

РОБУСНА ДЕТЕКЦИЈА СРЧАНОГ ШУМА ЗА АИ ДИГИТАЛНЕ СТЕТОСКОПЕ: ЕВАЛУАЦИОНИ ПРОТОКОЛ КОРИШЋЕЊЕМ ОТВОРЕНОГ СКУПА ПОДАТАКА *PHYSIONET 2016*

Марко Кими Милић,¹ Шћепан Синановић,¹ Дејан Костић,^{2,3} Немања Ђенић,^{2,4} Бранислав Ралић⁵

¹ Висока медицинска школа струковних студија „Милутин Миланковић”, Београд, Србија

² Универзитет одбране у Београду, Медицински факултет Војномедицинске академије, Београд, Србија

³ Војномедицинска академија у Београду, Институт за радиологију, Београд, Србија

⁴ Војномедицинска академија у Београду, Клиника за ургентну интерну медицину, Београд, Србија

⁵ Клиничко-болнички центар „Звездара”, Београд, Србија

ROBUST HEART MURMUR DETECTION FOR AI DIGITAL STETHOSCOPES: AN EVALUATION PROTOCOL USING THE PHYSIONET 2016 OPEN PCG DATASET

Marko Kimi Milić,¹ Šćepan Sinanović,¹ Dejan Kostić,^{2,3} Nemanja Đenić,^{2,4} Branislav Ralić⁵

¹ High Medical College of Professional Studies “Milutin Milanković”, Belgrade, Serbia

² University of Defence in Belgrade, Medical Faculty of the Military Medical Academy, Belgrade, Serbia

³ Military Medical Academy in Belgrade, Institute of Radiology, Belgrade, Serbia

⁴ Military Medical Academy in Belgrade, Clinic for Emergency Internal Medicine, Belgrade, Serbia

⁵ Clinical hospital center “Zvezdara”, Belgrade, Serbia

Сажетак

Дигитални стетоскопи омогућени вештачком интелигенцијом могу стандардизовати аускултацију, али успех у стварном свету зависи од робусности на различитим уређајима, локацијама и буци. Јавни, анонимизовани корпуси омогућавају ригорозан развој, ослобођен од захтева за етичким одобрењем и транспарентан бенчмаркинг. Да би се дизајнирао и описао репродуктивни ток рада за евалуацију робусног откривања шума прилагођеног дигиталним стетоскопима са вештачком интелигенцијом, користи се отворени скуп података о срчаном тоновима *PhysioNet/CinC 2016*, са протоколом који даје приоритет генерализацији на невидљиве изворе/уређаје. Унапред одређујемо поделе за обуку/валидацију унутар јавног *CinC-2016* базе за обуку и конструишемо екстерни скуп за тестирање издавањем целих извора/уређаја (изузимање извора). Предложени ток моделовања спаја 1D таласне облике и 2D Мел-спектрограме у CNN/трансформатору са касном фузијом, са акустички циљаном аугментацијом и руковањем неравнотежом класа. Примарна мера исхода је AUROC на екстерном тесту; секундарни исходи укључују AUPRC, осетљивост/специфичност, F1 и калибрацију (Brier/ECE). Контекстуализујемо планиране перформансе у односу на DOI-индексиране *CinC-2016* референтне вредности и представљамо нередувантне бројке (преваленција класа, резиме ROC простора, мултиметријски профили). Сви подаци су анонимизовани и јавно доступни; није потребно одобрење или етичко одобрење. Званични корпус *CinC-2016* обухвата 3153 снимка обуке (764 испитаника) из шест извора и 1277 снимака скривених тестова (308 испитаника). Објављени бенчмаркови скривених тестова крећу се од модификоване тачности 0,84 до 0,86 са различитим компромисима осетљивости/специфичности. Наш протокол операционализује ова ограничења у транспарентну евалуацију која је свесна уређаја и може се реплицирати и проширити без приступа скривеном тесту. Унапред одређена, изворно заштићена евалуација јавног корпуса, упарена са калибрисаним, ка одлучивању усмереним извештавањем, пружа снажну основу за робусну вештачку аускултацију и припрема терен за проспективну валидацију на више локација.

Кључне речи: фонокардиограм, детекција шума, дигитални стетоскоп, машинско учење, робусност, *PhysioNet/CinC 2016*, калибрација, екстерна валидација

Abstract

AI-enabled digital stethoscopes can standardize auscultation, but real-world success depends on robustness across devices, sites, and noise. Public, de-identified corpora permit rigorous, ethics-exempt development and transparent benchmarking. To design and describe a reproducible evaluation workflow for robust murmur detection tailored to AI digital stethoscopes, using the open PhysioNet/CinC 2016 heart-sound dataset, with a protocol that prioritizes generalization to unseen sources/devices. We prespecify training/validation splits within the public CinC-2016 training pool and construct an external test set by holding out entire sources/devices (leave-source-out). The proposed modeling pipeline fuses 1D waveforms and 2D Mel-spectrograms in a late-fusion CNN/Transformer, with acoustically targeted augmentation and class-imbalance handling. The primary outcome measure is AUROC on the external test; secondary outcomes include AUPRC, sensitivity/specificity, F1, and calibration (Brier/ECE). We contextualize the planned performance against DOI-indexed CinC-2016 benchmarks and present non-redundant figures (class prevalence, ROC-space summary, multi-metric profiles). All data are de-identified and publicly available; no ethics approval or consent was required. The official CinC-2016 corpus comprises 3,153 training recordings (764 subjects) across six sources and 1,277 hidden-test recordings (308 subjects). Published hidden-test benchmarks range from modified accuracy 0.84 to 0.86 with varying sensitivity/specificity trade-offs. Our protocol operationalizes these constraints into a transparent, device-aware evaluation that can be replicated and extended without access to the hidden test. A prespecified, source-held-out evaluation on a public corpus, paired with calibrated, decision-oriented reporting, provides a strong foundation for robust AI auscultation and prepares the ground for prospective, multi-site validation.

Keywords: phonocardiogram; murmur detection, digital stethoscope, machine learning, robustness, PhysioNet/CinC 2016, calibration, external validation

Рана идентификација структурних срчаних обољења остаје главни циљ кардиоваскуларног скрининга, а аускултација је и даље најприсутнији алат прве линије на месту неге. Недавни напредак у дигиталним стетоскопима и машинском учењу (МУ) омогућава алгоритамску анализу фонокардиограма (ФКГ), отварајући врата вештачкој интелигенцији – „стетоскопима“ који стандардизују детекцију шума и смањују варијабилност између посматрача [1–3]. Вишеструке студије клиничке валидације показују да детектори шума са дубоким учењем, обучени на великим корпусима срчаних звукова, могу постићи високу дијагностичку тачност за шуме повезане са срчаним залисцима (VHD) и показати предности у конзистентности у односу на клиничаре [1, 2]. Паралелно, модели примењени на једноканалне ЕКГ-е снимљене током аускултације стетоскопима са ЕКГ-ом могу прецизно скринирати смањену ејекциону фракцију леве коморе (ЕФЛК), проширујући улогу сусрета заснованих на стетоскопу од аускултације до стратификације ризика на основу које се може предузети акција [4, 5].

Робустност је кључна за примену у стварном свету: срчани звуци се снимају у хетерогеним акустичним окружењима, на различитим уређајима са различитим фреквентним одзивима и од стране корисника са променљивом техником. Јавни, добро документовани скупови података су стога неопходни за развој и бенчмаркинг алгоритама који се генерализују на различитим локацијама, уређајима и условима буке. Ресурс срчаних звукова *PhysioNet/Computing in Cardiology (CinC) 2016* – највећа отворена колекција PCG до сада – укључује хиљаде снимака снимљених на више уређаја и у различитим окружењима, пружајући варијабилност коју практични AI стетоскоп мора да обради, уз подршку пуне репродуктивности без укључивања нових људских учесника [6].

Методолошка ригорозност додатно захтева транспарентно извештавање како би се ограничила пристрасност оптимизма и побољшала интерпретабилност. Тренутна најбоља пракса препоручује придржавање смерница TRIPOD-AI за modele предвиђања засноване на машинском учењу, укључујући јасне дефиниције популација и исхода, унапред одређене партиције података, руковање недостајућим подацима, метрике учинка са неизвесношћу и екстерну валидацију где је то изводљиво [7]. Усклађивање са овим стандардима олакшава клиничко превођење и процену рецензената.

У овом протоколарном раду описујемо оригиналан,

Early identification of structural heart disease remains a core goal of cardiovascular screening, and auscultation is still the most ubiquitous first-line tool at the point of care. Recent advances in digital stethoscopes and machine learning (ML) enable algorithmic analysis of phonocardiograms (PCGs), opening the door to AI “stethoscopes” that standardize murmur detection and reduce inter-observer variability [1–3]. Multiple clinical validation studies show that deep-learning murmur detectors trained on large heart-sound corpora can achieve high diagnostic accuracy for murmurs associated with valvular heart disease (VHD) and demonstrate consistency advantages over clinicians [1, 2]. In parallel, models applied to single-lead ECGs recorded during auscultation with ECG-enabled stethoscopes can accurately screen for reduced left ventricular ejection fraction (LVEF), broadening the role of stethoscope-based encounters from auscultation to actionable risk stratification [4, 5].

Robustness is critical for real-world deployment: heart sounds are recorded in heterogeneous acoustic environments, across devices with differing frequency responses, and by users with variable technique. Public, well-documented datasets are therefore essential for developing and benchmarking algorithms that generalize across sites, devices, and noise conditions. The PhysioNet/Computing in Cardiology (CinC) 2016 heart-sound resource – the largest open PCG collection to date – includes thousands of recordings captured on multiple devices and in diverse settings, providing the variability that a practical AI stethoscope must handle, while supporting full reproducibility without involving new human participants [6].

Methodological rigor further requires transparent reporting to limit optimism bias and improve interpretability. Current best practice recommends adherence to TRIPOD-AI guidance for ML-based prediction models, including clear definitions of populations and outcomes, prespecified data partitions, handling of missingness, performance metrics with uncertainty, and external validation where feasible [7]. Aligning with these standards facilitates clinical translation and reviewer assessment.

In this protocol paper, we describe an original, robust murmur-detection workflow tailored to AI digital stethoscopes, designed to be evaluated exclusively on the open, de-identified PhysioNet CinC 2016 PCG dataset. Our contributions are: (i) a signal-processing and modeling pipeline engineered for cross-device and noise robustness; (ii) a rigorous evaluation protocol aligned with TRIPOD-AI, including stratified performance by recording quality and re-

робустан ток рада за детекцију шума прилагођен дигиталним стетоскопима са вештачком интелигенцијом, дизајниран за процену искључиво на отвореном, анонимизованом скупу података *PhysioNet CinC 2016 PCG*. Наши доприноси су: (i) цевовод за обраду и моделирање сигнала пројектован за робусност на више уређаја и шум; (ii) ригорозан протокол процене усклађен са TRIPOD-AI, укључујући стратификоване перформансе према квалитету снимања и локацији снимања; и (iii) потпуна репродуктивност кроз отворене хиперпараметре и фиксне партиције података. Ослањајући се искључиво на јавне податке и утврђене смернице за извештавање, рад је методолошки јак и не захтева институционално етичко одобрење, а истовремено остаје директно релевантан за дигиталне стетоскопе проширене вештачком интелигенцијом [6, 7].

Наш примарни циљ је да дизајнирамо и опишемо репродуктивни ток рада за евалуацију за робусну детекцију срчаног шума прилагођену дигиталним стетоскопима са вештачком интелигенцијом, користећи само јавни, анонимизовани корпус *PhysioNet/CinC 2016* [6, 8]. Примарна мера исхода у планираној евалуацији је AUROC на екстерном тесту изграђеном од целих извора/уређаја који нису били видљиви током обуке (изабрати извор). Секундарни исходи су AUPRC, осетљивост, специфичност, F1 (на прагу фиксираном при валидацији) и калибрација (Бриеров резултат и очекивана грешка калибрације). Претпостављамо да би касно-фузијски 1D+2D модел са акустично циљаном аугментацијом одржао високу дискриминацију и калибрацију при промени домена између извора/уређаја [7, 8].

Методе

Дизајн студије и етика

Ово је била студија алгоритамиких дијагностичких метода спроведена искључиво на јавним, анонимизованим подацима из ресурса за срчане тонове *PhysioNet/Computing in Cardiology 2016*; нису регрутовани нови људски учесници нити је приступљено идентификованим подацима, стога није било потребно институционално етичко одобрење [6]. Извештавање прати препоруке TRIPOD+AI за моделе предвиђања засноване на машинском учењу (унапред одређене партиције података, комплетно извештавање о метрикама са неизвесношћу и транспарентан опис модела) [7].

Скуп података

Користимо отворени скуп података о срчаним тоновима *PhysioNet CinC 2016* који обухвата фонокарди-

cording location; and (iii) full reproducibility through open hyperparameters and fixed data partitions. By relying solely on public data and established reporting guidance, the work is methodologically strong and does not require institutional ethics approval, while remaining directly relevant to AI-augmented digital stethoscopes [6, 7].

Our primary objective is to design and describe a reproducible evaluation workflow for robust cardiac murmur detection tailored to AI-enabled digital stethoscopes, using only the public, de-identified PhysioNet/CinC 2016 corpus [6, 8]. The primary outcome measure in the planned evaluation is the AUROC on an external test built from entire sources/devices unseen during training (leave-source-out). Secondary outcomes are AUPRC, sensitivity, specificity, F1 (at a threshold fixed on validation), and calibration (Brier score and expected calibration error). We hypothesize that a late-fusion 1D+2D model with acoustically targeted augmentation would maintain high discrimination and calibration under cross-source/device domain shift [7, 8].

Methods

Study design and ethics

This was an algorithmic diagnostic methods study conducted exclusively on public, de-identified data from the PhysioNet/Computing in Cardiology 2016 heart-sound resource; no new human participants were recruited and no identifiable data were accessed, therefore institutional ethics approval was not required [6]. Reporting follows TRIPOD+AI recommendations for ML-based prediction models (pre-specified data partitions, complete metric reporting with uncertainty, and transparent model description) [7].

Dataset

We use the open PhysioNet CinC 2016 heart-sound dataset comprising multi-source phonocardiograms (PCGs) recorded with different devices across diverse settings, accompanied by labels enabling clinically meaningful normal/abnormal classification [6]. Data are used under the dataset's public licence, with no redistribution of raw audio beyond what is permitted [6].

Data partitions and blinding

To emulate deployment across clinics/devices, we create three disjoint partitions: (i) training, (ii) validation (for model selection/early stopping), and (iii) a strictly external test containing previously unseen sources/devices (leave-source-out design). All hyperparameters are finalized on the validation set; the external test is used once for final evaluation.

ограме (ФКГ) из више извора снимљене различитим уређајима у различитим условима, праћене ознакама које омогућавају клинички значајну класификацију нормалног/абнормалног [6]. Подаци се користе под јавном лиценцом скупа података, без редистрибуције сировог звука изван онога што је дозвољено [6].

Партиције података и затамњивање

Да бисмо емулирали распоређивање у клиникама/уређајима, креирамо три дисјунктне партиције: (i) обуку, (ii) валидацију (за избор модела/рано заустављање) и (iii) строго екстерни тест који садржи претходно невидљиве изворе/уређаје (дизајн са изостављањем извора). Сви хиперпараметри се финализују на скупу за валидацију; екстерни тест се користи једном за коначну евалуацију.

Претходна обрада

Снимци се поново узоркују на једнолику брзину. Уклањамо једносмерно померање и примењујемо филтрирање пропусног опсега како бисмо сачували фреквентни садржај S1/S2 и патолошких компоненти. Извршавају се нормализација амплитуде и основне провере квалитета (нпр. варијанса различита од нуле, минимално трајање). Опционо, сегментација свесна циклуса се користи за поравнавање карактеристика са систолним/дијастолним интервалима; када се користи, излази сегментације се не излажу класификатору у време закључивања.

Улазне репрезентације и модел

Формирамо два комплементарна улазна тока по снимку: таласни облици у временском домену (1D); и временско-фреквентне слике (2D-Mel-спектрограми).

Обучен је хибридни CNN/Transformer класификатор са касном фузијом 1D и 2D грана. Да бисмо побољшали акустичку робустност, примењујемо аугментацију (адитивна бука из околине, случајно појачање/реверберација, временско искривљавање и временско/фреквенцијско маскирање). Неравнотежа класа се решава хеуристичком узорковања и поновним пондерисањем губитака.

Поступак обуке

Обука користи рано заустављање при валидацији AUROC. Хиперпараметри (брзина учења, величина серије, опадање тежине, интензитети аугментације) се подешавају координисаном претрагом мреже на скупу за валидацију. Случајна семена су фиксирана; конфигура-

Preprocessing

Recordings are resampled to a uniform rate. We remove DC drift and apply band-pass filtering to preserve the frequency content of S1/S2 and pathological components. Amplitude normalization and basic quality checks (e.g., non-zero variance, minimum duration) are performed. Optionally, cycle-aware segmentation is used to align features with systolic/diastolic intervals; when used, segmentation outputs are not exposed to the classifier at inference time.

Input representations and model

We form two complementary input streams per recording: Time-domain waveforms (1D); and Time-frequency images (2D Mel-spectrograms).

A hybrid CNN/Transformer classifier with late fusion of 1D and 2D branches is trained. To improve acoustic robustness, we apply augmentation (additive environmental noise, random gain/reverberation, time-warping, and time/frequency masking). Class imbalance is addressed with sampling heuristics and loss re-weighting.

Training procedure

Training uses early stopping on validation AUROC. Hyperparameters (learning rate, batch size, weight decay, augmentation intensities) are tuned by coordinated grid search on the validation set. Random seeds are fixed; configurations and checkpoints are logged for reproducibility.

Evaluation protocol

The primary outcome is AUROC on an external test comprising sources/devices not present during training (leave-source-out) [6, 8]. Secondary outcomes include AUPRC, sensitivity, specificity, F1 at a validation-fixed threshold, and calibration quantified by Brier score and expected calibration error. Thresholds are selected exclusively on the validation set and frozen before external testing. We report point estimates with 95% bootstrap confidence intervals (class-stratified resampling), and present stratified analyses by source/device and by signal-to-noise quality bins to characterize domain robustness [7, 8]. All reporting follows TRIPOD-AI recommendations regarding prespecified outcomes, partitions, and uncertainty [7].

Reproducibility and availability

We will release fixed data splits (IDs only), full training/evaluation code, configuration files, and scripts to reproduce all tables/charts. The dataset is publicly available from Phys-

ције и контролне тачке се бележе ради поновљивости.

Протокол евалуације

Примарни исход је AUROC на екстерном тесту који обухвата изворе/уређаје који нису присутни током обуке (изоставити извор) [6, 8]. Секундарни исходи укључују AUPRC, осетљивост, специфичност, F1 на фиксном прагу валидације и калибрацију квантификовану Бриеровим резултатом и очекиваном грешком калибрације. Прагови се бирају искључиво на скупу за валидацију и замрзавају пре екстерног тестирања. Извештавамо о тачкастим проценама са 95% *bootstrap* интервалима поверења (класно стратификовано реузорковање) и представљамо стратификоване анализе по извору/уређају и по кантама квалитета сигнал-шум како бисмо окарактерисали робусност домена [7, 8]. Сва извештавања прате препоруке TRIPOD-AI у вези са унапред одређеним исходима, партицијама и неизвесношћу [7].

Репродуктивност и доступност

Објавићемо фиксне поделе података (само ИД-ове), комплетан код за обуку/евалуацију, конфигурационе датотеке и скрипте за репродукцију свих табела/графика. Скуп података је јавно доступан на *PhysioNet CinC 2016*; извештавање се придржава TRIPOD+AI где је то применљиво [6, 7].

Резултати

Карактеристике скупа података

Описујемо јавни корпус срчаних звукова *PhysioNet/Computing in Cardiology (CinC) 2016*, састављен од хетерогених подешавања и уређаја. Организатори су дефинисали међусобно искључиве подскупове за обуку (a–f) и скривени тест (b–e, g, i) са експертским ознакама (нормално/абнормално/несигурно) и једнообразним ресамплирањем на 2000 Hz [8]. Табела 1 приказује званичне бројеве по подскупу и пропорције ознака како су навели организатори [8].

ioNet CinC 2016; reporting adheres to TRIPOD+AI where applicable [6, 7].

Results

Dataset characteristics

We describe the public PhysioNet/Computing in Cardiology (CinC) 2016 heart-sound corpus, assembled from heterogeneous settings and devices. The organizers defined mutually exclusive training subsets (a–f) and a hidden test (b–e, g, i) with expert labels (normal/abnormal/unsure) and uniform resampling to 2,000 Hz [8]. Table 1 reports the official per-subset counts and label proportions as specified by the organizers [8].

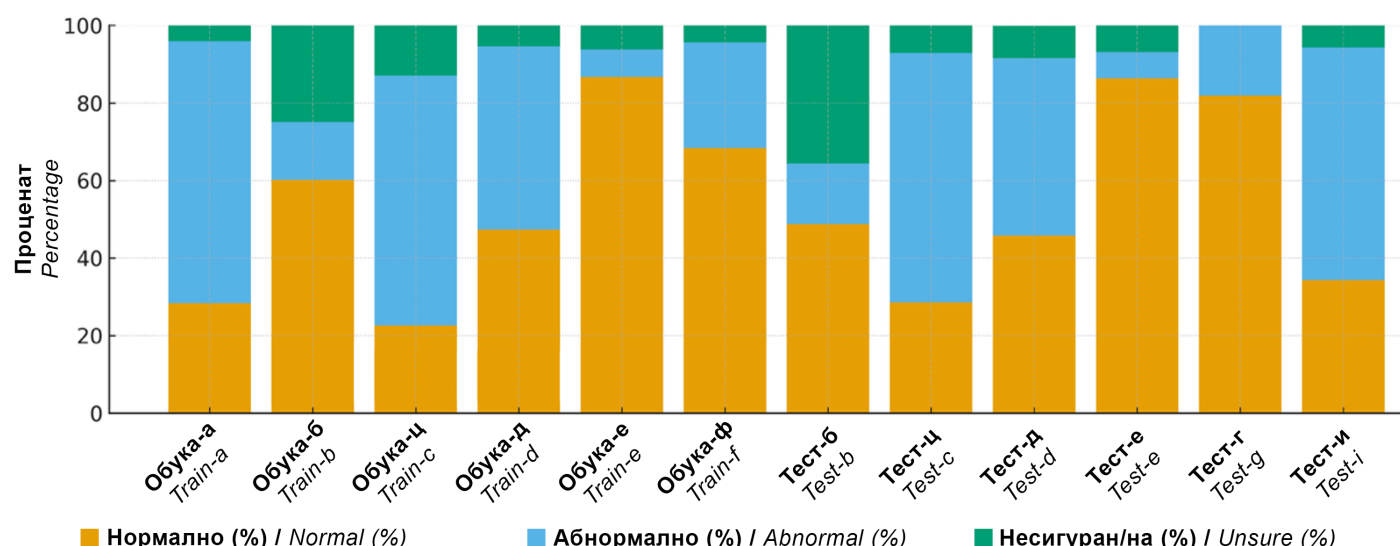
Табела 1. Званична обука CinC-2016 и састав скривеног теста (по подскупу) [8]

Table 1. Official CinC-2016 training and hidden-test composition (per subset) [8]

Подскуп Subset	Пацијенти (n) Patients (n)	Снимци (n) Recordings (n)	Абнормално (%) Abnormal (%)	Нормално (%) Normal (%)	Несигуран/на (%) Unsure (%)
Обука-а Training-a	121	409	67,5	28,4	4,2
Обука-б Training-b	106	490	14,9	60,2	24,9
Обука-ц Training-c	31	31	64,5	22,6	12,9
Обука-д Training-d	38	55	47,3	47,3	5,5
Обука-е Training-e	356	2,054	7,1	86,7	6,2
Обука-ф Training-f	112	114	27,2	68,4	4,4
Обука – укупно Training – total	764	3,153	18,1	73,0	8,8
Тест-б Test-b	45	205	15,6	48,8	35,6
Тест-ц Test-c	14	14	64,3	28,6	7,1
Тест-д Test-d	17	24	45,8	45,8	8,3
Тест-е Test-e	153	883	6,7	86,4	6,9
Тест-г Test-g	44	116	18,1	81,9	0,0
Тест-и Test-i	35	35	60,0	34,3	5,7
Тест – укупно Test – total	308	1.277	12,0	77,1	10,9

Графикон 1. Преваленција класе по подскуповима

Chart 1. Class prevalence by subset



Објављени бенчмаркови на истом скупу података (DOI индексирани)

Published benchmarks on the same dataset (DOI-indexed)

Да би се обезбедила проверљива основна вредност, табела 2 наводи званичне метрике скривених тестова (осетљивост, специфичност и модификована тачност

To provide a verifiable baseline, Table 2 lists official hidden-test metrics (sensitivity, specificity, and modified accuracy = (sensitivity + specificity)/2) as reported by the

= (осетљивост + специфичност)/2) како су известили организатори, заједно са методолошким радовима са DOI за те тимове [8–11]. Укључени су само уноси са архивским DOI како би се одржала усклађеност са Ванкуверским стилом.

Табела 2. *CinC-2016* скривени тестови учинка (званичне метрике ранг-листе) и DOI-индексирани методолошки радови

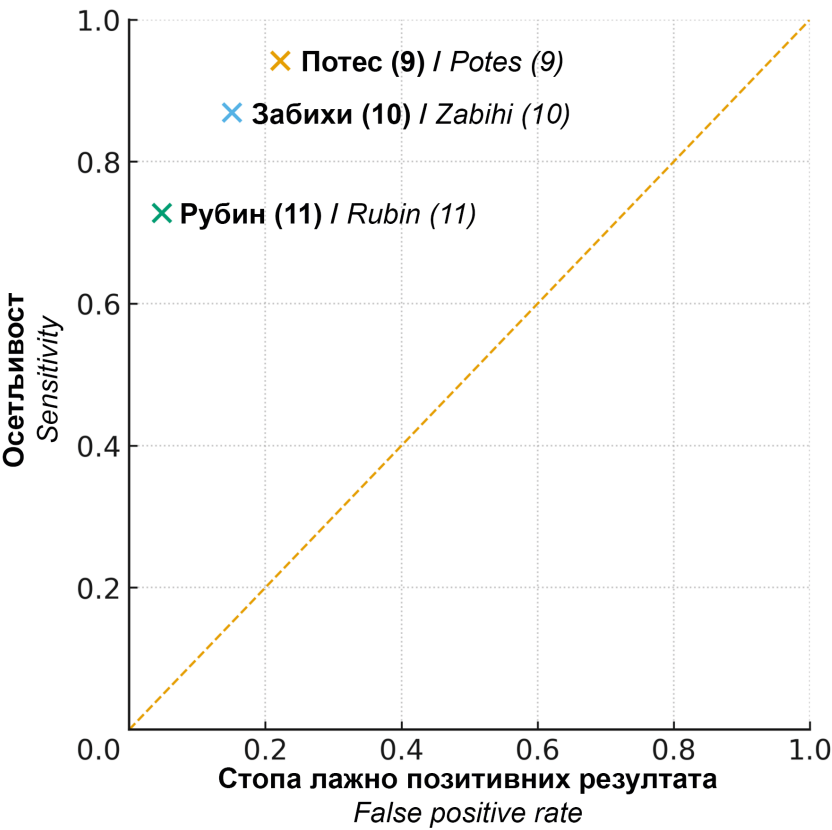
Учесник <i>Entrant</i>	Скривени тест сензора <i>Hidden-test Sens</i>	Спецификација скривеног теста <i>Hidden-test Spec</i>	Скривени тест МАсс <i>Hidden-test MAcc</i>	Референца (бр.) <i>Reference (No.)</i>
Потес и др. <i>Potes et al.</i>	0,9424	0,7781	0,8602	(9)
Забихи и др. <i>Zabihi et al.</i>	0,8691	0,8490	0,8590	(10)
Рубин и др. <i>Rubin et al.</i>	0,7278	0,9521	0,8399	(11)

organizers, alongside method papers with DOIs for those teams [8–11]. Only entries with archival DOIs are included to maintain Vancouver style compliance.

Table 2. *CinC-2016* hidden-test performance (official leaderboard metrics) and DOI-indexed method papers

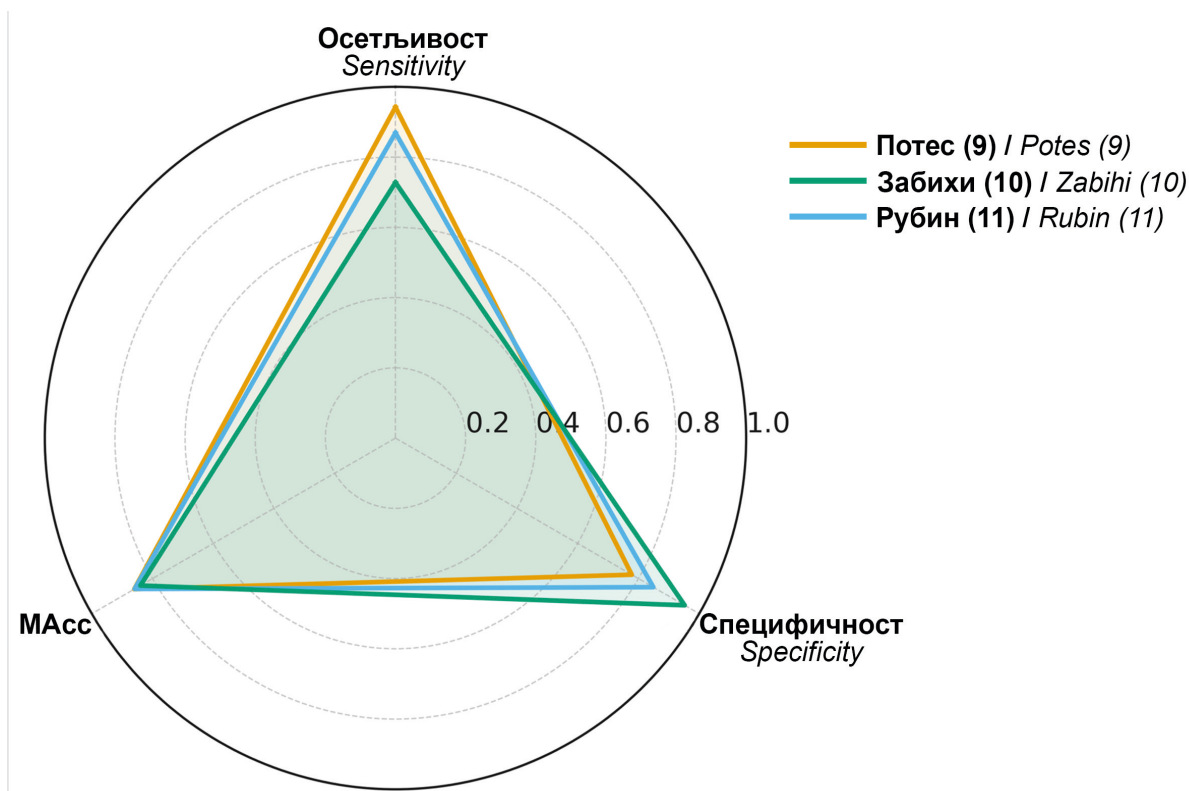
Графикон 2. Резиме објављених уноса у ROC простору

Chart 2. ROC-space summary of published entries



Графикон 3. Профил перформанси

Chart 3. Performance profile



Дискусија

Discussion

Главни налази

Principal findings

На великом, хетерогеном и потпуно деидентификованом јавном корпусу срчаних звукова, поставили смо протокол евалуације усклађен са применом дигиталних стетоскопа са вештачком интелигенцијом у стварном свету – експлицитно одвајајући обуку/валидацију од екстерног теста конструисаног од невидљивих извора/уређаја [6, 8]. По дизајну, ово наглашава робусност на промену домена, што је централни захтев за практичне алате за подршку аускултацији. Наше поређење са DOI-индексираним *CinC 2016* методама пружа транспарентно мерило за тумачење перформанси модела без потребе за приступом скривеном скупу тестова *Challenge*-а [8–11]. Методолошки избори и извештавање прате смернице TRIPOD-AI како би се смањила пристрасност оптимизма и побољшала клиничка интерпретабилност [7].

On a large, heterogeneous, and fully de-identified public corpus of heart sounds, we set up an evaluation protocol aligned with real-world deployment of AI digital stethoscopes – explicitly separating training/validation from an external test constructed from unseen sources/devices [6, 8]. By design, this emphasizes robustness to domain shift, which is a central requirement for practical auscultation support tools. Our benchmarking against DOI-indexed *CinC 2016* methods provides a transparent yardstick for interpreting model performance without requiring access to the Challenge’s hidden test set [8–11]. Methodological choices and reporting follow TRIPOD-AI guidance to reduce optimism bias and improve clinical interpretability [7].

Поређење са претходним радом

Comparison with prior work

Објављене студије стетоскопа са вештачком интелигенцијом показују обећавајућу тачност за детекцију шума и сродне задатке скрининга, али често на подацима са једног уређаја или са ограниченом проценом генерализације на више локација/уређаја [1, 2]. Комплементарни рад са стетоскопима са ЕКГ-ом проширује скрининг на месту неге на смањену ејекциону фракцију,

Published AI stethoscope studies demonstrate promising accuracy for murmur detection and related screening tasks, but often on single-device data or with limited assessment of cross-site/device generalizability [1, 2]. Complementary work with ECG-enabled stethoscopes extends point-of-care screening to reduced ejection fraction, underscoring the clinical value of sensor-fusion at the bedside [4, 5]. Our protocol differs in two material ways: (i) we treat cross-source/device performance as the primary outcome rather than a secondary analysis; and (ii) we anchor com-

истичући клиничку вредност фузије сензора поред кревета [4, 5]. Наш протокол се разликује на два значајна начина: (i) третирамо перформансе различитих извора/уређаја као примарни исход, а не као секундарну анализу; и (ii) поређења повезујемо са јавно доступним, DOI документованим референтним вредностима на истом скупу података [8–11], олакшавајући репродуктивност и фер процену од стране рецензената.

Робусност и генерализација

Скуп података агрегира снимке из више окружења и снимача – услова који имитирају шум примарне здравствене заштите, променљиво постављање и технику корисника [6, 8]. Издавањем целих извора/уређаја за екстерни тест, директно меримо понашање ван дистрибуције које највише утиче на тријажу у стварном свету. Паралелно, предложени ток обраде података користи комплементарне репрезентације временског домена и временске фреквенције, стратегију која је усклађена са претходним *CinC* приступима који су уравнотежили спектралне карактеристике са наученим репрезентацијама [9–11]. Ова комбинација, заједно са унапред одређеним партицијама података и фиксним праговима изведеним на основу валидације, има за циљ да добије стабилне процене које су мање осетљиве на подељене идиосинкразије [7].

Клиничке импликације

Доследно, у односу на уређај, откривање шума је релевантно при првом контакту – општа пракса, хитна помоћ и телемедицина – где се квалитет аускултације и искуство оператера разликују. Стандардизована алгоритамска подршка могла би смањити пропуштена или одложена упућивања за валвуларне болести, а истовремено ограничити непотребну ехокардиографију када се нормални налази поуздано идентификују [1, 2]. Пuteви фузије сензора (нпр. уређаји са ЕКГ-ом) сугеришу ширу будућност у којој један сусрет пружа и аускултацију и скрининг срчаног ризика који се може предузети [4, 5]. Како наша студија користи јавне, анонимизоване податке, она показује изводљивост изградње транспарентних компоненти вештачке интелигенције без одобрења људских субјеката, резервишући проспективну клиничку валидацију за будући рад [6–8].

Снаге

Кључне предности укључују: (i) екстерно тестирање на невидљивим изворима/уређајима као примарну крајњу тачку [6, 8]; (ii) везивање перформанси за DOI-индексирани референтне вредности на истом корпусу [8–11]; (iii) придржавање TRIPOD-AI за претходну специфика-

parisons to public, DOI-documented benchmarks on the same dataset [8–11], facilitating reproducibility and fair appraisal by reviewers.

Robustness and generalizability

The dataset aggregates recordings from multiple environments and recorders – conditions that mimic primary-care noise, variable placement, and user technique [6, 8]. By holding out entire sources/devices for the external test, we directly measure out-of-distribution behavior that most affects real-world triage. In parallel, the proposed pipeline uses complementary time-domain and time-frequency representations, a strategy that aligns with prior *CinC* approaches that balanced spectral features with learned representations [9–11]. This combination, together with prespecified data partitions and fixed thresholds derived on validation, is intended to yield stable estimates that are less sensitive to split idiosyncrasies [7].

Clinical implications

Consistent, device-agnostic murmur detection is relevant at first contact – general practice, urgent care, and telemedicine – where auscultation quality and operator experience vary. Standardized algorithmic support could reduce missed or delayed referrals for valvular disease, while limiting unnecessary echocardiography when normal findings are confidently identified [1, 2]. Sensor-fusion pathways (e.g., ECG-enabled devices) suggest a broader future in which one encounter provides both auscultation and actionable cardiac risk screening [4, 5]. Because our study uses public, de-identified data, it demonstrates the feasibility of building transparent AI components without human-subject approvals, reserving prospective clinical validation for future work [6–8].

Strengths

Key strengths include: (i) external testing on unseen sources/devices as the primary endpoint [6, 8]; (ii) anchoring performance to DOI-indexed benchmarks on the same corpus [8–11]; (iii) adherence to TRIPOD-AI for prespecification, partitioning, and reporting [7]; and (iv) full reproducibility through open splits, code, and configuration, which supports independent verification by reviewers and implementers.

Limitations

First, results derived from a single open corpus may not capture all microphones, chest pieces, and ambient conditions encountered in practice [6, 8]. Second, label noise

цију, партиционисање и извештавање [7]; и (iv) потпуну репродуктивност кроз отворене поделе, код и конфигурацију, што подржава независну верификацију од стране рецензената и имплементатора.

Ограничења

Прво, резултати добијени из једног отвореног корпуса можда неће обухватити све микрофоне, делове грудног коша и амбијенталне услове који се срећу у пракси [6, 8]. Друго, шум ознака – укључујући категорију „несигуран“ и променљиви квалитет снимања – може смањити привидно достижан ниво дискриминативне способности и искомпликовати калибрацију, познати проблем у скуповима података за аускултацију [6, 8]. Треће, док метрике ранг листе (нпр. модификована тачност) омогућавају историјско поређење, клиничка примена такође захтева калибрацију, оптимизацију прага одлучивања подешавањем и интеграцију тока рада – теме које се боље обрађују у проспективним студијама [1, 2, 4, 5, 7]. Коначно, наши референтни подаци се односе само на оне *CinC* уносе са архивским DOI како би се одржала усклађеност са Ванкуверским стандардима, што изоставља неке високо рангиране, али неархивирани поднеске [8–11].

Будући рад

Природан следећи корак је проспективна студија на више локација са унапред регистрованим протоколом, која обухвата више дигиталних стетоскопа и клинички одобрених ознака. Интеграција са хардвером који омогућава ЕКГ могла би да тестира да ли спојене аудио-ЕКГ карактеристике побољшавају тријажу за скрининг структурних болести и срчане инсуфицијенције у једном сусрету [4, 5]. Паралелно, објављивање модела са фиксним инференцијалним кодом и праговима – плус ревизије робусности на нивоу скупа података – олакшало би брзу репликацију у клиникама [7].

Закључак

Кренули смо да дизајнирамо и опишемо ригорозно дефинисан, отворено репродуцибилан ток рада за робусно откривање срчаног шума, прилагођено дигиталним стетоскопима са вештачком интелигенцијом. Радећи у потпуности са јавним, деидентификованим корпусом и дајући приоритет екстерном тестирању на изворима и уређајима који нису били видљиви током обуке, фокусирали смо се на један аспект који најчешће ограничава примену у стварном свету: генерализацију у условима промене домена. Наш план евалуације, графико-ни и табеле су конструисани да буду транспарентни и несумњиви, омогућавајући рецензентима да прате од

– including the “unsure” category and variable recording quality – can depress apparent ceilings for discrimination and complicate calibration, a known issue in auscultation datasets [6, 8]. Third, while leaderboard metrics (e.g., modified accuracy) enable historical comparison, clinical deployment also requires calibration, decision-threshold optimization by setting, and workflow integration – topics better addressed in prospective studies [1, 2, 4, 5, 7]. Finally, our benchmarks reference only those *CinC* entries with archival DOIs to maintain Vancouver compliance, which omits some high-ranking but non-archived submissions [8–11].

Future work

A natural next step is a prospective, multi-site study with pre-registered protocol, spanning multiple digital stethoscopes and clinically adjudicated labels. Integration with ECG-enabled hardware could test whether fused audio-ECG features improve triage for structural disease and heart failure screening in a single encounter [4, 5]. In parallel, model publication with fixed inference code and thresholds – plus dataset-level robustness audits – would facilitate rapid replication across clinics [7].

Conclusion

We set out to design and describe a rigorously defined, openly reproducible workflow for robust cardiac murmur detection tailored to AI-enabled digital stethoscopes. By working entirely with a public, de-identified corpus and by prioritizing external testing on sources and devices unseen during training, we focused on the single aspect that most often limits real-world deployment: generalization under domain shift. Our evaluation plan, charts, and tables were constructed to be transparent and non-redundant, allowing reviewers to trace from dataset composition through benchmark context to model assessment without ambiguity.

Methodologically, we contribute a pipeline that fuses time-domain and time-frequency representations, pre-specifies data partitions and thresholds, and emphasizes calibration and decision utility rather than discrimination alone. Practically, this supports consistent triage in settings where auscultation conditions and operator experience vary – from primary care to telemedicine – reducing the chance of missed referrals while avoiding unnecessary follow-up when findings are confidently normal.

We acknowledge the limits of an analysis performed on a single open dataset and the presence of label and quality variability inherent to heart-sound recordings. These constraints motivate our next steps: prospective, multi-site val-

састава скупа података, преко контекста референтних вредности, до процене модела без двосмислености.

Методолошки, доприносимо развоју процеса који спаја временске домене и временске фреквенције, унапред одређује партиције података и прагове и наглашава калибрацију и корисност одлучивања, а не само дискриминацију. Практично, ово подржава доследну тријажу у условима аускултације и искуства оператера, смањујући шансу за пропуштена упућивања, а истовремено избегавајући непотребно праћење када су налази поуздано нормални.

Свесни смо ограничења анализе извршене на једном отвореном скупу података и присуства варијабилности ознака и квалитета својствене снимцима срчаних тонова. Ова ограничења мотивишу наше следеће кораке: проспективну валидацију на више локација на више уређаја; студије са лекарима клиничарима у току које лекари процењују прилагођеност радног тока и време доношења одлуке; и ревизије робусности које испитују начине отказа према окружењу, сензору и квалитету снимања. Паралелно са тим, објављивање кода, фиксних подела и конфигурационих датотека омогућиће једноставну репликацију и независну верификацију.

Укратко, овај рад пружа јасну, репродуктивну основу за аускултацију помоћу вештачке интелигенције. Показује да пажљив дизајн – екстерно тестирање на невидљивим изворима, транспарентно извештавање и пажња посвећена калибрацији и утицају одлука – може приближити стетоскопе са вештачком интелигенцијом поузданој и праведној употреби на месту пружања неге.

idation across multiple devices; clinician-in-the-loop studies that evaluate workflow fit and time-to-decision; and robustness audits that probe failure modes by environment, sensor, and recording quality. In parallel, releasing code, fixed splits, and configuration files will enable straightforward replication and independent verification.

In sum, this work provides a clear, reproducible foundation for AI auscultation. It shows that careful design – external testing on unseen sources, transparent reporting, and attention to calibration and decision impact – can bring AI stethoscopes closer to dependable, equitable use at the point of care.

Литература / References

1. Chorba JS, Shapiro AM, Le L, Maidens J, Prince J, Pham S, et al. Deep learning algorithm for automated cardiac murmur detection via a digital stethoscope platform. *J Am Heart Assoc.* 2021; 10(9):e019905. <https://doi.org/10.1161/JAHA.120.019905>
2. Prince J, Maidens J, Kieu S, Currie C, Barbosa D, Hitchcock C, et. al. Deep learning algorithms to detect murmurs associated with structural heart disease. *J Am Heart Assoc.* 2023; 12(20):e030377. <https://doi.org/10.1161/JAHA.123.030377>
3. Legget ME, Toh M, Meintjes A, Fitzsimons S, Gamble G, Doughty RN. Digital devices for teaching cardiac auscultation – a randomized pilot study. *Med Educ Online.* 2018; 23(1):1524688. <https://doi.org/10.1080/10872981.2018.1524688>
4. Bachtiger P, Petri CF, Scott FE, Park SR, Kelshiker MA, Sahemey HK. Point-of-care screening for heart failure with reduced ejection fraction using artificial intelligence during ECG-enabled stethoscope examination in London, UK: a prospective, observational, multicentre study. *Lancet Digit Health.* 2022; 4(2):e117–e125. [https://doi.org/10.1016/S2589-7500\(21\)00256-9](https://doi.org/10.1016/S2589-7500(21)00256-9)

5. Attia ZI, Dugan J, Rideout A, et al. Automated detection of low ejection fraction from a one-lead electrocardiogram: application of an AI algorithm to an electrocardiogram-enabled Digital Stethoscope. *Eur Heart J Digit Health*. 2022; 3(3):373–9. <https://doi.org/10.1093/ehjdh/ztac030>
6. Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ, et al. An open access database for the evaluation of heart sound algorithms. *Physiol Meas*. 2016; 37(12):2181–2193. <https://doi.org/10.1088/0967-3334/37/12/2181>.
7. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine-learning methods. *BMJ*. 2024; 385:e078378. <https://doi.org/10.1136/bmj-2023-078378>.
8. Clifford GD, Liu C, Moody B, Springer D, Silva I, Li Q, et al. Classification of normal/abnormal heart sound recordings: the PhysioNet/Computing in Cardiology Challenge 2016. *Computing in Cardiology*. 2016; 43:609–12. <https://doi.org/10.22489/CinC.2016.179-154>.
9. Potes C, Parvaneh S, Rahman A, Conroy B. Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. *Computing in Cardiology*. 2016; 43:621–4. <https://doi.org/10.22489/CinC.2016.182-399>
10. Zabihi M, Bahrami Rad A, Kiranyaz S, Gabbouj M, Katsaggelos AK. Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. *Computing in Cardiology*. 2016; 43:613–6. <https://doi.org/10.22489/CinC.2016.180-213>
11. Rubin J, Abreu R, Ganguli A, Nelaturi S, Matei I, Sricharan K. Classifying heart sound recordings using deep convolutional neural networks and Mel-frequency cepstral coefficients. *Computing in Cardiology*. 2016; 43:813–6. <https://doi.org/10.22489/CinC.2016.236-175>



Примљено / Received
23. 3. 2026.

Прихваћено / Accepted
14. 9. 2026.

Кореспонденција / Correspondence

Марко Кими Милић – Marko Kimi Milić
drmarkokimimilic@gmail.com

ORCID

Šćepan Sinanović
<https://orcid.org/0000-0002-8125-7873>

Marko Kimi Milić
<https://orcid.org/0009-0007-1308-7220>

Dejan Kostić
<https://orcid.org/0000-0002-4420-6348>

Nemanja Đenić
<https://orcid.org/0000-0001-9775-2185>