpretation, verification of the underlying study data, and responsibility for the final manuscript remain with the authors.

Code availability

Custom code and the trained algorithm are available from the corresponding author on reasonable request, including for editorial and peer-review purposes.

Role of the funding source

There was no funding source for this study.

Author contributions

RS contributed to Conceptualization, Methodology, and Investigation. RR contributed to Conceptualization, Methodology, Investigation, and Validation, and directly accessed and verified the underlying data. SS contributed to Data Curation, Methodology, and Supervision. KSP contributed to Conceptualization, Methodology, and Investigation. SM contributed to Conceptualization, Methodology, and Investigation. SR contributed to Conceptualization, Methodology, and Investigation. AS contributed

to Software, Methodology, Validation, and Writing – Original Draft, and directly accessed and verified the underlying data. NM contributed to Formal Analysis, Visualization, and Writing – Original Draft.

Competing interests

The authors declare no competing interests.

Data availability

De-identified participant data and the data dictionary will be available upon reasonable request to the corresponding author after publication, subject to institutional data governance approval and a signed data access agreement. The study protocol and statistical analysis plan are available from the corresponding author upon request.

Acknowledgements

We thank Laveena Balani, Sreenivas Reddy Biravelly, Katepogu Prashanth, Deepak Thorat, Sunil Kumar Godara, Palash Rawlani, Yajnadatta Sarangi, Ankit Sahoo, Kanhaiya Lal Chaudhary, Raghavendra Ballal R, Divya Jain, Samiksha Rajvanshi, Sarthak Chaudhary, Sharukh Rauf, Suhail Batt, Sourish Mukherji, Juvin G, Pradeep, Anirudh Goyal, Rakesh Roshan, Sathvik Alladi, Payal Kaw, and Katalam Sai Krishna for clinical contributions to this study. We also thank Mohd Kaif for assistance in preparing the manuscript.

References

1. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2022. *CA Cancer J. Clin.* **72**, 7–33 (2022).
2. Valle, J. W., Kelley, R. K., Nervi, B., Oh, D. Y. & Zhu, A. X. Biliary tract cancer. *Lancet* **397**, 428–444 (2021).
3. Hundal, R. & Shaffer, E. A. Gallbladder cancer: epidemiology and outcome. *Clin. Epidemiol.* **6**, 99–109 (2014).
4. Stapley, S. et al. The risk of pancreatic cancer in symptomatic patients in primary care: a large case-control study using electronic records. *Br. J. Cancer.* **106**, 1940–1944 (2012).
5. Becker, A. E., Hernandez, Y. G., Frucht, H. & Lucas, A. L. Pancreatic ductal adenocarcinoma: risk factors, screening, and early detection. *World J. Gastroenterol.* **20**, 11182–11198 (2014).
6. Rawla, P., Sunkara, T. & Gaduputi, V. Epidemiology of pancreatic cancer: global trends, etiology and risk factors. *World J. Oncol.* **10**, 10–27 (2019).
7. Poruk, K. E. et al. The clinical utility of CA 19 – 9 in pancreatic adenocarcinoma: diagnostic and prognostic updates. *Curr. Mol. Med.* **13**, 340–351 (2013).
8. Brancato, L. et al. CA19-9 as a dynamic biomarker for continuous monitoring of therapeutic efficacy in pancreatic adenocarcinoma. *Cancers (Basel).* **17**, 3902 (2025).

9. Miller, L. J. et al. Performance of fluorescence in situ hybridization in biliary brushings with equivocal cytology: an institutional experience. *J. Am. Soc. Cytopathol.* **13**, 285–290 (2024).
10. Nakai, Y., Hakuta, R., Shimamatsu, Y., Otsuka, N. & Takayama, Y. Endoscopic approach to indeterminate biliary strictures. *Clin. Endosc.* **59**, 40–48 (2026).
11. Virgilsen, L. F., Rasmussen, L. A., Vedsted, P. & Jensen, H. Diagnosing pancreatic cancer in general practice: a cross-sectional study on associations between suspicion of cancer, urgent referral and time to diagnosis. *Scand. J. Prim. Health Care.* **40**, 78–86 (2022).
12. Mishra, A. K. et al. Machine learning models for pancreatic cancer risk prediction using electronic health record data: a systematic review and assessment. *Am. J. Gastroenterol.* **119**, 1112–1123 (2024).
13. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
14. Chen, J. H. & Asch, S. M. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N Engl. J. Med.* **376**, 2507–2509 (2017).
15. Cancer of the Liver Italian Program (CLIP). Investigators. A new prognostic system for hepatocellular carcinoma: a retrospective study of 435 patients. *Hepatology* **28**, 751–755 (1998).
16. Genç, C. G. et al. A new scoring system to predict recurrent disease in grade 1 and 2 nonfunctional pancreatic neuroendocrine tumors. *Ann. Surg.* **267**, 1148–1154 (2018).
17. Placido, D. et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nat. Med.* **29**, 1113–1122 (2023).
18. Banerjee, I. et al. Probabilistic prognostic estimates of survival in metastatic cancer patients (PPES-Met) utilizing free-text clinical narratives. *Sci. Rep.* **8**, 10037 (2018).
19. World Health Organization. *Ethics and governance of artificial intelligence for health* (World Health Organization, 2021). <https://www.who.int/publications/i/item/9789240029200> (accessed 25 August 2026).
20. Sanjeevi, S. et al. Impact of delay between imaging and treatment in patients with potentially curable pancreatic cancer. *Br. J. Surg.* **103**, 267–275 (2016).
21. Geng, C. X. et al. Effects of treatment delays on outcomes in pancreatic adenocarcinoma. *Ann. Pancreas Cancer.* **7**, 8 (2024).

Figures

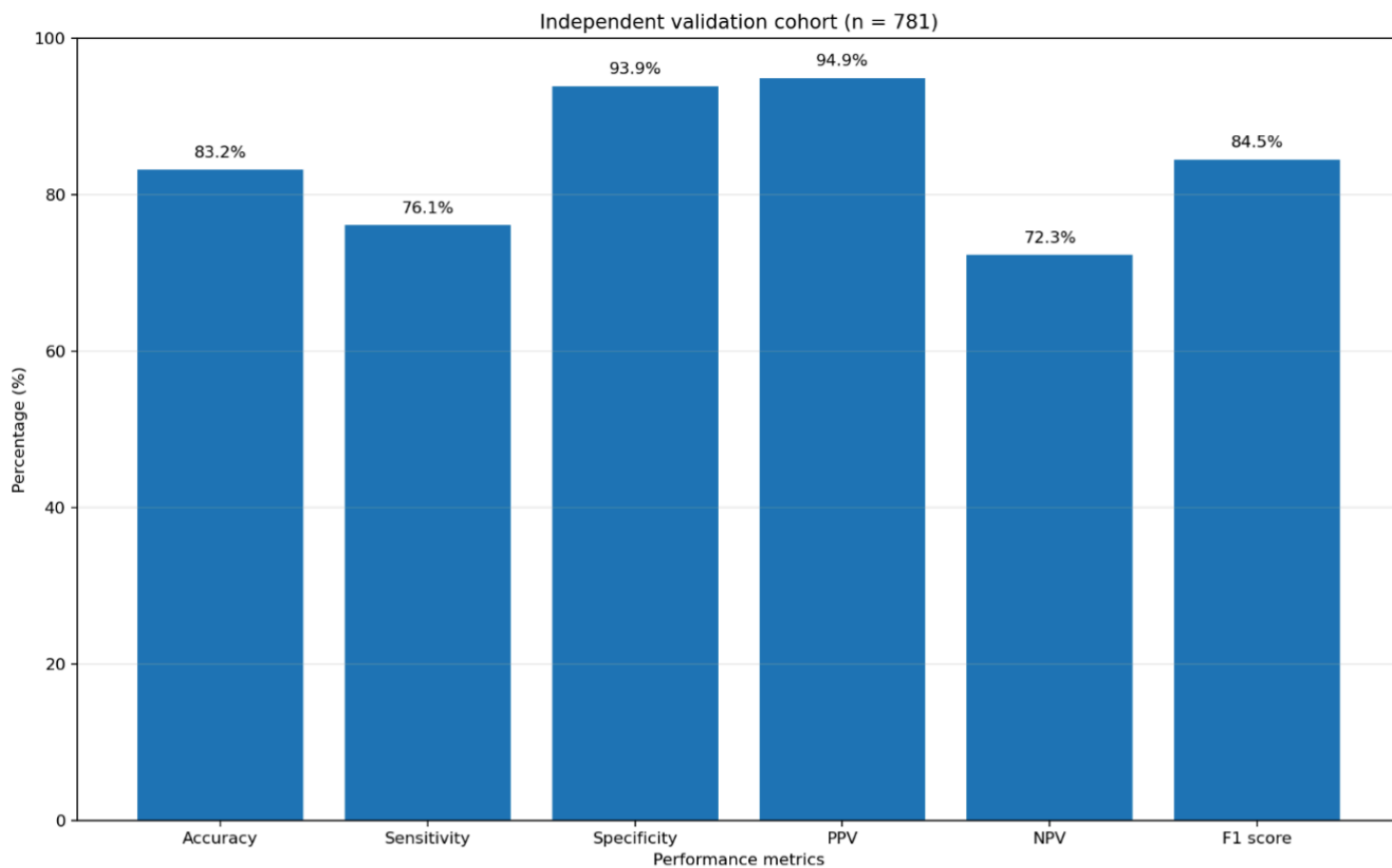


Figure 1

Performance of the ensemble machine-learning model in the independent validation cohort. Bars represent accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score derived from the confirmed validation counts (n = 781).

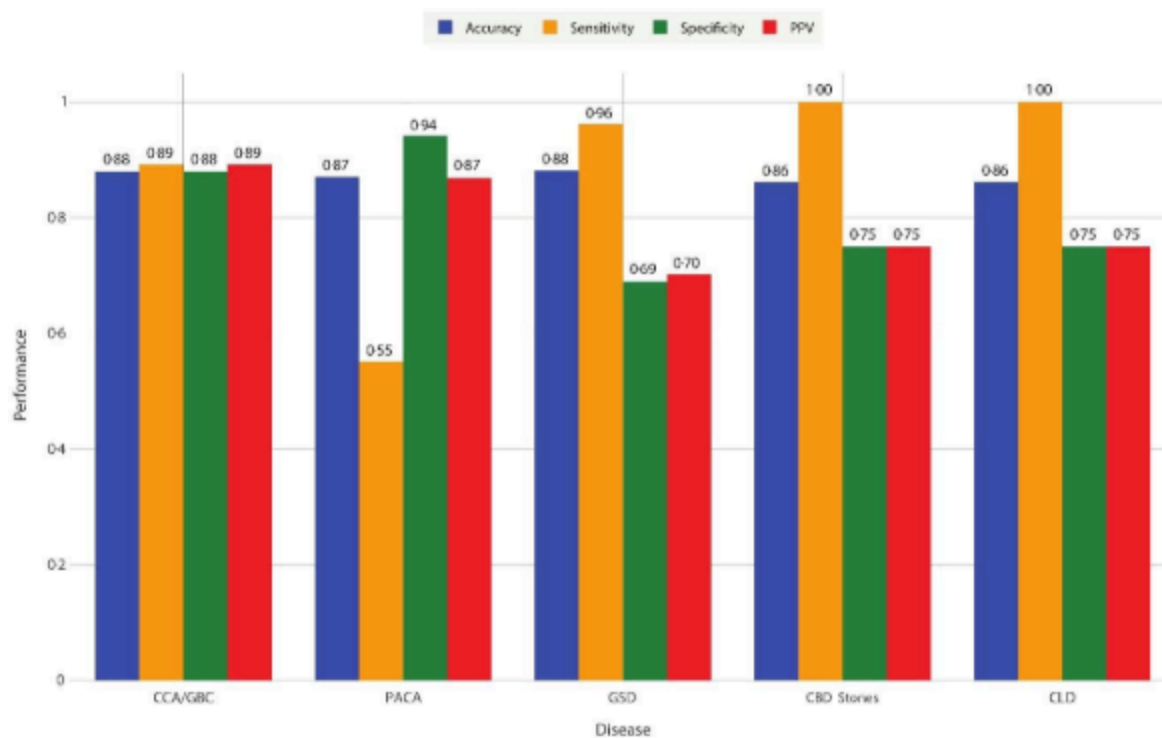


Figure 2

Disease-specific algorithm performance metrics. The bar chart compares accuracy, sensitivity, specificity, and PPV across five HPB disease categories. Gradient Boosting performed best for CCA/GBC, SVM for PACA and GSD, and AdaBoost for CBD stones and CLD in the reported disease-specific analyses.

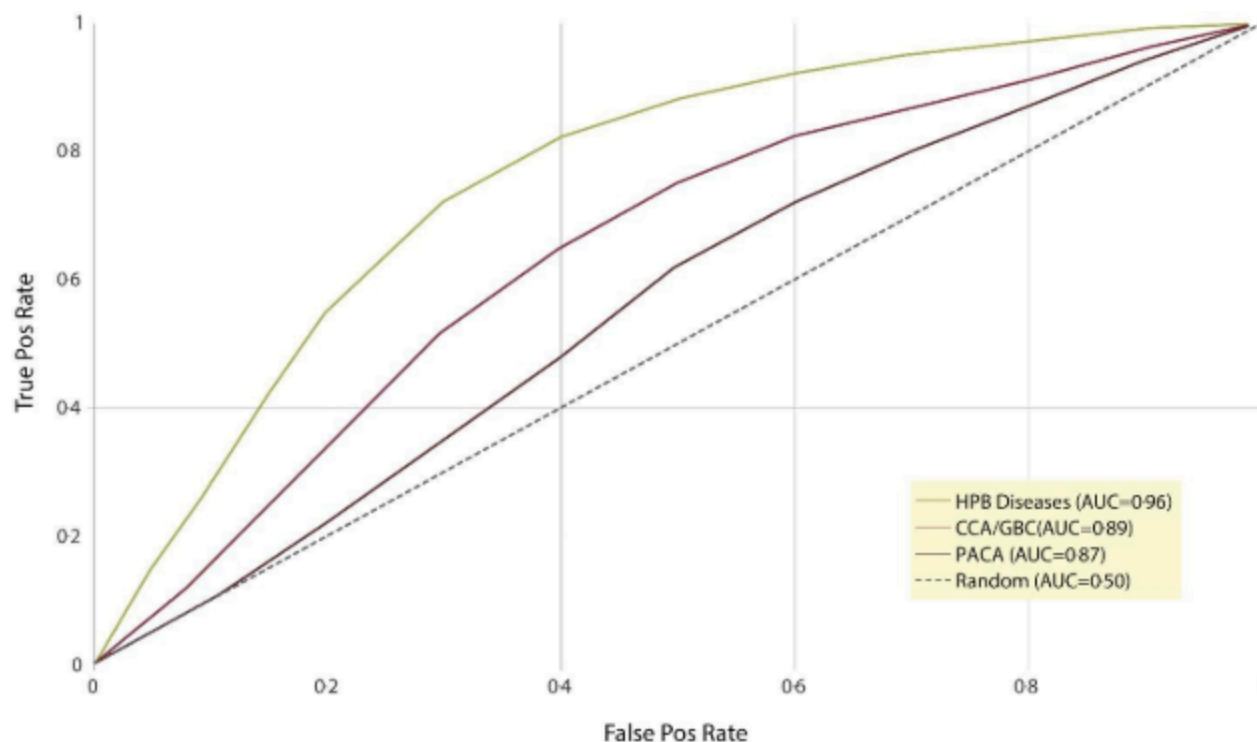


Figure 3

Receiver operating characteristic (ROC) curves for HPB disease classification models. The curves show reported AUC values for overall HPB diseases (0.96), CCA/GBC (0.89), and PACA (0.87).

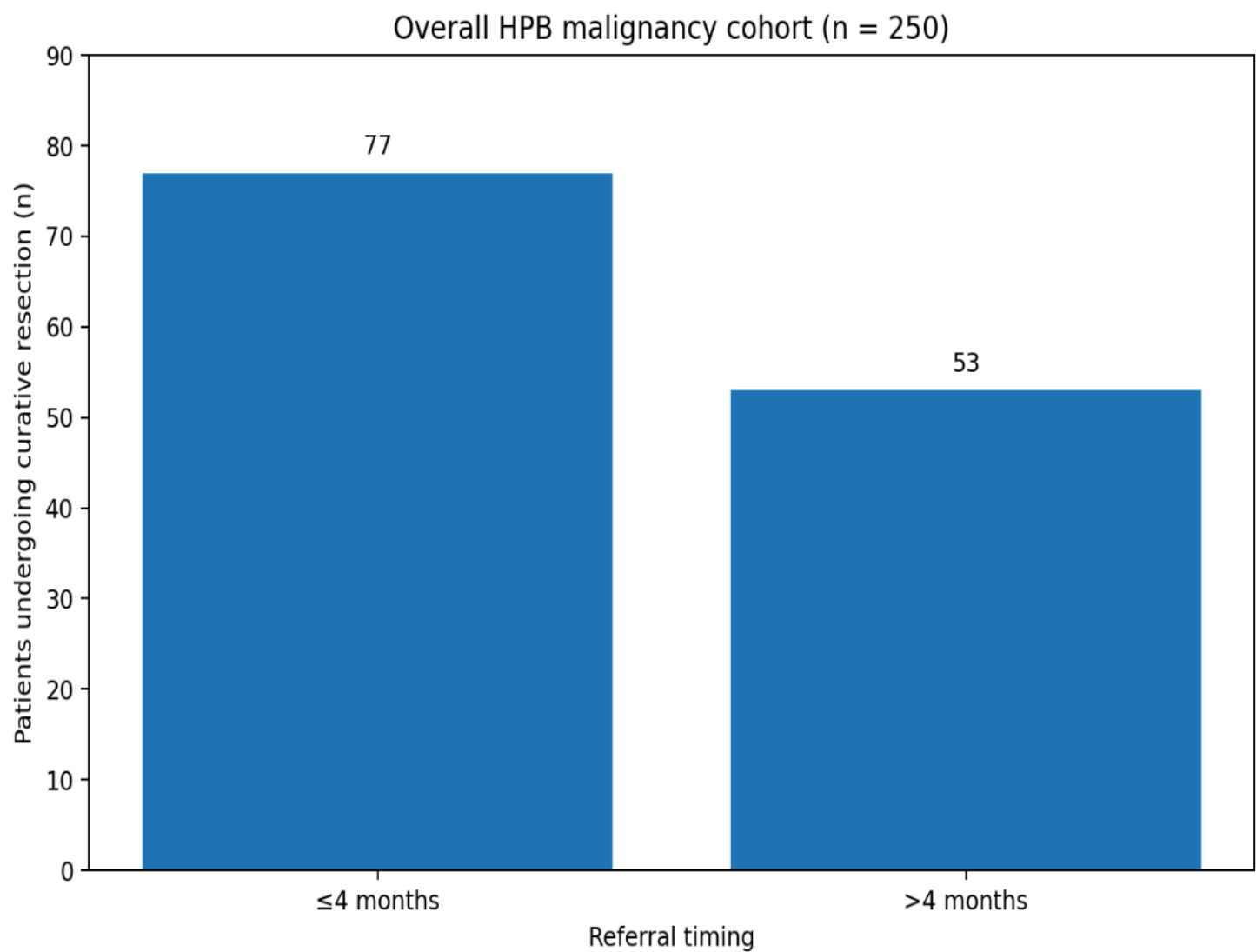


Figure 4

Reported curative resection counts by referral timing in the overall HPB malignancy cohort. Bars show the numbers of patients undergoing curative resection among those referred within four months (n = 77) and those referred later (n = 53), as reported for the overall malignancy cohort (n = 250).

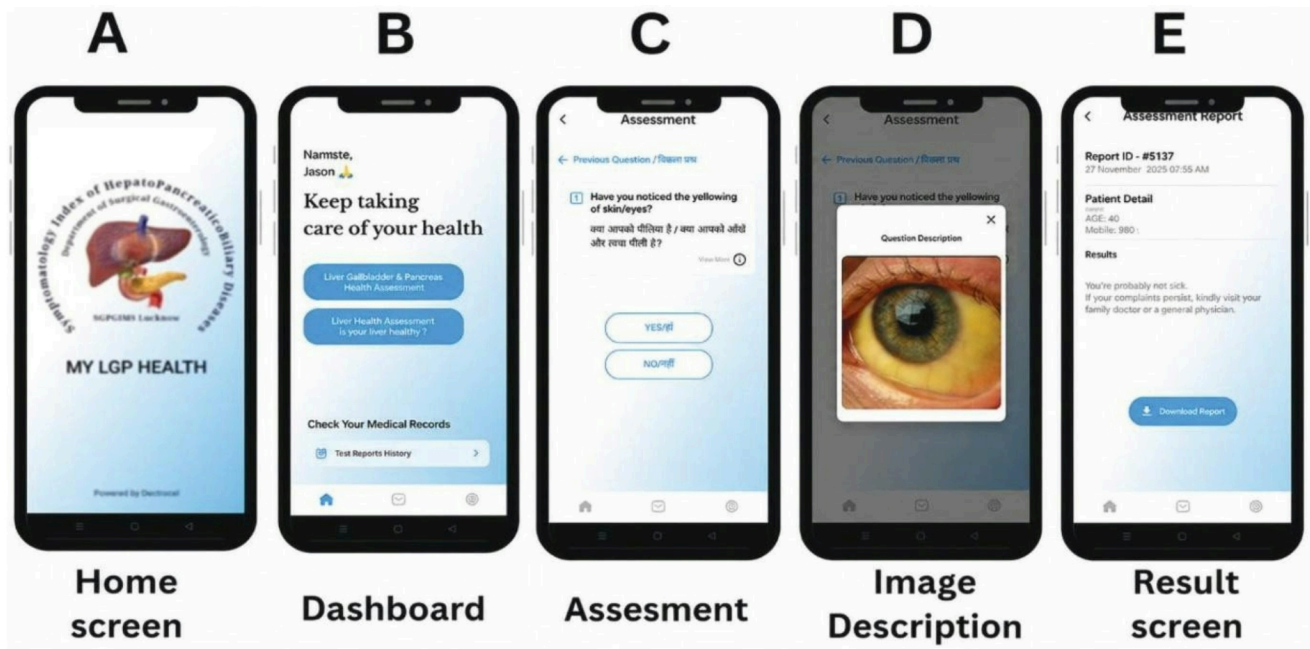


Figure 5

Clinical workflow of the hepatopancreatobiliary disease assessment application. Panels show the home screen (A), dashboard (B), symptom assessment interface (C–D), and final assessment report (E), illustrating structured symptom assessment and machine-learning-based risk stratification.

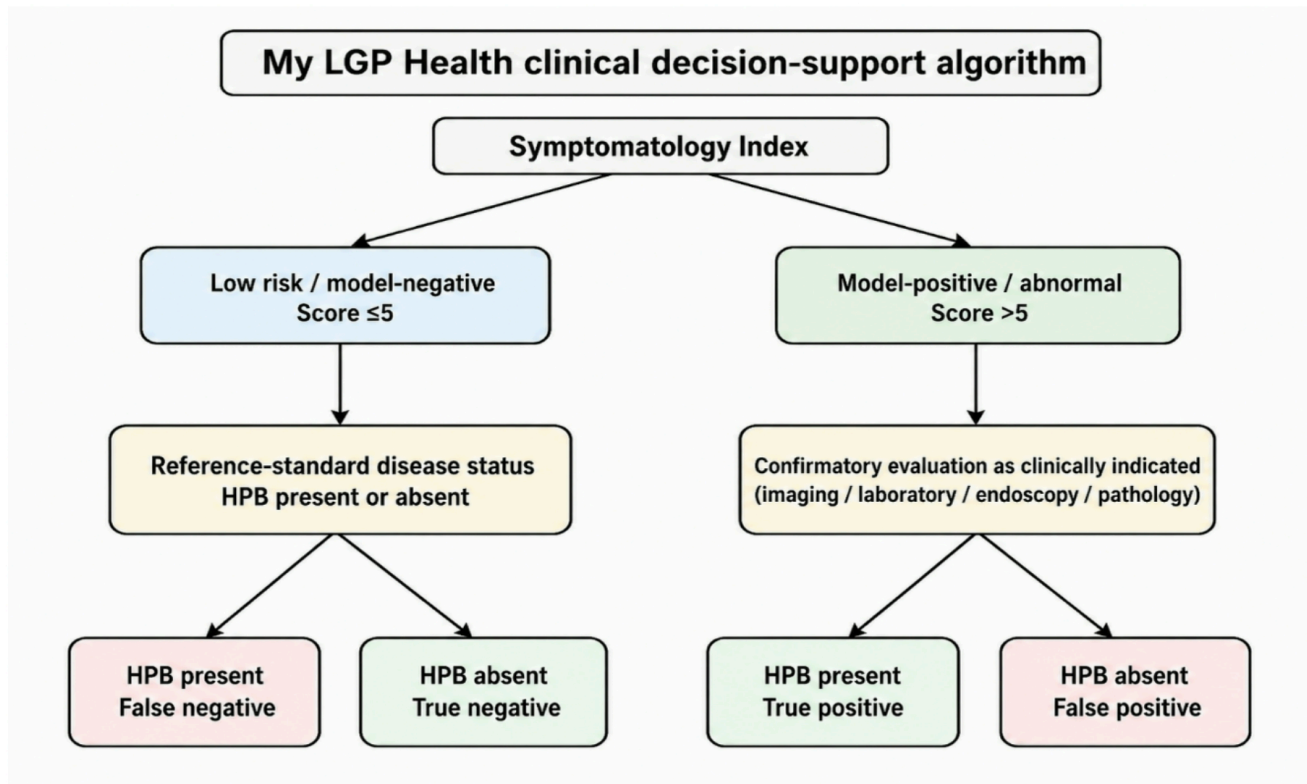


Figure 6

Diagnostic workflow of the My LGP Health clinical decision-support algorithm. For model evaluation, scores >5 were classified as model-positive and scores ≤5 as model-negative. Final reference-standard disease status was used to classify outcomes as true positive, false positive, false negative, or true negative.

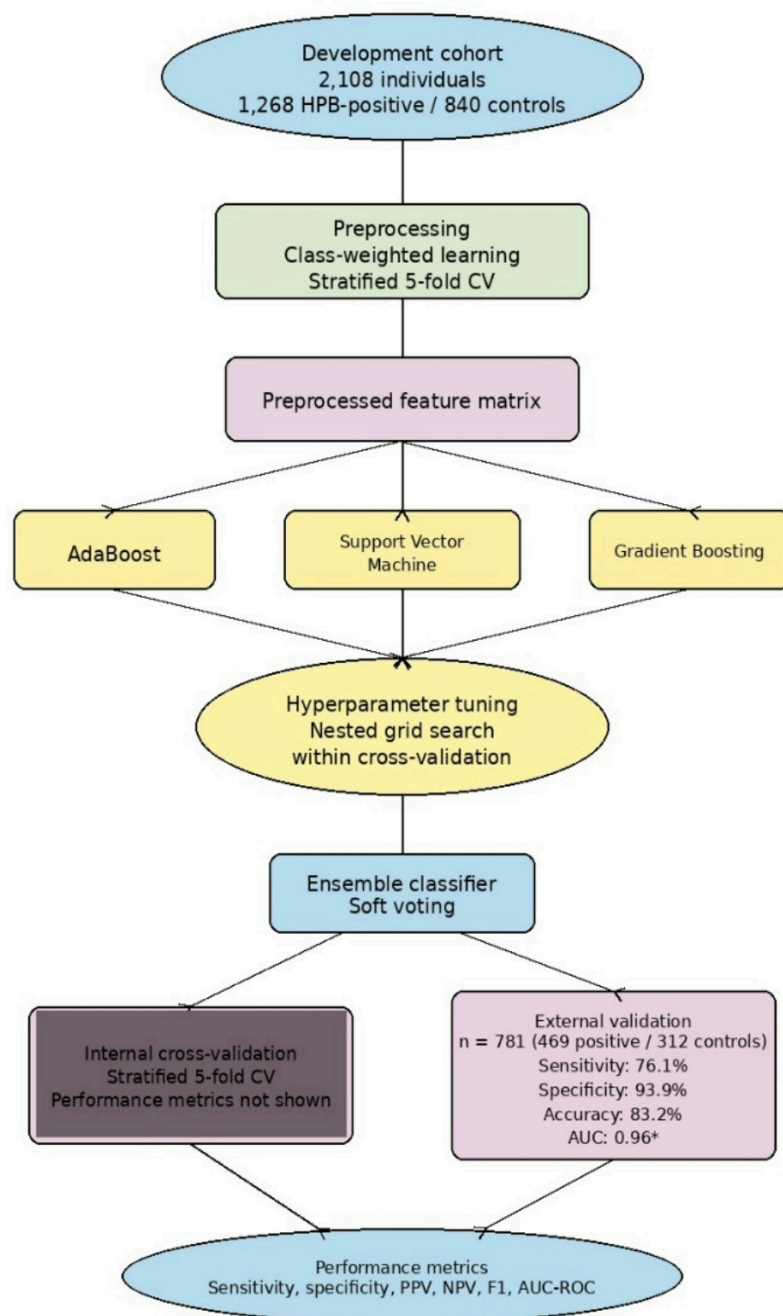


Figure 7

Development and validation workflow of the ensemble machine-learning model. The workflow comprised cohort selection, data preprocessing, feature engineering, model development using Gradient Boosting, AdaBoost, and Support Vector Machine classifiers, hyperparameter optimization, soft-voting ensemble construction, internal cross-validation, and independent validation.