electronics

MDPI

Systematic Review

Evaluation Protocols and On-Chip Feasibility in Embedded Electrocardiogram Classification: A Systematic Review

Moisés Arredondo-Velázquez

Center for Research and Advanced Studies of the National Polytechnic Institute (Cinvestav), Campus Tamaulipas, Carretera Victoria–Soto la Marina Km 5.5, Ciudad Victoria 87130, Tamaulipas, Mexico; juan.arredondo@cinvestav.mx

Abstract

Embedded electrocardiogram (ECG) classifiers are compared through classification quality and hardware cost, while the evaluation protocol that produced the quality value is treated as a matter of statistical fairness without hardware consequence. The opposite position is developed here. The data partition (intra-subject, patient-specific, inter-patient, inter-session) fixes the variability a classifier must absorb, and therefore the capacity it requires, the memory it occupies and the class of device able to run it. Web of Science, Scopus, IEEE Xplore, PubMed and ScienceDirect were searched (last on 21 September 2026) and 54 studies were included: 30 embedded implementations on microcontroller, FPGA and ASIC targets and 24 algorithmic studies spanning five model families. Risk of bias was assessed with a five-domain instrument and the results were synthesized narratively without pooling. The primary evidence is paired. Five studies evaluated one model under more than one partition with database and preprocessing held constant, and the loss on unseen subjects ranged from 6.2 percentage points of accuracy to a macro-averaged F1 score falling from 0.81 to 0.35 in a model whose accuracy remained near 0.98. Across studies, group mean accuracy was 97.9% for intra-subject ($n = 12$) and 92.5% for inter-patient ($n = 2$) evaluation, a descriptive difference of 5.4 points between groups that differ in database, class set and preprocessing. Deployed footprints spanned 1.3 to 186.2 kB, and five of the twelve designs with a determinable footprint exceeded the 16 kB on-chip budget of an entry-class microcontroller, beyond which every inference requires external accesses one to two orders of magnitude more costly. The capacity cost of subject-separated evaluation is formulated as a mechanism-based hypothesis (the generalization tax), shown to be consistent with every paired observation, confronted with the on-chip memory of named devices, and tied to the measurement that would quantify it, a capacity sweep under a fixed inter-patient partition. The conclusions apply to the beat-level classification of individuals not seen in training.

Keywords: convolutional neural networks; edge computing; electrocardiogram; embedded systems; evaluation protocol; inter-patient evaluation; model capacity; PRISMA; spiking neural networks; systematic review; transformers

Check for updates

Academic Editors: Yu Feng, Ning Li and Han Huang

Received: 15 September 2026
Revised: 17 October 2026
Accepted: 22 October 2026
Published: 30 October 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Cardiovascular disease continues to be the leading cause of mortality worldwide, with annual deaths expected to increase from 20.5 million in 2025 to 35.6 million by 2050 [1]. Three quarters of these deaths take place in low- and middle-income countries, where a shortage of eleven million health workers is projected by 2030 [1,2]. In rural areas, a single cardiologist may be responsible for several hundred thousand individuals, and referrals frequently rely

on unstable connections. As a result, a screening device must acquire, classify, and report findings locally, operate on battery power, and function without onsite interpretation.

This context fixes the evidence a classifier must provide. A device operating in such districts encounters individuals absent from any training set, so the meaningful performance value is the one obtained on subjects excluded from training, a requirement stated in general terms by the clinical machine learning literature as the need to approximate the use case in the cross-validation design [3], and documented for deep ECG classifiers as a drop from 98.4% to 90.2% mean accuracy in the eleven studies that report both an intra-patient and an inter-patient result [4]. The closest approximation to this requirement is found in two groups of studies. The first adopts the inter-patient paradigm of De Chazal et al. [5]. The second, a small set of embedded implementations, combines that paradigm with the AAMI standard [6] and with evaluation on physical hardware [7]. Both groups report accuracies below those obtained under partitions that share subjects between training and test, and this difference has direct consequences for hardware design.

A substantial body of the literature presents ECG classifiers implemented on embedded platforms [8,9], typically evaluated with respect to classification quality, often measured by accuracy on the MIT-BIH Arrhythmia Database [10], and execution cost, which can include latency, resource utilization, power, or energy [11]. Both aspects are frequently regarded as inherent characteristics of the system, leading to judgments that a model achieving 99 percent accuracy in 15 kilobytes is superior to one achieving 92 percent accuracy in 40 kilobytes. This method of comparison fails to recognize that accuracy depends on the evaluation partition, and that partitions can vary significantly in difficulty, making a direct comparison of values derived from different partitions invalid. De Chazal et al. demonstrated two decades ago that mixing beats from a single patient between training and testing phases artificially increases performance, as classifiers tend to memorize patient-specific morphology. They introduced the record-separated DS1/DS2 split, which has since become the benchmark for rigorous evaluation [5]. Luz et al. reinforced this point [12], and Silva et al. observed that only a small proportion of embedded studies comply with both the inter-patient paradigm and the AAMI standard [7].

The hardware implication remains insufficiently addressed. When evaluation on previously unseen subjects increases the complexity of the task, the classifier must possess greater capacity, and this requirement incurs a concrete cost in devices with memory fixed during fabrication. A model with weights and activations that reside within on-chip static memory incurs only arithmetic energy consumption, while one that exceeds this limit incurs the additional expense of external memory accesses, which are one to two orders of magnitude more costly [13–15]. As a result, the selection of evaluation protocol directly influences the capacity needed and the category of device capable of supporting the classifier.

This phenomenon is herein referred to as the generalization tax, denoting the additional model capacity (parameters, memory footprint, peak activation memory and latency) required for a classifier to maintain diagnostic utility when evaluated on previously unseen subjects. This term is used throughout as the name of a hypothesis about a mechanism. The evidence available in the literature measures the loss of quality when the partition changes and the capacity is held fixed, whereas the tax is the capacity that would recover that loss, and the two are related through a capacity-versus-quality curve, available in the literature under intra-subject partitions and identified here as the open measurement under the inter-patient one. The hypothesis is stated precisely, every observation that bears on it is assembled, and the result is organized into a decision procedure that links the deployment scenario, the evaluation protocol and the platform able to host the classifier.

Three contributions are made in this work, mapped in Figure 1 against the existing literature. First, the evaluation protocol is treated as an independent variable whose con-

sequence for the capacity required in hardware is argued from the cost structure of each classifier family and bounded by the paired observations available, whereas prior reviews used the protocol as a criterion of study fairness [7,12]. Second, five classifier families are placed in one cost model expressed in parameters, multiply–accumulate operations and peak activation memory, which allows transformers [16], recurrent networks, spiking networks and the convolutional architectures of the embedded corpus to be compared directly. Third, the protocol-driven capacity requirement is confronted with the on-chip memory of representative hardware targets, and research directions for reducing complexity under inter-patient constraints are outlined.

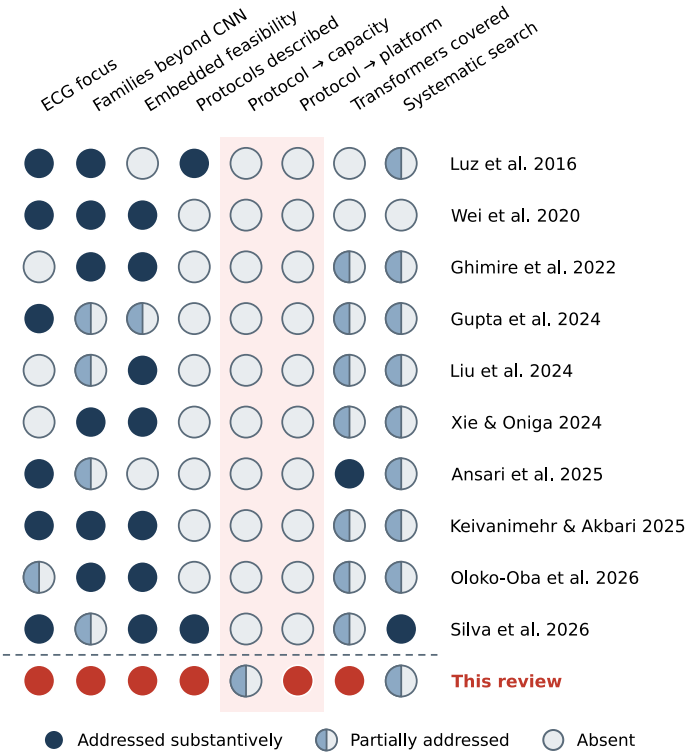


Figure 1. Coverage of the available reviews against the dimensions addressed in the present work. Rows correspond to Luz et al. 2016 [12], Wei et al. 2020 [11], Ghimire et al. 2022 [17], Gupta et al. 2024 [9], Liu et al. 2024 [18], Xie and Oniga 2024 [19], Ansari et al. 2025 [16], Keivanimehr and Akbari 2025 [20], Oloko-Oba et al. 2026 [21] and Silva et al. 2026 [7]. A filled marker indicates a dedicated section with systematic extraction, a half marker a discussion without extraction, and an open marker no coverage. The present work receives a half marker in the column on required capacity because the effect is argued mechanistically, and in the column on systematic search because screening was performed by a single reviewer. The shaded columns are the two least covered by the available reviews.

The work most closely related is that by Silva et al., who screened 1427 records to 122 studies published between 2017 and 2024 and found that 68 (55.7%) adhered to AAMI guidelines, 37 (30.3%) employed an inter-patient split, 38 (31.1%) discussed embedded feasibility and 5 (4.1%) fulfilled all three [7]. That study established the prevalence of subject-separated evaluation. Complementary to that work, the memory a classifier occupies under each protocol and the devices able to host it are examined here. The deployed footprints of the corpus are confronted directly with the on-chip memory of named devices, and the inter-patient regime is reached through the mechanism and the paired evidence assembled in this systematic review.

Figure 1 shows a clear division. Reviews on algorithm and hardware co-design focus on accelerator technologies [11,17,18], with comparable analyses for other biopotentials [19]. Assessments of ECG classification quality concentrate on deep learning [8,9,22] and on

transformer and large language model methods [16]. Surveys of TinyML for cardiovascular disease [20] and of edge machine learning for wearables [21] describe the deployment pipeline. Adjacent topics such as data leakage in machine learning [23], dataset shift in clinical artificial intelligence [24] and patient-level validation of ECG classifiers [25] are cited where they support a specific argument, and general-purpose accelerator design is outside the scope of the corpus, which is restricted to embedded ECG classifiers.

The remainder of the paper is organized as follows. Section 2 describes the search, selection and extraction procedures under the PRISMA 2020 statement. Section 3 presents the four protocols, the leakage modes that blur them, and the metrics under which an imbalanced beat-level result is read at the clinical level. Section 4 reviews the five classifier families. Section 5 relates model complexity to the constraints of an embedded target, with a common vocabulary for the resources of microcontroller, reconfigurable and custom targets, the accounting of the deployment footprint and the real-time budget. Section 6 compiles the evidence on the generalization-tax hypothesis, with the paired within-study comparisons first. Section 7 introduces the decision framework, Section 8 the reporting conventions and Section 9 the research directions, and Section 10 concludes.

2. Review Methodology

This review was prepared and reported following the PRISMA 2020 statement [26]. The completed checklists are provided as Supplementary Materials. The strings used for each interface are presented in Table 1, the flow of records appears in Figure 2, and the extraction sheet containing all recorded fields for each included study is available in comma-separated format.

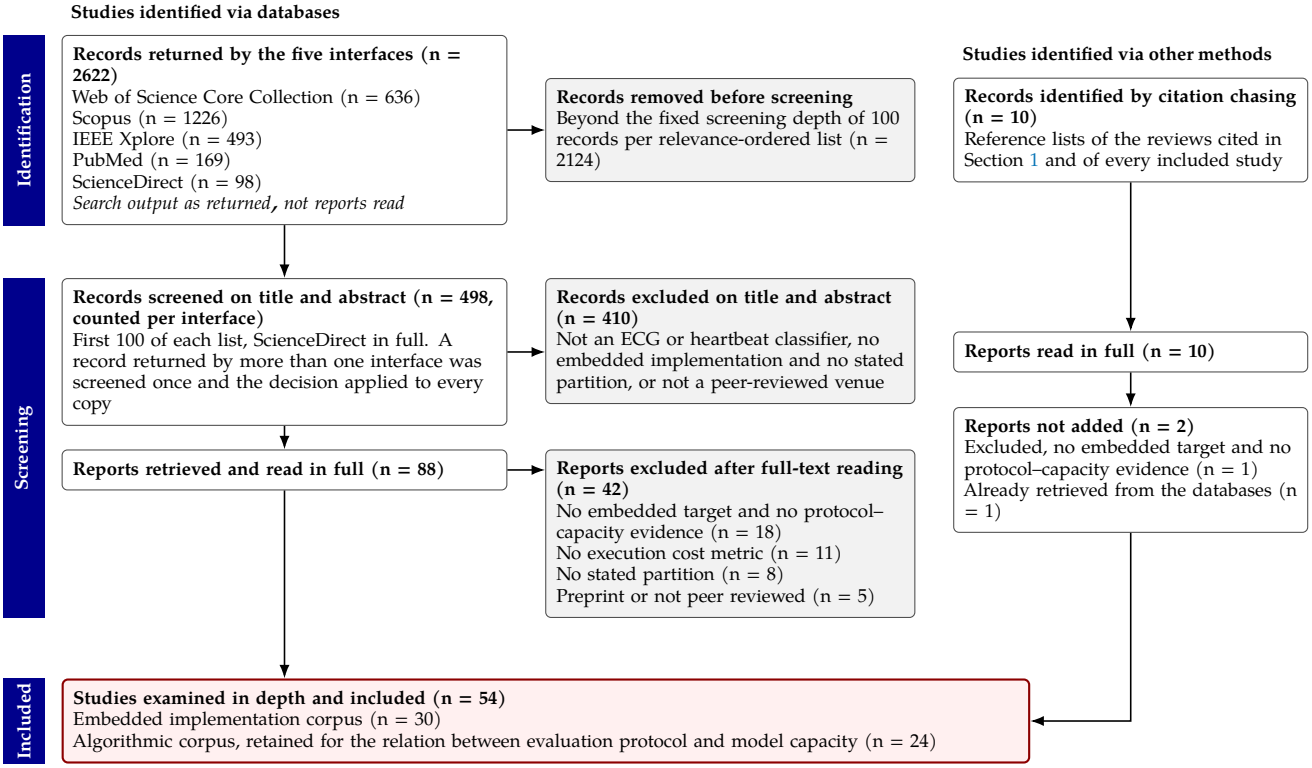


Figure 2. Flow of records through the review, following the PRISMA 2020 template [26]. The left branch reports the records returned by the five interfaces in Table 1, the records removed before screening and the records excluded at each stage. The right branch reports the records found by citation chasing. Each report discarded after full-text reading was assigned exactly one reason, the first to apply in the order listed in Section 2.

2.1. Scope of the Corpus and Conduct of the Review

The corpus is restricted to ECG classifiers carried to an embedded implementation, because the object of the analysis is the consequence of the evaluation protocol for the hardware that hosts the classifier. Studies of ECG processing and classification that report no hardware target contribute no execution cost, footprint or platform information, and were retained only where they document the relation between the evaluation protocol and the capacity a classifier requires, as a separate corpus. The review was conducted by a single author, and four decisions were taken to keep every step verifiable. First, the screening depth of the relevance-ordered lists was fixed in advance and the extraction was made verbatim, with every field published in the Supplementary Materials, so that the screened set and every extracted value can be reconstructed. Citation chasing in the reference lists of every included study complemented the screening and favored the retention of studies whose inter-patient result appears only in the full text. Second, the risk-of-bias assessment was made with an instrument built for the combination of clinical evaluation and hardware reporting examined here, with each of its five domains tied to one field of the extraction sheet so that every judgement can be reproduced from the published data. Third, the results were synthesized narratively, because the surveyed studies differ in platform, technology, database, number of classes and partition, and the group means reported for the corpus are descriptive summaries of it. Fourth, the review was not entered in a prospective register, and the eligibility criteria, search strings, field restrictions, screening depth, extraction fields and risk-of-bias instrument were reported in full, in the manuscript and in the Supplementary Materials, so that every decision can be audited against a fixed specification.

2.2. Sources and Search Strings

Records were retrieved using five interfaces. Web of Science Core Collection served as the main source, while Scopus, IEEE Xplore, PubMed, and ScienceDirect provided additional coverage. The search focused on ECG classifiers implemented on embedded platforms. Studies lacking an implementation were included only when they documented the relationship between the evaluation protocol and the capacity required by a classifier. These were reported as a separate corpus. Queries combined three facets using conjunction, and terms within each facet were joined using disjunction. The signal facet included electrocardiogram, ECG, arrhythmia, and heartbeat. The model facet included classification, neural network, machine learning, convolutional, transformer, recurrent, support vector machine, and spiking. The implementation facet included embedded, microcontroller, FPGA, ASIC, accelerator, edge computing, and TinyML. A fourth facet (comprising evaluation, inter-patient, patient-specific, partition, and generalization) was used to stratify the retained records and to identify studies that reported more than one partition. Table 1 presents the string used for each interface, the field restriction, the date of the final search, and the number of records returned. ScienceDirect permits a maximum of eight Boolean connectors (a reduced string was used in that case). Searches were completed on 21 September 2026 and resulted in the retrieval of 2622 records in total.

The total number is the output of the five interfaces and not a count of reports examined. Each interface returns its records ordered by relevance, and the first 100 records from each list were screened on title and abstract, which gives 498 records once the 98 records from ScienceDirect are counted in full. Fixing the depth allows the screened set to be reconstructed by re-executing the strings in Table 1.

Table 1. Search strings as executed, with the field to which each was restricted, the date of the last execution and the number of records returned. All five interfaces were searched on the same date. The three facets were combined by conjunction and the terms within each facet by disjunction.

Interface	Field	String Executed	Records
Web of Science	Topic	`TS=(("ECG" OR "electrocardiogram" OR "arrhythmia" OR "heartbeat") AND ("classification" OR "neural network" OR "machine learning" OR "convolutional" OR "transformer" OR "recurrent" OR "support vector machine" OR "spiking") AND ("embedded" OR "microcontroller" OR "FPGA" OR "ASIC" OR "accelerator" OR "edge computing" OR "TinyML")) AND PY=(2021-2026)`	636
Scopus	Title, abstract, keywords	`TITLE-ABS-KEY(("ECG" OR "electrocardiogram" OR "arrhythmia" OR "heartbeat") AND ("classification" OR "neural network" OR "machine learning" OR "convolutional" OR "transformer" OR "recurrent" OR "support vector machine" OR "spiking") AND ("embedded" OR "microcontroller" OR "FPGA" OR "ASIC" OR "accelerator" OR "edge computing" OR "TinyML")) AND PUBYEAR > 2020 AND PUBYEAR < 2027`	1226
IEEE Xplore	Abstract	`(("Abstract":ECG OR "Abstract":electrocardiogram OR "Abstract":arrhythmia OR "Abstract":heartbeat) AND ("Abstract":classification OR "Abstract":"neural network" OR "Abstract":"machine learning" OR "Abstract":convolutional OR "Abstract":transformer OR "Abstract":recurrent OR "Abstract":"support vector machine" OR "Abstract":spiking) AND ("Abstract":embedded OR "Abstract":microcontroller OR "Abstract":FPGA OR "Abstract":ASIC OR "Abstract":accelerator OR "Abstract":"edge computing" OR "Abstract":TinyML))`, year range 2021–2026	493
PubMed	Title/Abstract	`(electrocardiogram[Title/Abstract] OR ECG[Title/Abstract] OR arrhythmia[Title/Abstract] OR heartbeat[Title/Abstract]) AND (classification[Title/Abstract] OR "neural network"[Title/Abstract] OR "machine learning"[Title/Abstract] OR convolutional[Title/Abstract] OR transformer[Title/Abstract] OR recurrent[Title/Abstract] OR "support vector machine"[Title/Abstract] OR spiking[Title/Abstract]) AND (embedded[Title/Abstract] OR microcontroller[Title/Abstract] OR FPGA[Title/Abstract] OR ASIC[Title/Abstract] OR accelerator[Title/Abstract] OR "edge computing"[Title/Abstract] OR TinyML[Title/Abstract]) AND 2021:2026[dp]`	169
ScienceDirect	Title, abstract, keywords	`(ECG OR electrocardiogram OR arrhythmia) AND classification AND (embedded OR FPGA OR microcontroller OR ASIC)`, year range 2021–2026.	98
Total			2622

Last executed September 2026.

2.3. Deduplication, Screening Workflow and Exclusion Criteria

The screening was proceeded in four stages with a fixed rule each, and Figure 2 reports the records entering and leaving every stage with the reason for each exclusion. In the first stage, the five relevance-ordered lists were exported with their bibliographic fields and truncated at the stated depth, and the 2124 records beyond it were reported as removed before screening. In the second stage, records returned by more than one interface were identified by exact match to the digital object identifier and, where none was available, by normalized title and first author. Such a record was screened once, the decision was applied to every copy, and the per-interface counts in Figure 2 were kept so that the screened set could be reconstructed interface by interface. In the third stage, each record was screened on title and abstract against three questions (whether the work concerns an ECG or heartbeat classifier, whether it reports an embedded implementation or a result under a stated partition, and whether it is a peer-reviewed journal or conference publication). A record failing any question was excluded, which removed 410 records, and a record whose abstract left any question open was carried to full-text reading. In the fourth stage, the 88 reports retrieved from the databases and the 10 reports reached by citation chasing were read in full against the eligibility criteria, and each excluded report was assigned exactly one reason, the first to apply in the following order. No embedded target and no evidence

on the relation between protocol and capacity (18 reports), an embedded target without any execution cost metric (11 reports), a classification result without a stated partition (8 reports), and a preprint or otherwise non-peer-reviewed record that had passed the earlier stages on its metadata (5 reports). Of the 10 reports reached by citation chasing, 1 was excluded for the first reason and 1 had already been retrieved from the databases. The first reason gathers the studies of ECG classification without a hardware target, whose results carry no execution cost or platform information for the analysis developed here. The 54 studies that constitute the two corpora were examined in depth, and every quantity reported in this review was extracted from that reading.

2.4. Eligibility Criteria and Data Extraction

Two categories of work were included, according to distinct criteria. The implementation corpus consists of ECG classifiers executed or synthesized for a physical embedded target, with at least one execution cost metric (latency, memory footprint, resource utilization, power, or energy). The algorithmic corpus contains studies without an implementation that report a classification outcome under a specified partition, since the relationship between model complexity and evaluation protocol can be inferred from such results and the documented architectural choices influence hardware mapping. The primary window spans 2021 to 2026, with earlier works retained if they establish protocols still in use. Only peer-reviewed journal and conference publications were considered. For each study, the following data were extracted verbatim: platform and technology, model family and topology, operation count, parameter count, arithmetic width, memory footprint, peak activation memory, latency, clock frequency, database, number of classes, all reported classification metrics, and the partitioning strategy. Multiply accumulate counts were converted using Ops = 2 MAC and annotated, since most studies employ fixed-point arithmetic. Data not reported were recorded as missing and never inferred, since the completeness of reporting is itself a finding of this review. The corpus comprises 30 embedded implementations (10 on microcontrollers, 14 on reconfigurable devices, and 6 custom circuits) and 24 algorithmic studies.

2.5. Risk-of-Bias Assessment

Each implementation was assessed across five domains, each selected to ensure that every judgement is based on a field extracted verbatim. The first domain examines whether the evaluation protocol is explicitly stated. The second evaluates whether the protocol separates subjects between training and test groups, with a low rating for inter-patient setups, an unclear rating for patient-specific arrangements, and a high rating in other cases. The third determines whether a per-class or prevalence-insensitive metric accompanies the aggregate value. The fourth considers whether the quality of the variant that is actually deployed can be distinguished from the best software result, given that several studies report a reduction sequence where only the final stage is implemented on the device. The fifth assesses whether a parameter count or a memory footprint is disclosed. Each domain was tied to a deployment consequence. A high rating in the first or second domain means that the reported quality cannot be mapped to the deployment scenario of a device, in the third that a collapse of minority-class recall would be invisible, in the fourth that the quality attributed to the device may belong to a software stage, and in the fifth that the on-chip residency check cannot be applied. The instrument is therefore a reporting audit whose ratings describe what each study makes verifiable. Table 2 gives the assessment in full. Twenty of the thirty implementations state a protocol, three separate subjects and five adopt a patient-specific arrangement, so twenty-two report subjects whose beats were available during training. Twelve report a per-class metric, twelve permit the deployed variant to

be identified and twelve report a capacity quantity. These distributions are the reason the group means of the corpus are reported under two conventions.

Table 2. Risk-of-bias assessment of the surveyed implementations, listed in two blocks to save space. Five domains are scored as low (L), unclear (U) or high (H). D1: the evaluation protocol is stated. D2: the protocol separates subjects between training and test, scored low for an inter-patient partition, unclear for a patient-specific one and high otherwise. D3: a per-class or prevalence-insensitive metric is reported. D4: the quality of the variant actually deployed can be separated from the best software result. D5: a parameter count or a memory footprint is reported. Every judgement rests on a field extracted verbatim, and the assessment was applied by a single reviewer.

Work	D1	D2	D3	D4	D5	Work	D1	D2	D3	D4	D5
[27]	L	L	L	U	H	[28]	L	U	L	L	L
[29]	H	H	H	U	L	[30]	L	H	H	L	H
[31]	H	H	H	U	H	[32]	L	H	H	U	L
[33]	L	H	L	L	L	[34]	L	H	H	L	L
[35]	L	H	H	U	L	[36]	H	H	H	U	H
[37]	H	H	L	U	H	[38]	H	H	H	U	H
[39]	H	H	H	L	H	[40]	L	H	H	L	L
[41]	L	U	L	U	H	[42]	L	U	H	U	L
[43]	L	L	L	L	H	[44]	H	H	L	U	H
[45]	L	H	H	L	H	[46]	H	H	H	U	H
[47]	L	H	H	L	H	[48]	H	H	L	U	H
[49]	L	H	H	U	H	[50]	L	H	L	U	H
[51]	L	U	L	U	L	[52]	L	H	H	L	L
[53]	H	H	H	U	L	[54]	L	L	L	U	H
[55]	L	H	H	L	H	[56]	L	U	L	L	L
Total, L/U/H	20/0/10	3/5/22	12/0/18	12/18/0	12/0/18	across the 30 implementations					

3. Evaluation Protocols in ECG Classification

The way in which a database is partitioned determines which sources of variability appear in the test set and consequently influences the interpretation of the reported value. Four strategies were identified in the surveyed literature. However, the enforcement of these strategies is frequently undermined by preprocessing decisions that inadvertently transfer information from the test records to the training set. Both aspects are outlined in this section, as the analysis relies on distinguishing between protocols as described and protocols as implemented.

3.1. The Four Partitioning Strategies

The AAMI EC57 standard establishes the beat classes, the metrics, and the databases used for arrhythmia evaluation [6]. However, it does not specify how records are divided between training and test sets, which is the omission responsible for the divergence discussed here. Figure 3 illustrates the four strategies as distributions of beats into partitions.

Within the intra-subject strategy, beats from all records are pooled and divided randomly, so that beats of a test subject also appear in training (Figure 3a). This estimate demonstrates the ability to distinguish beat types within familiar recordings and typically overstates performance when applied to new patients [5,12]. The patient-specific strategy involves training a distinct model for each individual. A common pool, assembled from a fixed set of records, supplies the population component, and an initial labeled segment of the test record provides the personal component. According to the formulation of Kiranyaz et al., the pool contains 245 beats from records 100 to 124, while the adaptation beats are taken from the first five minutes of each test record, with the remainder serving as the test set [57]. In this approach, the division occurs along the time axis within each record (Figure 3b). First introduced for embedded deployment

by Kiranyaz et al. and later extended by Malik et al. with operational neurons [58], this protocol describes post-enrollment scenarios such as Holter monitors and implanted devices. The inter-patient strategy ensures that the record sets remain mutually exclusive, as in the DS1/DS2 division of MIT-BIH [5] (Figure 3c). This method accounts for morphological variability among subjects and is the only approach that evaluates a screening device deployed on an individual who has never been enrolled. The inter-session strategy retains the same subjects in both partitions and separates recordings by elapsed time or electrode re-application (Figure 3d). This approach is standard in surface electromyography, where the Ninapro datasets employ multi-day acquisitions [59], and is also used in ECG biometrics. Although no arrhythmia standard currently exists and the surveyed corpus does not contain any implementation evaluated under this protocol, the wearable patch represents the device best described by its use. Figure 3e arranges the four protocols according to the extent of variability accounted for by the resulting estimate.

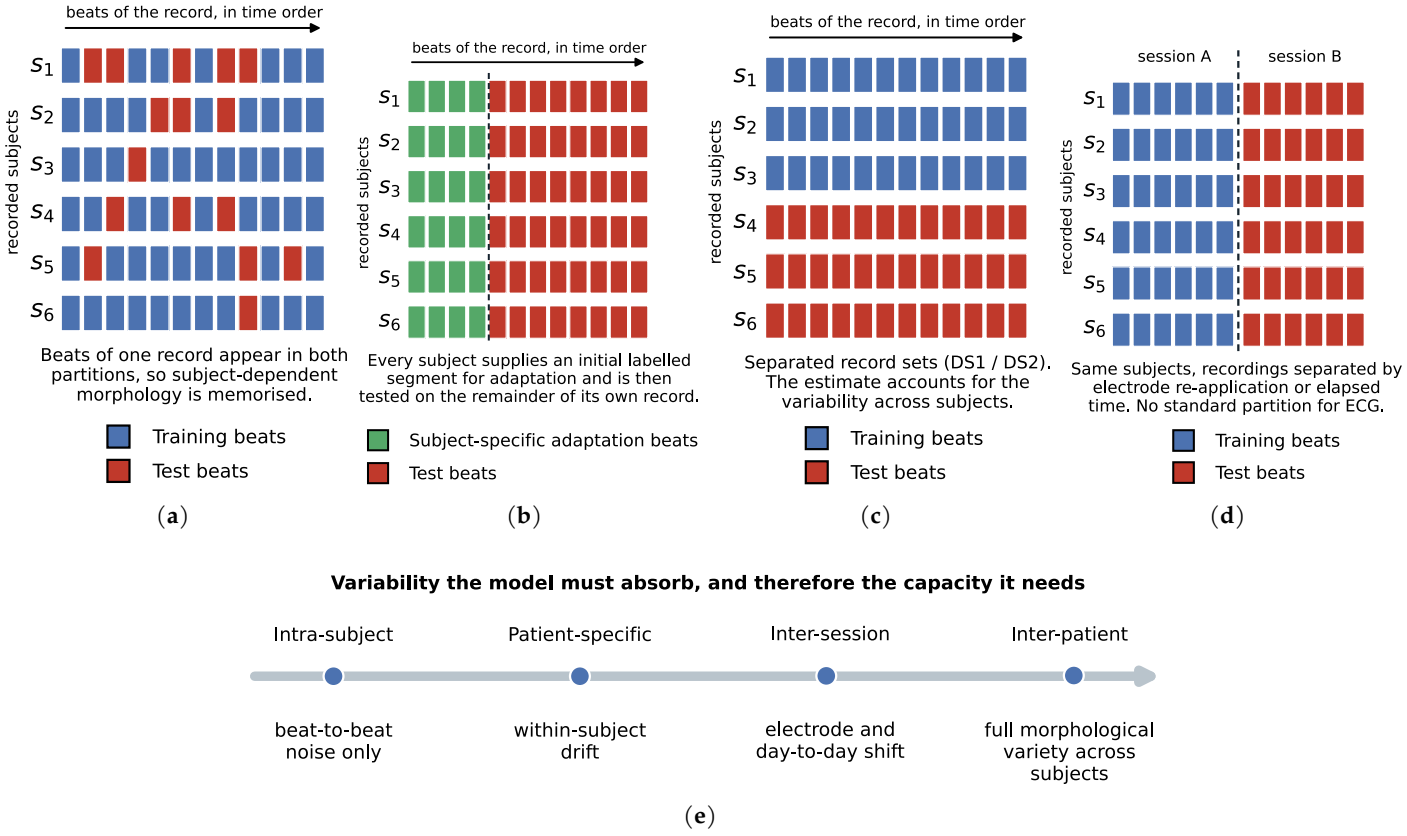


Figure 3. A visual taxonomy of the four evaluation protocols is presented. In each matrix, a row corresponds to one recorded subject $s_1, \ldots, s_6$ and a column represents one beat of that subject's record, arranged in chronological order. Each cell depicts a single beat and its color indicates the partition to which that beat has been assigned. (**a**) Intra-subject protocol ensures that beats from every subject are present in both partitions. (**b**) Patient-specific protocol uses a common pool compiled from a fixed set of records to supply the population component, and the initial segment of each test record provides the personal component for each individual, so the division occurs along the time axis within the test records and the records of the common pool do not contribute an adaptation segment. (**c**) Inter-patient protocol divides the subjects into two groups so that no subject contributes to both. (**d**) Inter-session protocol splits the same subjects along the time axis at a boundary defined by electrode re-application or elapsed time. (**e**) The four protocols are arranged in order according to the extent of variability accounted for by the resulting estimate.

3.2. Leakage Modes That Blur the Boundaries

A stated protocol and an enforced protocol are not synonymous. The general taxonomy of leakage (a test set not independent of training, preprocessing that uses the whole dataset,

a feature that proxies the label) has been shown to affect published results across more than a dozen fields [23], and the case of one subject on both sides of the boundary is the one for which the mismatch between record-wise and subject-wise cross-validation was quantified in clinical signal classification [3]. Four instances of that taxonomy were found in the surveyed corpus and are drawn in Figure 4. Segment-level random splitting results in nearly identical consecutive beats being allocated to both sides of the partition (Figure 4a). When the window stride is shorter than the window length, adjacent windows contain overlapping samples (Figure 4b). Minority oversampling conducted prior to partitioning introduces test information into training examples (Figure 4c). This issue is especially relevant for transformer studies that implement SMOTE-Tomek [60] or generative augmentation [61]. Normalization statistics calculated using the entire database transfer the distribution of the test set into the preprocessing phase (Figure 4d).

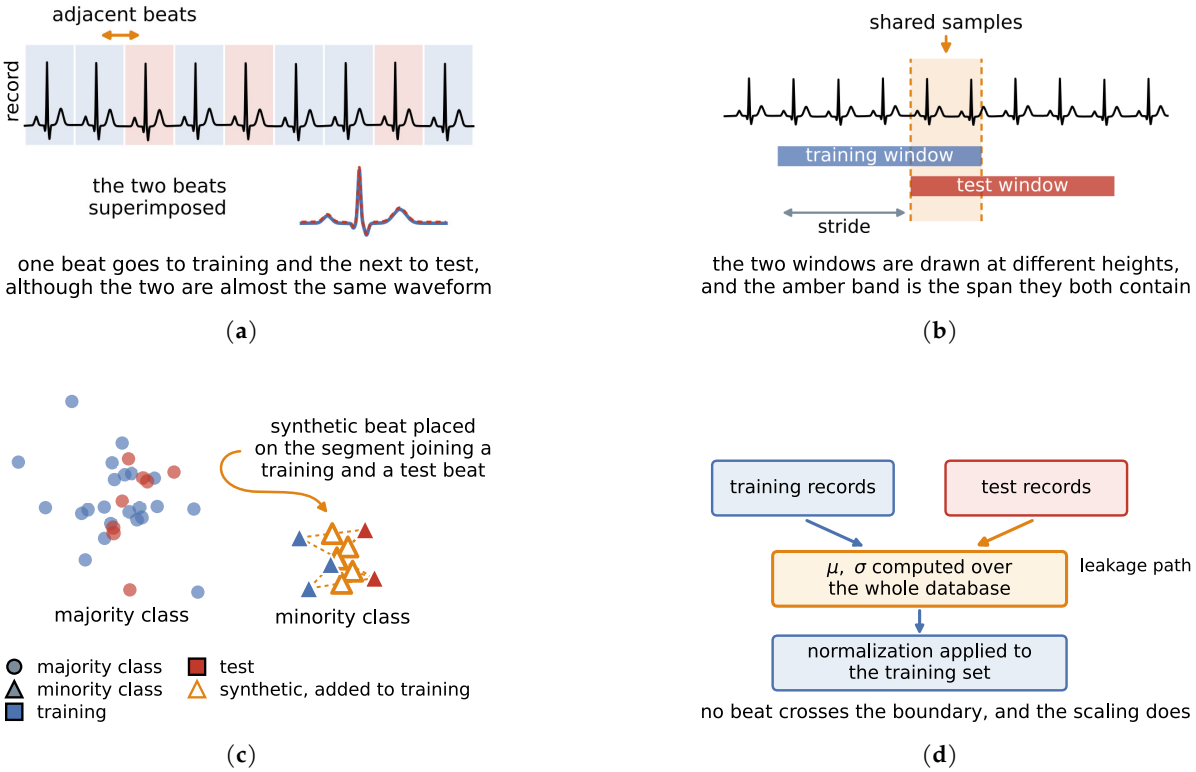


Figure 4. Certain practices transfer information from test records into the training process. Blue indicates training data, red represents test data, and orange illustrates the path taken by the transferred information. (**a**) Beat-level random splitting places nearly identical consecutive beats from one subject on both sides of the boundary due to random assignment. (**b**) When the window stride is shorter than the window length, two windows contain overlapping raw samples. (**c**) Minority-class oversampling conducted prior to partitioning generates a synthetic beat that falls between a training beat and a test beat. (**d**) Normalization statistics calculated using the entire database result in no beat crossing the boundary, yet the scaling does.

The resulting inflation is not consistent across classes, and the per-class impact possesses greater clinical significance than the aggregate value. In MIT-BIH, the aggregate accuracy is primarily influenced by the normal class, so a model may experience a substantial decline in its ability to identify supraventricular and fusion beats even as the overall accuracy remains nearly unchanged. Prevalence-insensitive metrics such as the Matthews correlation coefficient reveal this effect, as they reach high values only when all four quadrants of the confusion matrix demonstrate strong performance [62]. Another source of incommensurability is seldom disclosed. Studies employing MIT-BIH vary in whether the

four paced records are omitted, as required by the AAMI standard [6], in whether the unknown class is included, in whether five, four, or two classes are assessed, and in the beat-to-annotation association criterion. Luz and Menotti demonstrated that the selection of samples alters reported performance by a margin comparable to the differences ascribed to the models [63]. The group means reported for the corpus aggregate works that diverge in these decisions and are therefore descriptive.

3.3. Metrics for an Imbalanced Beat Population and for the Patient Level

The partition fixes which subjects the estimate accounts for and the metric fixes which errors it counts. In the DS2 test set of MIT-BIH the normal class accounts for roughly 89% of the beats, the supraventricular class for under 4% and the fusion class for under 1% [5], so a classifier that loses the supraventricular class entirely loses fewer than four points of accuracy. This arithmetic explains results such as the one in which an accuracy near 0.98 coexists with a macro-averaged F1 of 0.35 [64]. The AAMI standard requires sensitivity and positive predictivity per class [6], and twelve of the thirty implementations honor it. Three further families of metric are relevant to screening deployments. The first is the area under the precision–recall curve. The receiver operating characteristic normalizes the false-positive rate by the number of negatives, so a large majority class yields an area near unity while minority precision remains poor, and the precision–recall curve is the representation recommended for imbalanced classification [65]. The single area in the corpus, 0.9946 for a Cortex-M33 deployment, is a receiver operating characteristic area [36]. The second family is the pair of predictive values at a stated prevalence. Positive predictive value, under that name or as precision, is reported by six implementations [27,33,37,50,51,54], always at the prevalence of the test partition. Both values depend on prevalence through

$$\mathrm{PPV} = \frac{\mathrm{Se}\,\pi}{\mathrm{Se}\,\pi + (1-\mathrm{Sp})(1-\pi)}, \qquad \mathrm{NPV} = \frac{\mathrm{Sp}\,(1-\pi)}{\mathrm{Sp}\,(1-\pi) + (1-\mathrm{Se})\,\pi}, \tag{1}$$

where π is the prevalence of the class. A supraventricular detector with a sensitivity of 0.85 and a specificity of 0.99 has a positive predictive value of 0.77 at the 3.7% prevalence of DS2 and of 0.46 at a prevalence of 1%, which is closer to what an ambulatory screening device encounters, while its negative predictive value remains above 0.99 in both cases. A device whose purpose is to refer individuals to a cardiologist is judged by the first number, which cannot be read from a result reported at the prevalence of a research database. The third family is the set of patient-level metrics. Every value reported by the surveyed implementations pools beats across records, so a sensitivity of 0.9 can describe nine subjects detected in ten or nine beats in ten within every subject, and the beat-level value does not distinguish the two. The clinically interpretable quantities are the per-record sensitivity and specificity and the predictive values at the prevalence of the population screened, the level at which the ambulatory classifier validated against cardiologists was evaluated [66], at which the subject-separated evaluation of ECG abnormality classification has most recently been carried out [25], and at which the current scientific statement on artificial intelligence in clinical electrophysiology asks that performance be reported [67]. These three families are an aspect to consider in the evaluation of embedded classifiers. Since the works of the corpus report their quality through accuracy, per-class sensitivity, positive predictivity and F1, the description given here is informative, and the relation of these metrics with the hardware implications examined in this review will be established once embedded reports adopt them. The analysis in this systematic review is therefore kept to the metrics the corpus reports.

4. Classifier Families and the Techniques Employed

Five model families were identified in the corpus, with each associating its computational cost with a distinct resource, which determines the response to increased capacity requirements. The computational structure of the fully connected, convolutional, recurrent, transformer and spiking families is depicted in Figure 5a–e, respectively. A unified notation is employed and outlined in Figure 5f. Let $\mathbf{x}^{(i)} \in \mathbb{R}^{d_{\text{in}}}$ denote the input of sample i, where d_{in} represents the number of samples in one segmented beat, S_ℓ signifies the units of layer ℓ for $\ell = 1, \ldots, L$, $\omega_{j,k}^{(\ell)}$ indicates the weight from unit j in layer ℓ to unit k in layer $\ell + 1$, and $\mathbf{a}^{(\ell,i)} = f(\mathbf{z}^{(\ell,i)})$ is the activation vector with a fixed entry $a_0 = 1$ so that $\omega_{0,k}^{(\ell)}$ functions as the bias. The function $f_o(\cdot)$ denotes the output softmax that produces the class probabilities $P_c^{(i)}$. Three cost metrics are monitored for each family: the parameter count P, the multiply–accumulate operations per inference, and the peak activation memory M_{peak}, defined as the maximum volume of intermediate data simultaneously present during the forward pass. For every family, the quality reported by previous studies appears with the metrics provided by the respective authors, along with the mechanisms used to improve it.

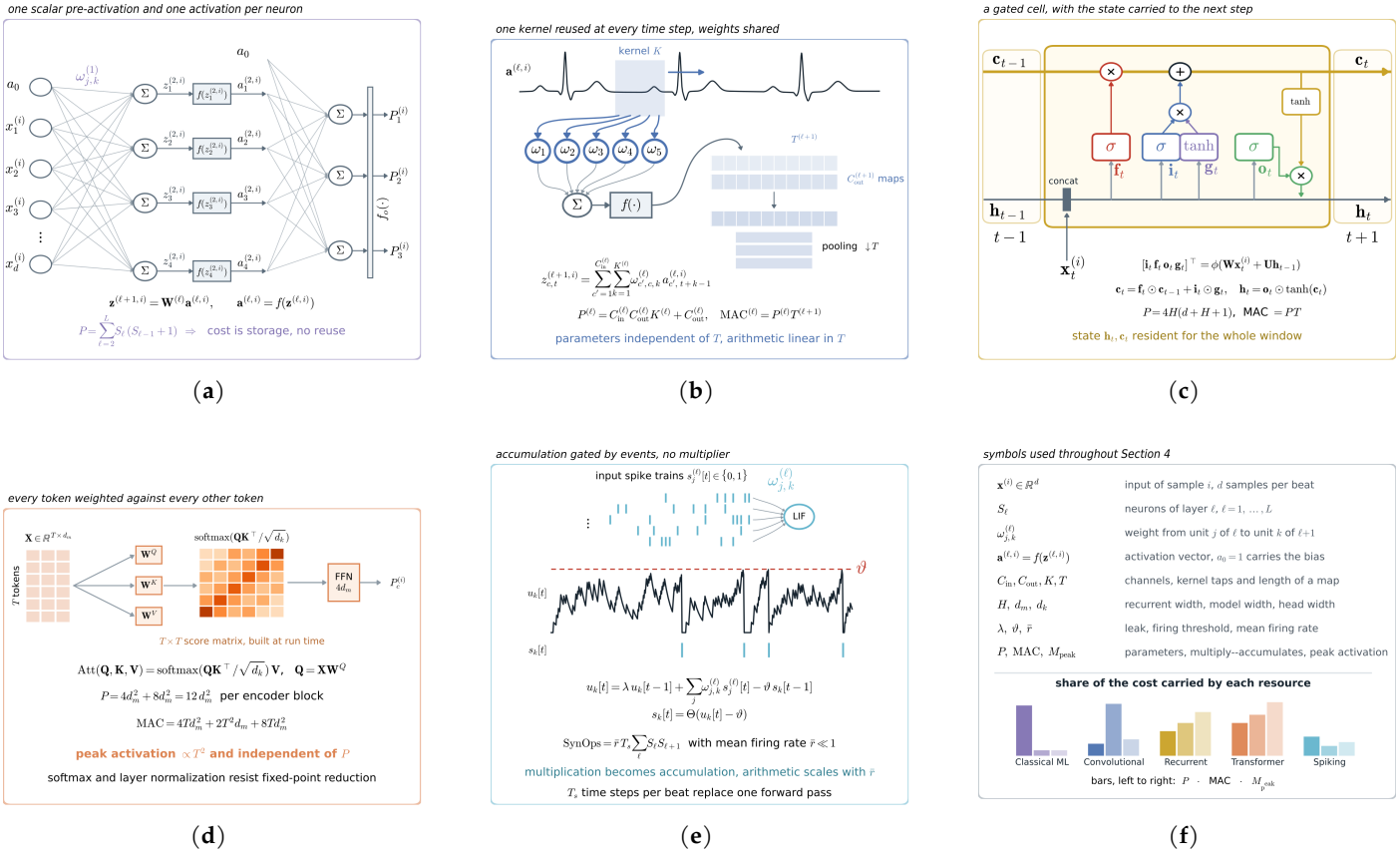


Figure 5. The computational structure of the five classifier families identified in embedded ECG work is depicted, with each family illustrated using the visual grammar specific to its operator. (**a**) The fully connected network assigns each neuron the computation of a scalar pre-activation $z_k^{(\ell,i)}$ followed by a single activation. (**b**) The convolutional network employs a kernel with K taps that is applied at every time step across a feature map of length T. (**c**) The recurrent network is shown unrolled over three time steps, with gate vectors and the state progressing forward. (**d**) The transformer encoder generates a $T_{\text{tok}} \times T_{\text{tok}}$ score matrix at run time from the projections of the token matrix. (**e**) The spiking network features a leaky integrate-and-fire unit that accumulates weighted input spikes and emits an event once the membrane potential surpasses the threshold ϑ. (**f**) The notation employed throughout Section 4 is presented along with the resource that contributes the majority of the computational cost in each family, specified as parameters P, multiply–accumulate operations, and peak activation memory M_{peak}.

4.1. Classical Machine Learning

A dense network computes a single weighted sum per unit followed by an activation:

$$\mathbf{z}^{(\ell+1,i)} = \mathbf{W}^{(\ell)}\mathbf{a}^{(\ell,i)}, \qquad a_k^{(\ell+1,i)} = f\big(z_k^{(\ell+1,i)}\big) \tag{2}$$

This relationship establishes that $P = \sum_{\ell=2}^{L} S_\ell(S_{\ell-1}+1)$ with $S_1 = d_{\text{in}}$ and MAC $\approx P$ accurate to within one percent for the widths present in the corpus. A kernel machine substitutes the layers with a decision function $g(\mathbf{x}^{(i)}) = \sum_{m=1}^{M} \alpha_m y_m \kappa(\mathbf{x}^{(i)}, \mathbf{x}_m) + b$, where the computational cost depends on the number M of support vectors. In both cases, each weight is employed once per inference, which results in computation and storage scaling together, and an increase in the number of classes leads to a doubling of both. Raj and Ray integrated a discrete orthogonal Stockwell transform with a least-squares twin support vector machine on an ARM9 processor, optimized using an artificial bee colony search, and evaluated it using three configurations on the same platform. The class-oriented configuration reports a sensitivity and positive predictivity of 96.29% with an F1 score of 76.06%, the personalized configuration achieves 96.08% across the three metrics, and the inter-patient configuration following de Chazal attains 86.89% [27]. The 9.4-point difference is achieved without any modification to the model, database, or preprocessing, and the inter-patient result is the one retained for this systematic review. At the opposite end of the spectrum, a two-layer dense network implemented on an Arduino Nano requires 2628 operations and 664 parameters, occupies 1.27 kB on an ATmega328 with 2 kB of static memory, and introduces a temporary dequantization scheme that maintains eight-bit weights in memory [33]. This reporting is informative, as an overall accuracy of 96.84% is accompanied by a macro-averaged F1 of 78.3%, with per-class values of 97.5% precision, 99.3% recall, and 98.4% F1 for the normal class; 78.3%, 52.3%, and 62.7% for the supraventricular class; and 93.8%, 45.3%, and 61.1% for the fusion class. A fusion recall of 45.3% in conjunction with 96.84% accuracy demonstrates the extent to which an aggregate metric may obscure deficits in minority classes. According to Equation (2), this family exhibits the lowest cost per parameter and the least advantageous scalability, which confines its application to problems involving a small number of well-separated classes.

4.2. Convolutional Networks

A convolutional layer applies a kernel of $K^{(\ell)}$ taps at every position of the input map:

$$z_{c,t}^{(\ell+1,i)} = \sum_{c'=1}^{C_{\text{in}}^{(\ell)}} \sum_{k=1}^{K^{(\ell)}} \omega_{c',c,k}^{(\ell)}\, a_{c',t+k-1}^{(\ell,i)}, \tag{3}$$

Here, $C_{\text{in}}^{(\ell)}$ and $C_{\text{out}}^{(\ell)}$ represent the input and output channel counts, and $T^{(\ell+1)}$ denotes the length of the output map. The parameter count is given by $P^{(\ell)} = C_{\text{in}}^{(\ell)} C_{\text{out}}^{(\ell)} K^{(\ell)} + C_{\text{out}}^{(\ell)}$, and the number of operations is $\text{MAC}^{(\ell)} = P^{(\ell)} T^{(\ell+1)}$. The parameter count remains unaffected by T, whereas the operation count increases with it. Consequently, a receptive field spanning an entire beat requires storage that is determined by K and the channel counts, and increasing depth enhances discriminative capacity without a corresponding rise in memory usage. Depthwise separable factorization modifies $C_{\text{in}}C_{\text{out}}K$ to $C_{\text{in}}K + C_{\text{in}}C_{\text{out}}$ [68].

Convolutional models are predominant in the embedded literature, spanning four orders of magnitude in complexity. Binarized architectures substitute multiplication with exclusive-NOR and population count, as demonstrated by Courbariaux and Wu [69,70]. Zhang and collaborators trained a patch-based binarized network on DS1 and recalibrated it on chip using the initial segment of each DS2 record. This approach, constituting a patient-specific classifier grounded in an inter-patient framework, increased accuracy from 93.81% to 97.55%, specificity from 68.00% to 89.15%, and F1 score from 94.39% to 97.42% [28]. The

21-point improvement in specificity, compared to a 3.7-point increase in accuracy, illustrates that the advantage of adaptation appears most clearly where the aggregate metric is least sensitive. A related design incorporating binarization, depthwise separable convolution, and a combined convolution–pooling operator achieves 96.61% accuracy across five classes with a macro-averaged F1 of 89.08% under patient-specific partitioning, and 96.96% accuracy, 97.82% sensitivity, and 98.21% positive predictive value in a two-class task [51]. Lu and colleagues developed a VGG-inspired accelerator that achieves 99.10% accuracy and sensitivity across five beat types, employing 11,065 parameters, a 70:30 beat partition, and sixteen-bit fixed-point arithmetic, along with a sign-bit dynamic activation strategy [35]. Liu and co-authors implemented a four-layer one-dimensional network directly on hardware without multiplexed units and report, under record-level partitioning of the Chapman database, an accuracy of 93.24% and a macro-F1 of 92.28% in floating point, and 92.95% accuracy and 92.05% macro-F1 after eight-bit quantization, representing decreases of 0.29 and 0.23 points, respectively [43]. Rawal and associates documented a complete reduction sequence under an 80:20 beat partition, progressing from an unconstrained network through architecture search and pruning to a multiplier-free design with logarithmic shifters. The sequence yields overall accuracies of 96.77%, 94.26%, and 90.80% across the three software stages and 86.37% on hardware, indicating a 10.4-point loss in the absence of any protocol change and providing the only comprehensive ablation identified in this review [45]. At the higher end of complexity, a dual-branch two-dimensional network requires approximately 2×10^7 operations per inference and achieves 97.79% accuracy in floating point, 97.54% accuracy at eight bits, and 96.89% at four bits, with a performance of 91.72% on an additional dataset [34]. Two principal observations emerge from Equation (3). Increasing the width of a layer scales P and MAC quadratically according to $C_{\text{in}}C_{\text{out}}$, whereas extending the kernel length affects them linearly. Designs achieving more than 97% accuracy at minimal computational cost tend to maintain modest channel counts [28,35], while increasing operations by another two orders of magnitude yields no performance improvement under comparable protocols [34].

4.3. Recurrent and Hybrid Models

A recurrent layer of width H updates an internal state at every time step:

$$[\mathbf{i}_t\ \mathbf{f}_t\ \mathbf{o}_t\ \mathbf{g}_t]^\top = \phi\big(\mathbf{W}\mathbf{x}_t^{(i)} + \mathbf{U}\mathbf{h}_{t-1}\big), \quad \mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t, \quad \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \tag{4}$$

The parameter count is given by $P = 4H(d_{\text{in}} + H + 1)$, and the number of multiply–accumulate operations is MAC $= PT$ over a window of T steps. The receptive field encompasses the window at a parameter cost that does not depend on T, a characteristic that aligns well with rhythm abnormalities governed by interval sequences. However, the states $\mathbf{h}_t$ and $\mathbf{c}_t$ persist throughout the window, which means M_{peak} includes a term proportional to H that scheduling cannot eliminate. The dependence of step t on $t-1$ precludes the systolic mapping described in Equation (3). Hou and colleagues introduced a long short-term memory autoencoder incorporating an unsupervised representation stage prior to the classifier [71]. Wu and co-authors implemented the family in silicon by combining a bidirectional recurrent stage with a preceding convolutional stage, taking advantage of the similarity between consecutive beats to reduce computation. They report 186.2 kB of on-chip static memory, 2.13 mW, and 2.78 µJ per beat on a 40 nm process [42], which constitutes the largest memory usage in the corpus and directly results from the recurrent state. Saadatnejad and associates distributed wavelet features across multiple recurrent branches, ensuring that the aggregate latency remains compatible with the beat rate on wearable processors [44]. Additionally, a convolutional–recurrent accelerator on an Intel Arria device achieves 95.3% accuracy with approximately 7.6×10^6 operations, utilizing

vendor-level quantization, pruning, and pipelining [46]. Temporal convolutional networks represent an intermediate approach, offering an extended receptive field combined with the parallelism of convolution. This approach has been validated for biopotentials on parallel ultra-low-power processors [72].

4.4. Transformers

A transformer encoder processes T_{tok} tokens of width d_m through three learned projections, a recombining projection and a pairwise comparison:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{Q}, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^{K}, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^{V}, \qquad \text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}, \tag{5}$$

This is followed by a position-wise feed-forward stage with an inner width of $4d_m$. Each block contains approximately $4d_m^2$ parameters allocated to the attention projections and $8d_m^2$ to the feed-forward stage, resulting in a total of $12\,d_m^2$ parameters. The block requires $4T_{\text{tok}}d_m^2 + 2T_{\text{tok}}^2 d_m + 8T_{\text{tok}}d_m^2$ multiply–accumulate operations. The $T_{\text{tok}} \times T_{\text{tok}}$ score matrix is computed at run time and contributes a term proportional to T_{tok}^2 to M_{peak}, independent of P. As a result, a transformer can be parameter-efficient and still surpass the memory limitations of its intended target, a characteristic unique to this model family. In exchange, every token can access every other token in a single operation, a feature the ECG literature leverages to establish correlations among beats throughout entire recordings [16,73]. The mechanisms introduced are categorized into four groups, each with distinct hardware requirements.

The first group introduces modifications to the stem responsible for generating the embedding. Hu and colleagues presented adaptive segmentation guided by the local RR interval (F1 99.83%) [74]. Guan and co-authors proposed a low-dimensional denoising embedding of wavelet and Fourier features preceding eight encoder layers (specificity 98.39%, sensitivity 98.41%) [75]. Chon and associates incorporated a multi-kernel residual stem with heart-rate-variability positional encoding, achieving 89.8% and 80.7% [76]. Liu and collaborators replaced fully connected layers with gated recurrent units fused with a convolutional branch, resulting in 89.93% and 84.06% [77]. Din and colleagues implemented a majority vote approach with convolutional, recurrent, and attention branches, attaining an F1 score of 99.34%, albeit at the expense of employing three parallel branches [78]. The second group focuses on modifications to the attention mechanism. Meng and associates substituted self-attention with a lightweight convolutional attention mechanism, eliminating 72% of the parameters in the attention block while maintaining 99.32% accuracy under a 7:1:2 intra-patient partition. This result provides strong evidence that attention performance can be preserved even with substantial reductions in capacity [50]. Tao and co-authors integrated refined diagonal and gated linear attention, achieving specificity of 95.22%, sensitivity of 97.50%, and approximately 50% faster convergence [79]. Wu and colleagues introduced a dilated stem combined with residual gated attention, reaching 95.7% accuracy, 88.1% sensitivity, and an average F1 of 82.6% across five classes. This design exhibits the largest gap between accuracy and F1, amounting to 13.1 points, among the attention mechanisms examined [80]. Wang and collaborators aligned ECG tags with segmented signals, reporting 99.40% accuracy, 99.41% specificity, and 97.36% sensitivity over twenty repetitions [81]. Ji and co-authors combined multi-scale grid attention with self-attention, reporting 99.28% accuracy under ten-fold cross-validation on twelve-lead data, compared with an average F1 of 0.847 and accuracy of 0.827 on a multi-hospital partition. They note that computational demand constrains practicality in resource-limited environments [82].

The third group addresses class imbalance. El-Ghaish and Eldele introduced a context-aware loss function with dynamic class weights, achieving 99.35% accuracy and 94.26% macro-F1 under an 80:20 partition, with class F1 scores of 99.41% for normal, 89.22% for

supraventricular, 96.69% for ventricular, 86.78% for fusion, and 99.18% for unknown, as well as fusion recall of 82.68% compared to 99.46% for normal beats [83]. Islam and co-authors applied SMOTE-Tomek resampling within a convolution and attention framework, obtaining 99.14% accuracy and 94.69% macro-F1 on MIT-BIH and 99.58% accuracy with 96.15% macro-F1 on INCART, and observe that minority-class performance continues to limit overall results [60]. Xia and colleagues generated synthetic minority beats using a transformer generator, reaching 94.69% accuracy [61], while Che and associates employed a link constraint loss function, achieving an F1 of 0.786 and accuracy of 83.66%, though they considered the computational burden impractical for portable devices [84]. The fourth group focuses on deployment. Busia and collaborators optimized the number of blocks, embedding size, and heads to yield 6,649 parameters within 49 kB, achieving 98.97% accuracy at eight bits compared to 99.05% in floating point under an intra-patient partition, with 0.97 million operations, 4.28 ms, and 0.09 mJ per inference on a parallel ultra-low-power microcontroller. This configuration uses sixty times fewer parameters and three hundred times fewer operations than comparable transformers [52]. Shan and co-authors minimized stem parameters through depthwise separable convolutions and squeeze-and-excitation blocks on a twelve-lead device, achieving specificity and sensitivity of 99.91% [48]. Two transformer studies provide inter-patient results. Xia and colleagues combined a lightweight transformer with convolutional and denoising autoencoder embeddings and RR features, reporting specificity of 99.37%, sensitivity of 99.91%, and performance gains concentrated in minority classes [85]. Liu and collaborators report a vision transformer for congestive heart failure, achieving 98.88% under the inter-patient scheme and above 99.8% under the intra-patient scheme [86]. Reporting results for both schemes using a single architecture exemplifies the recommended practice.

The per-class results for this family provide the most compelling evidence of the discrepancy between aggregate metrics and those that are clinically decisive. The lightweight fusing transformer, achieving 99.32% accuracy, reports 99.75% precision, 99.84% sensitivity, and 99.79% F1 for the normal class; 91.58%, 94.47%, and 93.00% for the ventricular class; and 93.86%, 83.00%, and 88.10% for the supraventricular class [50]. The supraventricular class, which is most critical for screening purposes, consistently exhibits the lowest sensitivity, ranging from 82% to 87%, in every attention design that reports per-class results, while aggregate accuracy remains above 99% [50,60,83].

4.5. Spiking Networks

A leaky integrate-and-fire unit accumulates weighted input events and emits an event when its membrane potential crosses a threshold:

$$u_k[t] = \lambda\, u_k[t-1] + \sum_j \omega_{j,k}^{(\ell)}\, s_j^{(\ell)}[t] - \vartheta\, s_k[t-1], \qquad s_k[t] = \Theta(u_k[t] - \vartheta), \tag{6}$$

In this context, λ denotes the leak, ϑ the threshold, and $\Theta(\cdot)$ the Heaviside function. Given that $s_j^{(\ell)}[t] \in \{0, 1\}$, the product simplifies to a conditional addition, which eliminates the need for a multiplier in the datapath. The computational workload is expressed as SynOps $= \bar{r}\, T_s \sum_\ell S_\ell S_{\ell+1}$, where T_s represents the number of time steps per beat and $\bar{r}$ is the mean firing rate. The sparsity of the ECG waveform ensures that $\bar{r}$ remains well below unity, resulting in T_s passes per beat and necessitating storage of each potential $u_k[t]$ throughout these passes. Xing and colleagues integrated a spiking network with an attention module on an Artix-7 device and report 92.07% accuracy under an inter-patient partition, compared to 98.26% accuracy, 94.75% sensitivity, 84.69% precision, and 89.09% F1 for the same model evaluated with a random 70:30 split. The energy consumption and inference latency are 346.33 µJ and 1.37 ms per classification, respectively [54]. The observed 9.2-point difference between accuracy and F1 under the permissive protocol indicates that errors

accumulate in minority classes, with the stricter partition further reducing performance by 6.2 points. Mao and co-authors trained an artificial network on DS1, converted it to a spiking network, and applied incremental on-chip training using the first labelled beats of each DS2 record, increasing accuracy from 93.67% to 97.36%, specificity from 55.54% to 82.30%, and F1 from 96.78% to 98.66%, while reducing multiply–accumulate operations by a factor of fifty [56]. Synaptic operations are not directly comparable with the operation counts defined in Equations (3) and (5), so these designs are discussed qualitatively and omitted from the quantitative comparison.

4.6. Predominance of Convolution in the Embedded Corpus

Convolutional models constitute the majority of embedded implementations, whereas attention mechanisms prevail in the algorithmic literature of the same period, for three principal reasons. The first pertains to structural characteristics. Equation (3) describes a fixed dataflow without sequential dependencies, featuring a static access pattern and weights that are reused at every position, which synthesis tools efficiently map onto systolic arrays or streaming pipelines. In contrast, Equation (5) generates a run-time intermediate that scales with T_{tok}^2 and incorporates a softmax and a layer normalization, both of which possess dynamic ranges that resist fixed-point reduction. Additionally, Equation (4) introduces step dependence, which precludes effective pipelining. Several transformer studies characterize their designs as impractical for portable devices [82,84]. The second reason involves arithmetic properties. Convolutional designs accommodate extreme quantization, achieving 97.55% accuracy after on-chip adaptation and 93.81% before it for the patch-based binarized network [28], and 96.61% for the depthwise separable binarized variant [51]. The binarized classifiers of Wong and colleagues attain 91.3% sensitivity, 98.1% specificity, and 86.7% precision on the ventricular task, as well as 85.4% sensitivity, 99.5% specificity, and 98.5% accuracy on the two-dimensional variant [37,41]. Furthermore, 98.59% accuracy is achieved using four-bit power-of-two quantization on a 65 nm processor [53]. In contrast, the score matrix and the softmax exponential restrict the bit width of attention mechanisms. The third reason is chronological in nature. Transformers were introduced to ECG classification in 2021, and all embedded transformer implementations in the corpus are of recent origin [48,50,52], suggesting that the technological gap is diminishing.

4.7. Trends Across Families

Three distinct patterns are evident. First, the majority of studies introduce a new topology accompanied by a single aggregate metric, and modifications in depth, width, kernel size, branch count, and fusion strategy alter quality within a range much narrower than the underlying complexity, which indicates that architecture selection is not the determining factor. Second, techniques with transferable benefit operate on the loss function or the data itself, such as context-aware weighting [83], resampling [60], generative augmentation [61], and constraint-based embedding regularization [84]. All of these methods improve minority-class metrics without increasing the parameter count and represent the only interventions in the corpus that enhance screening metrics while maintaining an unchanged computational footprint. Third, cost-reduction methods including depthwise separable factorization [48,51,68], lightweight convolutional attention [50], binarization [28,69], pruning [45,87,88], and quantization [34,89] decrease capacity through well-understood mechanisms while maintaining accuracy. Among these, only the fully mapped accelerator reports a minority-sensitive metric under a subject-separated partition, achieving 92.05% macro-F1 after eight-bit quantization [43]. The gap between overall accuracy and macro-averaged F1 is systematic and remains independent of model family, ranging from 18.5 points for a dense network on a microcontroller [33] to 0.9 points for the record-level

accelerator [43], including 9.2 points for a spiking accelerator [54], 7.5 points for a binarized accelerator [51], 5.1 points for a bidirectional transformer [83], and 4.5 points for a convolution and attention network [60]. The smallest discrepancy is observed in the single case evaluated at the record level, where the aggregate metric is least influenced by the majority class.

5. Model Complexity and the Binding Constraints of an Embedded Target

5.1. Three Quantities Conflated Under Complexity

The term complexity encompasses three distinct quantities that impose constraints on different resources. The operation count determines latency when parallelism and clock speed are set, the parameter count dictates the requirements for weight storage, and peak activation memory, defined as the largest volume of intermediate data present at any point during the forward pass, determines whether the model can be scheduled on a fixed datapath. A convolutional layer typically exhibits a high operation count with relatively few parameters, while a dense layer displays the opposite characteristics. Additionally, a wide early feature map may require substantially more activation memory than suggested by its parameter count. Studies focused on microcontroller deployment have consistently identified peak memory as the primary limiting factor, and inference runtimes address this constraint using patch-based scheduling and operator reordering [90,91]. This distinction is illustrated in two reports. Mendoza and co-authors deployed a quantized residual network on an ARM Cortex-M33 system on chip equipped with 256 kB of RAM and 1.5 MB of flash, reporting 709.6 kB of flash and 24.7 kB of runtime RAM, with inference times of 119 ms and an energy consumption of 0.95 mJ per inference at 78 MHz [36]. Another study from the architecture search literature describes a binary classifier requiring 86.6 kB of flash and 8.5 kB of peak RAM [92]. In both cases, weight storage exceeds volatile memory requirements by a factor ranging from seven to twenty-nine, indicating that a design evaluated solely on parameter count would be assessed for the incorrect resource.

5.2. A Common Resource Vocabulary Across Platform Classes

The surveyed implementations operate on three categories of target, each with resources defined and reported using different units. A microcontroller provides fixed static memory, fixed flash storage, and a processor with multiply–accumulate throughput determined by its instruction set and clock speed. A reconfigurable device offers block memory, look-up tables, and multiplier slices, which are utilized in proportions that depend on the specific mapping. A custom circuit features macros sized at the design stage and reports area, power, and energy consumption. Table 3 aligns the quantities required by a model, such as weight storage, live activation memory, signal buffering, program and runtime, operation throughput, latency, and energy, with the resource that accommodates each on every class, including the number of implementations that specify it. Table 4 presents the nominal resource budgets for each target in the corpus. On-chip storage is the only category where all classes report using the same unit, which justifies its use in Figure 6 and in the comparison of footprints with the on-chip memory of each device. The remaining categories elucidate why two designs with identical footprints may differ substantially in latency or power consumption, and clarify that a utilization percentage reported against a device chosen by the authors does not reveal the minimum resource requirements of the design.

Table 3. Correspondence between the quantities a classifier demands and the resource that carries each of them on the three platform classes of the corpus, with the number of the thirty surveyed implementations that state the quantity.

Quantity Demanded	Microcontroller or SoC	Reconfigurable Device	Custom Circuit	Reported By
Weight set, $P \times$ width	Flash, read in place or copied to SRAM at start-up	Block RAM or distributed LUT memory initialized at configuration (external DDR when exceeded)	On-chip SRAM, ROM or embedded non-volatile macro	12 of 30 (parameter count or footprint)
Peak activation memory M_{peak} and recurrent or membrane state	SRAM, shared with stack and runtime	Block RAM and pipeline registers (a streaming mapping holds one row of a feature map)	SRAM macro and datapath registers	0 of 30
Input buffer, $f_s \times$ window $\times$ sample width	SRAM, typically double-buffered under DMA	Block RAM or FIFO	SRAM or FIFO	Implied by the window length; never stated jointly with the sampling rate and the latency
Program, interpreter and runtime overhead	Flash for code, SRAM for stack and heap	Control logic in LUT and flip-flops, or a soft processor	Finite-state controller	Separated from the weight set by 1 of 30 [36], folded into a total footprint by 5 others
Operation throughput, MAC per inference	Cycles on the core (DSP and SIMD extensions on Cortex-M4 and later cores)	DSP slices and LUT, in parallel to the degree the mapping allows	Multiplier count and width fixed at design time	13 of 30 (operation count)
Latency and energy per inference	Cycles divided by clock, active and sleep current	Pipeline depth and clock, static plus dynamic power	Clock, supply voltage and area	Latency by 4, power or energy by 12

Table 4. Nominal budgets of the targets that appear in the surveyed corpus, as stated in vendor documentation (a range is given where the work does not identify the exact part). Volatile memory is the quantity drawn as a vertical rule in Figure 6. Values are rounded and intended for the residency check.

Class	Target (Works)	Compute	Volatile On-Chip Memory	Non-Volatile or Configuration Memory	MAC Resource
8-bit MCU	ATmega328P, Arduino Nano [33]	AVR, 16 MHz	2 kB SRAM	32 kB flash	8×8 hardware multiplier, two cycles
Cortex-M3 MCU	STM32F103 [40]	72 MHz	20–96 kB SRAM by part	64 kB–1 MB flash	32-bit single-cycle multiplier, no SIMD
Cortex-M4 MCU	STM32F429 [49]	180 MHz	256 kB SRAM	2 MB flash	DSP extension, single-cycle MAC, FPU
Cortex-M33 SoC	As reported [36]	78 MHz as run	256 kB RAM	1.5 MB flash	DSP extension, FPU
Parallel ULP	GAP9 [52]	9 RISC-V cores, to 370 MHz	128 kB L1, 1.5 MB L2	2 MB embedded MRAM	Per-core SIMD MAC and a convolution engine

Table 4. *Cont.*

Class	Target (Works)	Compute	Volatile On-Chip Memory	Non-Volatile or Configuration Memory	MAC Resource
Low-cost FPGA	iCE40 UltraPlus UP5K [37]	5280 LUT4	120 kbit block RAM plus 1 Mbit single-port RAM, about 143 kB	External SPI flash	8 DSP, 16×16
Mid-range FPGA	Artix-7 XC7A35T class, Arty [38,54]	20,800 LUT6	1.8 Mbit block RAM, about 225 kB	External flash	90 DSP48E1
SoC FPGA	Zynq-7020, PYNQ-Z2 and Zybo Z7-20 [30,32,34]	53,200 LUT6 and dual Cortex-A9	4.9 Mbit block RAM, about 630 kB	External DDR3	220 DSP48E1
SoC FPGA	Cyclone V SE, DE1-SoC class [28,43,51]	32,070 ALM and dual Cortex-A9	4.5 Mbit M10K, about 560 kB	External DDR3	87 variable-precision DSP
Large SoC FPGA	Zynq XC7Z045, ZC706 [35,39,47]	218,600 LUT6	19.1 Mbit block RAM, about 2450 kB	External DDR3	900 DSP48E1
Custom circuit	28–65 nm processes [29,31,41,42,53,56]	Fixed datapath	2.22–186.2 kB SRAM in the corpus	Design-defined	Multiplier count and width fixed at design time, reported as µW, µJ and mm^2

5.3. Accounting for the Deployment Footprint

The parameter count commonly serves as a proxy for memory and accounts for six of the twelve entries in Figure 6. It represents the first of five components. The memory occupied by a deployed classifier is

$$M_{\text{dep}} = \underbrace{P\, b_w/8}_{\text{weights}} + \underbrace{M_{\text{peak}}}_{\text{activations and state}} + \underbrace{f_s\, T_w\, b_s/8}_{\text{input buffer}} + \underbrace{M_{\text{prog}}}_{\text{code and runtime}} + \underbrace{M_{\text{rt}}}_{\text{stack, heap, descriptors}}, \quad (7)$$

where b_w denotes the weight width in bits, f_s represents the sampling rate, T_w refers to the window held in memory, and b_s signifies the sample width. These terms occupy distinct types of storage. On a microcontroller, the weights and code are stored in flash memory, while the remaining three terms reside in static memory. Mendoza and colleagues report 709.6 kB of flash and 24.7 kB of RAM, with the latter being the quantity most constrained by a Cortex-M budget [36]. On a reconfigurable device, all five terms contend for block memory except when the weights are streamed from external DDR memory, in which case the first term is moved off-chip and incurs the external access cost during each inference. In a custom circuit, the five terms are dimensioned independently, and the reported on-chip SRAM equals their sum. Two implications arise for Figure 6. An entry estimated solely from the parameter count represents only the first term and thus serves as a lower bound on the deployment footprint. The input-buffer term is particularly significant for a windowed classifier, as a 256-sample single-lead beat at 360 Hz occupies 512 bytes at sixteen bits, while a ten-second twelve-lead window at 500 Hz requires 120 kB, which exceeds the static memory available in the smallest microcontroller class shown in Table 4. It is therefore advisable that the five terms be specified individually, each assigned to the storage resource that contains it.

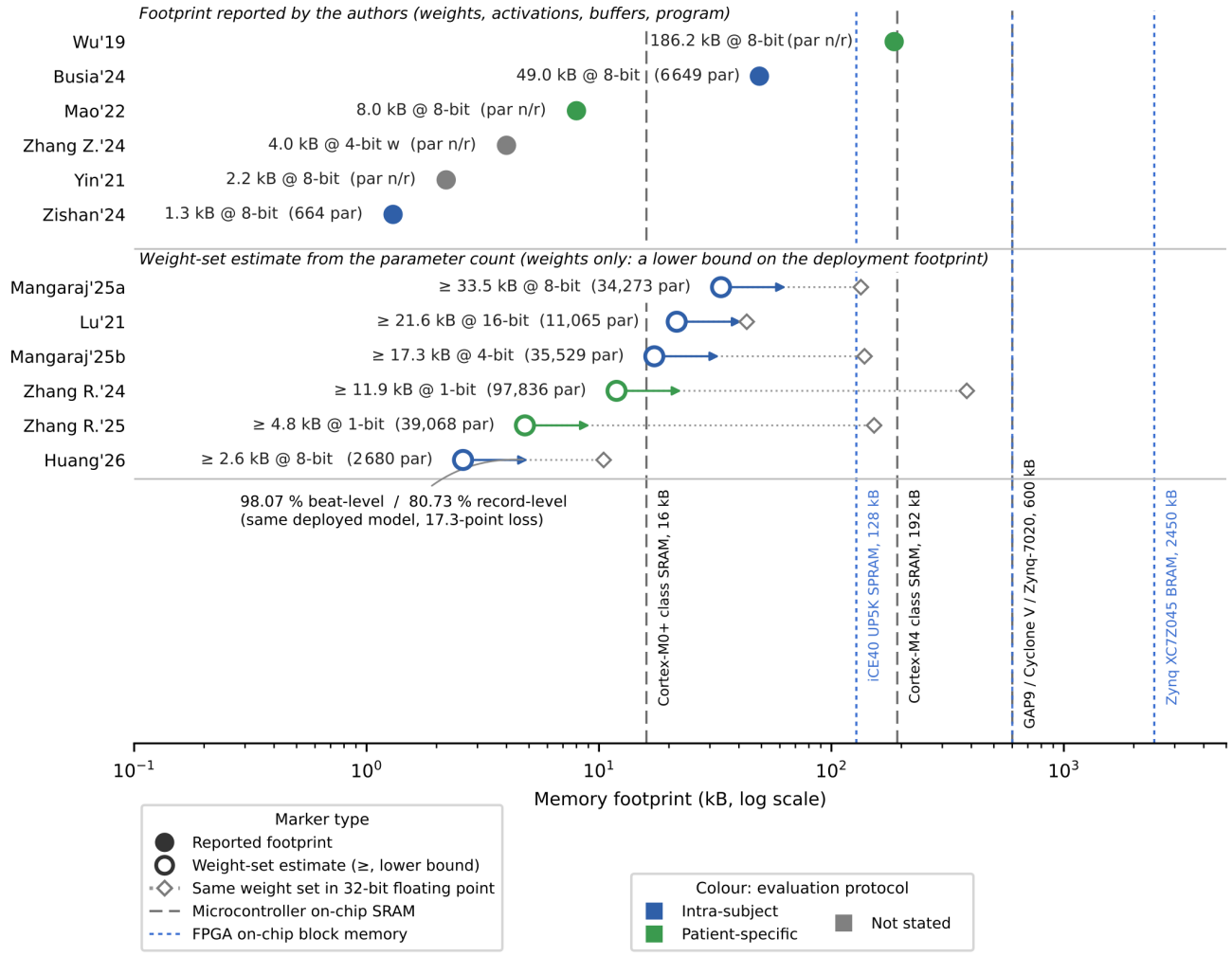


Figure 6. Memory footprint of the surveyed implementations is compared with the on-chip memory of representative embedded targets, focusing on the best-performing variant of each work. The entries are Wu'19 [42], Busia'24 [52], Mao'22 [56], Zhang Z.'24 [53], Yin'21 [29], Zishan'24 [33], Mangaraj'25a [32], Lu'21 [35], Mangaraj'25b [34], Zhang R.'24 [51], Zhang R.'25 [28] and Huang'26 [40]. Two kinds of entry are shown with distinct markers. Filled markers represent author-reported footprints, including activations, buffers and program storage. Hollow markers indicate weight-set estimates based on parameter count and deployed width, corresponding to the first term of Equation (7), and each is drawn with a right-pointing tail to indicate a lower bound for deployment. The grey diamond joined to each estimate represents the same weight set in 32-bit floating point, before quantization. Labels specify footprint, deployed width and parameter count if available. Vertical rules mark the on-chip memory of each device, as stated in Table 4. A model positioned to the right of a rule must fetch part of its weight set externally during inference. Colour shows the evaluation protocol. Twelve entries are included, six reported and six estimated. The entry with a second accuracy is the only implementation reporting both a capacity quantity and a result under a subject-separated partition, as discussed in the text.

5.4. On-Chip Residency

The role of memory is fundamentally energetic. The energy required for data access increases sharply with greater distance from the datapath. Reading from external dynamic memory consumes approximately one hundred times more energy than reading from a local register file [13]. This substantial difference has shaped accelerator design for the past decade [14,15]. When a model resides entirely on chip, only arithmetic energy is expended. If a model exceeds the available on-chip budget, the system incurs the external memory access cost during each inference, and this penalty does not increase in a continuous manner

with the size of the overflow. Residency is a binary property. Figure 6 compares the surveyed memory footprints with the capacities of representative target devices.

Four key features in Figure 6 support the argument. The deployed footprints range from 1.3 kB to 186.2 kB. Five of the twelve entries surpass the 16 kB limit of the smallest microcontroller class, and one exceeds the 128 kB of the low-cost reconfigurable device, indicating that the residency threshold is crossed within the corpus. When read at training width, the six weight sets estimated from parameter counts occupy between 10.5 kB and 382.2 kB in 32-bit floating point. Four of these exceed 128 kB and one exceeds the 192 kB of the mid-range microcontroller, demonstrating that quantization is essential for placing these models on chip. The segment joining the two markers of each row in Figure 6 quantifies the contribution of arithmetic to the residency condition. The third feature is that a footprint can be computed for twelve of the thirty implementations, while those reporting only an operation count span from 54,306 to 7.6×10^6 operations per inference. The fourth feature is that three implementations evaluated with an inter-patient partition report quality without parameter counts or footprints. Of the twelve entries, six are intra-subject, four patient-specific, and two do not specify a protocol. The application of the residency check to the inter-patient regime relies on the mechanism of the generalization-tax hypothesis stated in Section 1. One entry permits both conditions to be observed together. The pruned network of Huang et al. has 2680 parameters, an estimated weight set of 2.6 kB at eight bits, and executes on an STM32F103 with ample static memory. This model achieves 98.07% accuracy under a beat-level partition and 80.73% under a record-level, patient-independent partition [40]. The point is depicted in Figure 6 with both results, illustrating the predicted relationship. Residency is achieved, the model remains among the smallest in the corpus, yet the quality needed for a screening device is not attained. The capacity required to recover the 17.3-point difference is the measure that the generalization-tax hypothesis leaves open, and, for all other entries, the operative question concerns the extent to which the footprint can increase before crossing the device threshold.

The circuits literature provides the necessary quantitative data for this comparison. Zhang and colleagues implemented a two-stage network on a 65 nm process with 4 kB of on-chip SRAM, achieving 98.59% accuracy on the five AAMI classes using four-bit weights and consuming 0.15 µJ per inference at 1 V [53]. Cai and co-authors report an event-driven core operating at 1.3 µW and 1.8 µJ per classification with 98.58% accuracy [31]. Yin and collaborators describe an implantable processor consuming 2.63 µW, equipped with a 2.22 kB data buffer, and achieving 98.2% sensitivity [29]. Wu and associates report 186.2 kB of SRAM and 2.13 mW for a recurrent and convolutional classifier [42]. The binarized designs of Wong and colleagues operate with a 256 kbit RAM at 55.4 and 34.9 µW [37,41]. Consequently, the on-chip memory reported in the corpus ranges from 2.22 kB to 186.2 kB, which defines the interval within which the residency check is applied.

5.5. The Real-Time Budget

A latency value acquires significance only with respect to the deadline determined by the acquisition process. In beat-by-beat classification, the device receives one segmented beat as input, and the subsequent beat arrives following one RR interval. The deadline corresponds to the shortest interval the device must accommodate, expressed as $T_d = 60/\mathrm{HR}_{\max}$, which is 333 ms at 180 beats per minute and 273 ms at 220 beats per minute. For windowed classification, the input consists of a window of length T_w advanced by a hop T_h, making the hop duration the relevant deadline. In both modes, the processing chain must satisfy

$$t_{\mathrm{acq}} + t_{\mathrm{pre}} + t_{\mathrm{inf}} \le T_d, \qquad U = \frac{t_{\mathrm{pre}} + t_{\mathrm{inf}}}{T_d} < 1, \qquad \bar{P} \approx U\,P_{\mathrm{active}} + (1 - U)\,P_{\mathrm{sleep}}, \quad (8)$$

where t_{acq} denotes the delay before the input is complete, which corresponds to the post-R segment of the window for a beat centered on its R peak or to T_w itself for a windowed classifier. The variable t_{pre} represents filtering and segmentation time, t_{inf} is the inference latency, U indicates processor utilisation, and $\bar{P}$ signifies the average power when the processor is in sleep mode between inferences. The first relation expresses the real-time condition, the second determines the available margin, and the third converts this margin into the quantity by which a battery-powered device is dimensioned. The sampling rate influences both the input buffer in Equation (7) and t_{pre}. Evaluated against this budget, the 119 ms latency reported for Cortex-M33 deployment [36] is feasible in both classification modes with varying margins. For a beat-by-beat classifier operating at 180 beats per minute, 64% of the deadline remains unused, resulting in the processor being active for 36% of the time, which means the average power is determined by the active term in Equation (8). For a classifier executed once per ten-second window, only about one percent of the deadline is occupied, so the average power is governed by the sleep state. Other reported latencies are lower on this scale, including 25 ms on STM32F103 at 7.5% of the 333 ms deadline [40], 4.28 ms on GAP9 at 1.3% [52], and 1.37 ms on Artix-7 at 0.4% [54]. The recurrent design by Saadatnejad et al. was sized so that its aggregate latency across multiple branches remains below the beat rate [44]. The utilizations provided above are calculated under the beat-by-beat assumption, since sampling rate, window length, and inference period are reported separately from latency. The four quantities should therefore be reported together, and the real-time condition constitutes a second criterion alongside the residency check.

6. The Generalization Tax

6.1. What the Hypothesis States

The generalization tax, introduced in Section 1, refers to the additional capacity required to maintain constant quality when evaluating on unseen subjects. Observations in the literature typically keep the model fixed and alter the partition, thereby measuring the loss of quality at a constant capacity. In contrast, the generalization tax concept maintains quality and examines the capacity required, linking the two through a capacity-versus-quality curve under the inter-patient partition. The central claim of this review is therefore presented as a hypothesis regarding an underlying mechanism, and the relevant observations are organized in this section according to the strength of the control each provides.

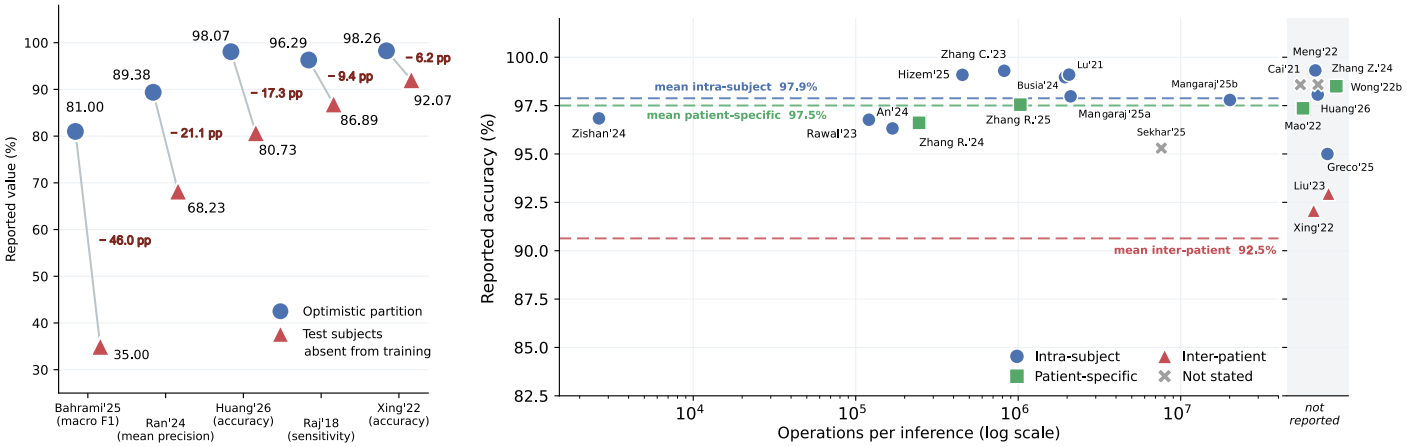
Hypothesis 1 (Generalization tax)**.** *For a fixed classifier family, database, and preprocessing, achieving a given minority-class quality requires greater capacity under an inter-patient partition compared to intra-subject or patient-specific partitions, since subject-specific morphology at test time must be replaced by an invariant representation. The additional capacity acts as a threshold to be surpassed, as further increases only improve fit to training subjects without aiding generalization. The hypothesis predicts a loss of quality at constant capacity when the partition changes, especially affecting minority classes, and asserts that this loss is not recovered simply by increasing capacity. The relationship between loss and capacity is parameterized by a coefficient that can be determined by varying capacity within a fixed inter-patient partition.*

The first two predictions are testable with the corpus and are examined in this section. The third requires a measurement not yet available in the literature.

6.2. Paired Evidence Within Single Studies

Comparisons across studies are complicated by differences in database, class set, preprocessing, and model, so the quantitative argument begins with studies that evaluate the same model under multiple partitions while keeping all other factors constant. Five such

cases were identified and are summarized in Figure 7a. These cases provide the primary evidence, and the reported accuracies of the corpus against operation count, shown in Figure 7b, provide the cross-study means that serve as a descriptive complement.



(**a**) (**b**)

Figure 7. Evidence bearing on the generalization-tax hypothesis. (**a**) paired results reported inside a single study, where model, database and preprocessing are held constant and only the partition changes (the primary quantitative argument): Bahrami'25 [64], Ran'24 [93], Huang'26 [40], Raj'18 [27] and Xing'22 [54]. (**b**) reported accuracy against operation count for every surveyed implementation that reports an accuracy, with marker shape and colour denoting the protocol, dashed rules giving the group means, and works without an operation count in the shaded column. The entries are Zishan'24 [33], An'24 [49], Rawal'23 [45], Hizem'25 [55], Zhang C.'23 [47], Lu'21 [35], Busia'24 [52], Zhang R.'25 [28], Zhang R.'24 [51], Mangaraj'25a [32], Mangaraj'25b [34], Sekhar'25 [46], Meng'22 [50], Cai'21 [31], Zhang Z.'24 [53], Wong'22b [41], Mao'22 [56], Huang'26 [40], Greco'25 [30], Liu'23 [43] and Xing'22 [54]. The groups differ in database, class set and preprocessing, and the inter-patient mean rests on two entries.

Bahrami and Fotouhi trained a convolutional network on MIT-BIH and evaluated it under three paradigms with the five AAMI classes. The macro-averaged F1 declines from 0.81 in the patient-specific paradigm to 0.66 in the intra-patient and 0.35 in the inter-patient setting, with patient-specific accuracy reaching 0.98. Under the inter-patient paradigm, the model classifies nearly every beat as normal [64], providing a clear illustration of the divergence between aggregate and per-class performance. Ran and colleagues assessed a multi-path convolutional network under both paradigms, observing accuracy decrease from 98.73% to 91.22%, average sensitivity from 94.89% to 89.91%, and average precision from 89.38% to 68.23%, a loss of 21.1 points, occurring mainly in the minority classes [93]. Huang and co-authors deployed a pruned network with 2680 parameters on an STM32F103, reporting 98.07% accuracy for a beat-level partition and 80.73% for a record-level, patient-independent partition, a 17.3-point loss with model, database, and preprocessing unchanged, representing the largest drop in aggregate accuracy in this corpus [40]. Xing and collaborators observed a decrease from 98.26% on a random split to 92.07% under an inter-patient partition, a difference of 6.2 points [54]. Raj and Ray reported a sensitivity of 96.29% under a class-oriented scheme and 86.89% in the inter-patient scheme of de Chazal on one platform, a loss of 9.4 points [27]. These five cases span a range of metrics and magnitudes, from 6.2 points of accuracy to 46 points of macro-averaged F1, but exhibit a consistent pattern. The loss is minimal on aggregate metrics and greater on minority-class-weighted metrics, which are critical for screening devices. All five cases are consistent with the first prediction of the hypothesis, and the two that report a metric weighted for minority classes support its concentration on these classes. None of the five

varies the capacity, so the second prediction is assessed using the one case that reports a capacity quantity, the network of Huang et al., in Section 5. Assertions that compression or quantization preserve diagnostic quality must therefore use a metric that a degenerate majority-class classifier would fail [62].

6.3. Cross-Study Group Means

Figure 7b displays the reported accuracies of the embedded corpus, demonstrating that the relationship between accuracy and complexity remains weak and non-monotonic across four orders of magnitude. A caveat pertains to the operation-count axis of Figure 7b. Binarized designs are included because the operation count for a binarized convolution is determined by its topology and remains comparable with its full-precision counterpart. However, the energy required per operation is not equivalent, so the axis quantifies structural operations only and does not reflect relative power consumption. The horizontal rules indicate the group means. For intra-subject, the mean is 97.9% ($n = 12$, SD 1.33, range 95.00–99.32). For patient-specific, the mean is 97.5% ($n = 4$ out of five entries, SD 0.67, range 96.61–98.50). For inter-patient, the mean is 92.5% ($n = 2$, SD 0.44, range 92.07–92.95), resulting in a separation of 5.4 percentage points. The inter-patient group comprises three works. Two report accuracy, with 92.95% for the fully mapped accelerator by Liu et al. [43] and 92.07% for the spiking design by Xing et al. [54], so the mean is calculated over these two. The third, the platform of Raj and Ray, provides sensitivity, positive predictivity, and F1 [27], and therefore is included in the paired evidence and risk-of-bias assessment, not in the mean. A similar approach is used for the patient-specific group, in which the processor of Wu et al. does not report accuracy [42]. The separation of 5.4 points based on twelve and two works constitutes the only cross-study comparison presented in this review and is highlighted in the abstract and the conclusions.

These means are calculated by selecting the best value reported in each work, following the convention established in comparison tables within this literature. This approach favors the intra-subject group, since most works documenting a full reduction sequence are found there and their initial stage operates in software. When recalculated using the actually deployed variant, which alters five entries, the intra-subject mean shifts to 96.2% and its standard deviation increases from 1.33 to 3.74. As a result, the separation narrows from 5.4 to 3.7 points, while the other groups remain unchanged. The separation persists under this stricter criterion, and the observed sensitivity of the intra-subject mean is informative, as one design loses 10.4 points between its unconstrained and multiplier-free forms [45]. Three cautions are warranted. The inter-patient and patient-specific groups consist of only two and four works respectively, so their means are not statistically robust. These groups also aggregate varying databases, class counts, and decisions about the paced records of MIT-BIH, factors that can affect performance by a similar margin [6,63]. Furthermore, the dispersion increases when per-class results are considered. The direction and magnitude of the observed separation are supported by findings outside this corpus. A systematic review that identified eleven studies reporting both intra-patient and inter-patient results found mean accuracies of 98.39% and 90.15%, yielding a separation of 8.2 points in an independently assembled sample [4]. The means reported in this review align with that result.

6.4. Why Quality Does Not Rise with Capacity Within a Protocol

Figure 7b presents an apparent contradiction. The intra-subject entries extend across the entire complexity axis, from 2628 operations to 2×10^7, while the inter-patient entries do not report operation counts and therefore lack a coordinate on this axis. Three factors

account for this pattern. First, intra-subject partitions impose no minimum capacity, as beats from test subjects present in training enable models of any size to leverage subject-specific morphology, and near-maximal accuracy can be achieved with a few hundred parameters [33]. Models at the higher end of the axis are intended for objectives other than accuracy, such as increased class count, two-dimensional or multi-branch inputs, or enhanced throughput [32,34,35]. Second, selection bias is present, since the reporting of inter-patient results entails a more stringent evaluation, and only studies focused on evaluation or generalization-oriented architectures pursue this, resulting in a small and moderate-capacity inter-patient sample [27,43,54]. Third, within the inter-patient group, increased capacity does not improve quality. Liu et al. report 92.95% with a four-layer network [43], and Xing et al. achieve 92.07% with a spiking network [54]. The primary challenge is the distribution shift across subjects, a phenomenon described in the clinical literature on dataset shift for models transferred between populations [24], since additional parameters fit the training subjects more closely without enhancing generalization. Effective interventions in this protocol influence the representation and class balance [83,85]. This observation motivates the hypothesis to be defined with a floor, not a slope. This form is suggested by the two inter-patient entries and the behavior of the intra-subject group, and aligns with interventions that enhance inter-patient quality. Two entries without a capacity coordinate illustrate the shape of the curve without determining it, and the floor is the aspect of the hypothesis that an inter-patient capacity sweep will address.

6.5. What the Evidence Measures and What It Does Not

The five paired cases assess a loss of quality at constant capacity, while the generalization tax is defined as the additional capacity needed to maintain constant quality. Converting between these two measures requires a capacity-versus-quality curve under a fixed protocol. Three published studies provide such curves for embedded ECG classification under intra-subject partitions. Risso and collaborators derive a Pareto front ranging from 0.91k to 5.36k parameters and from 50.3k to 293.5k operations, within which the optimal architecture increases accuracy by 1.03 points while eliminating 64.7% of parameters and 85.8% of operations compared to the hand-tuned baseline [94]. Berriche presents an integer-constrained search that achieves 98.21% at 51.5k multiply–accumulate operations for five classes and 98.7% at 38.3k for the binary classification task [92]. Ribeiro and co-authors evaluate the same network before and after quantization, reporting 93 kB, 7.65 ms, and 5.85 mJ per prediction on a smartphone at a cost of at most 0.3 points [95]. These curves define the shape of the trade-off within a protocol, and applying the same sweep under an inter-patient partition identifies the capacity threshold. The current status of the hypothesis is as follows. Its mechanism is derived from the cost structure of the model families in Section 4 and from the arithmetic of imbalance described in Section 3. The first prediction is confirmed by every paired observation found here and by the independent sample of Xiao et al. The second prediction aligns with the two inter-patient entries. The coefficient is the quantity that the inter-patient sweep will estimate, and the claims made throughout this work are limited to the mechanism, the direction of the association, and that measurement.

7. Decision Framework

The preceding sections delineate a procedure spanning from the clinical context to the platform, as illustrated in Figure 8. The deployment scenario determines the set of permissible protocols. Devices intended for the ambulatory screening of individuals who have not previously enrolled must be evaluated using inter-patient assessment. The capacity implication is the one formulated as a hypothesis in Section 6, a floor to be cleared before the problem can be represented, beyond which capacity yields no measurable return. The design

rule that follows holds for any position of the floor. The objective is to reach it at the lowest footprint and to spend the remaining effort on the representation and the class balance, which the paired evidence shows to act on the minority classes. A device that monitors one enrolled patient over time may use a patient-specific configuration, a less demanding task that admits a more compact model at the cost of an enrollment procedure and the memory for on-device adaptation [56,57]. Patches removed and reapplied over several days occupy an intermediate position without a standard protocol, and the corresponding row in Figure 8 is drawn as a hypothesis.

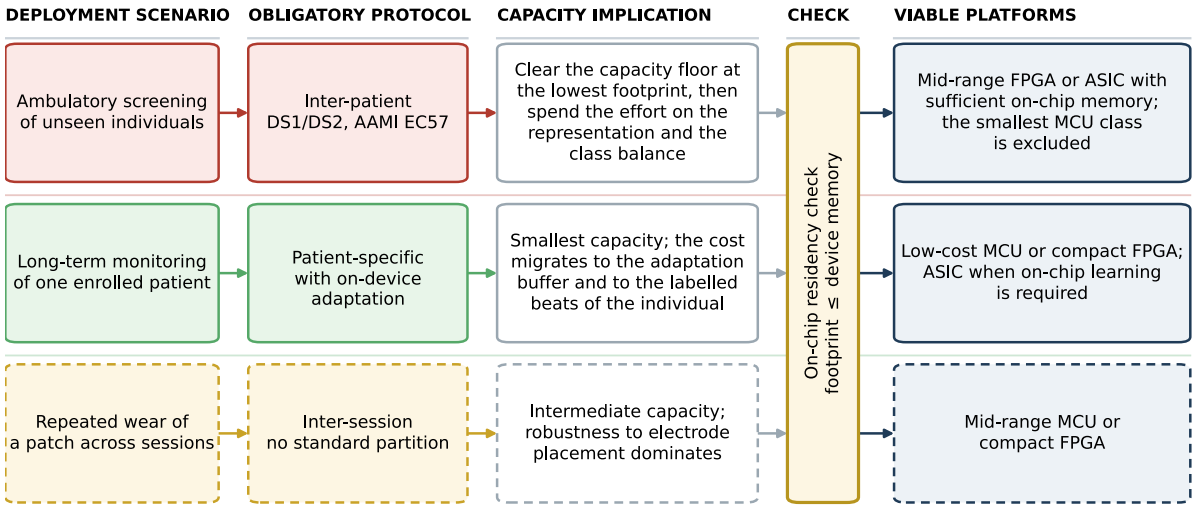


Figure 8. Decision framework linking the deployment scenario to the evaluation protocol it obliges, the capacity that protocol implies, and the platform classes that remain viable once the on-chip residency check is applied. The inter-session row is drawn from the physiological argument of Section 3 and not from surveyed evidence, since the corpus contains no implementation evaluated under that protocol.

The residency check constitutes the final stage. After the protocol establishes a target capacity, the footprint at the intended width is compared with the on-chip memory of the candidate device. If the footprint remains below the threshold, the device qualifies. If it exceeds the threshold, the device is deemed unsuitable or the design is modified through quantization, structured pruning, or distillation from a larger teacher [49,87–89,96]. From this perspective, the compression literature functions to restore residency once the protocol has increased the capacity requirement. The residency check is applied together with the real-time condition in Equation (8). The operating mode fixes the deadline, and the latency of the candidate design on the candidate device determines the utilization and the average power. A design passes this second check when it meets the deadline and keeps the average power within the battery budget. The two checks are independent, since a model can be resident on chip and slow, or fast and larger than the on-chip memory, so both must be satisfied.

A statement of clinical scope precedes the application of the framework, because the evidence in Section 6 pertains to beat-level data, while the screening decision is formulated at the patient level. Each quantity in Table 5, every paired comparison in Figure 7a, and every group mean describe the classification of beats or short windows pooled across records. The inter-patient requirement derived from these analyses constitutes a necessary, though not sufficient, condition for a screening device. It is necessary because a classifier that fails to maintain beat-level performance on unseen subjects cannot yield a reliable patient-level output through any aggregation method. It is not sufficient because the patient-level outcome also depends on the aggregation rule that converts beat decisions into a record-level decision, on the condition prevalence as expressed in Equation (1), and

on per-record sensitivity and specificity. Consequently, the conclusions of this review are articulated for beat-level classification under inter-patient evaluation, and their extension to patient-level screening requires a validation step, as exemplified by the ambulatory classifier compared with cardiologists [66] and by the patient-level evaluation of ECG abnormality classification [25], a step required of any device asserting screening applicability.

Table 5 summarizes the corpus as the framework requires, with the protocol given the same weight as the hardware specification and every reported metric included.

Table 5. This table surveys embedded ECG classifiers, categorized by classifier family. Operation counts, parameter counts, and classification metrics are presented as originally reported by the respective authors. For studies reporting multiple stages, the variant actually implemented on the target platform is indicated in the final column. Acc = accuracy, Se = sensitivity, Sp = specificity, Pr = precision, Re = recall, PPV = positive predictive value, F1 = F1 score, macro-F1 = macro-averaged F1 across the AAMI classes. A dash denotes that the quantity was neither reported nor inferred. Values marked † were converted from a multiply–accumulate count as Ops = 2 MAC.

Ref./yr	Family	Platform	Ops	Params	Protocol	Reported Classification Quality (%)
[27] 18	Classical ML	ARM9	–	–	Inter	Se 86.9, PPV 86.9, F1 86.9 (inter-patient); Se 96.3, PPV 96.3, F1 76.1 (class-oriented)
[29] 21	Classical ML	ASIC 65 nm	–	–	n/s	Se 98.2; 2.22 kB on-chip buffer, 2.63 µW
[31] 21	Classical ML	ASIC	–	–	n/s	Acc 98.58; 1.3 µW, 1.8 µJ per classification
[33] 24	Classical ML	Arduino Nano	2628 †	664	Intra	Acc 96.84, macro-F1 78.3; N: Pr 97.5, Re 99.3, F1 98.4; S: Pr 78.3, Re 52.3, F1 62.7; V: Pr 92.8, Re 89.2, F1 91.0; F: Pr 93.8, Re 45.3, F1 61.1
[35] 21	Convolutional	Zynq XC7Z045	2,056,000 †	11,065	Intra	Acc 99.10, Se 99.10 (16-bit fixed point)
[37] 22	Convolutional	iCE40UltraPlus	54,306 †	–	n/s	Se 91.3, Sp 98.1, Pr 86.7 (ventricular ectopic beat task)
[39] 22	Convolutional	ZC706	–	–	n/s	Acc 98.99 at 70% sparsity, 0.1 pp below the dense baseline
[41] 22	Convolutional	ASIC	–	–	Patient	Acc 98.5, Se 85.4, Sp 99.5, Pr 93.3, F1 89.2
[43] 23	Convolutional	Cyclone V	–	–	Inter	Acc 93.24, macro-F1 92.28 (FP); Acc 92.95, macro-F1 92.05 (8-bit, deployed)
[45] 23	Convolutional	Zynq ZCU106	120,516 †	–	Intra	Acc 96.77/94.26/90.80 (software stages), 86.37 (multiplier-free, deployed); AF 99.17/98.95/98.70
[47] 23	Convolutional	Zynq ZC706	820,000 †	–	Intra	Acc 98.90 (2-conv)/99.30 (3-conv), 2.05 and 14.27 µJ per inference
[49] 24	Convolutional	STM32F429	168,368 †	–	Intra	Acc 96.32
[51] 24	Convolutional	Cyclone V	245,394 †	97,836	Patient	Acc 96.61, macro-F1 89.08 (5-class); Acc 96.96, Se 97.82, PPV 98.21 (2-class)
[53] 24	Convolutional	ASIC 65 nm	–	–	n/s	Acc 98.59 at 4-bit weights; 4 kB SRAM, 0.15 µJ per inference
[55] 25	Convolutional	Raspberry Pi	453,664 †	–	Intra	Acc 99.09 (orig.)/96.01 (quant.)/92.90 (pruned, deployed)
[28] 25	Convolutional	Cyclone V	1,030,950 †	39,068	Patient	Acc 93.81→97.55, Sp 68.00→89.15, F1 94.39→97.42 after on-chip adaptation
[30] 25	Convolutional	Zybo Z7-20	–	–	Intra	Acc 95.00 (baseline)/91.90 (7-bit, deployed)
[32] 25	Convolutional	PYNQ Z2	2,101,248 †	34,273	Intra	Acc 97.98 (10-fold cross-validation)
[34] 25	Convolutional	PYNQ Z2	20,086,538 †	35,529	Intra	Acc 97.79 (FP32)/97.54 (INT8, deployed)/96.89 (INT4); 91.72 (second database)
[36] 25	Convolutional	Cortex-M33	–	–	n/s	AUC 0.9946; 709.6 kB flash, 24.7 kB runtime RAM, 119 ms, 0.95 mJ
[38] 26	Convolutional	Arty	–	–	n/s	Acc 99.5; 200 mW dynamic, 0.0767 mJ per inference
[40] 26	Convolutional	STM32F103	–	2680	Intra	Acc 98.07 (beat-level); Acc 80.73 (record-level, patient-independent); 0.156 mJ, 25 ms
[42] 19	Recurrent/hybrid	ASIC 40 nm	–	–	Patient	186.2 kB on-chip SRAM, 2.13 mW, 2.78 µJ per beat, 1.40 mm^2
[44] 20	Recurrent/hybrid	Wearable SoC	–	–	n/s	Quality reported per AAMI class, no aggregate value given
[46] 25	Recurrent/hybrid	Intel Arria 10	7,600,000 †	–	n/s	Acc 95.30
[48] 22	Transformer	Custom device	–	–	n/s	Sp 99.91, Se 99.91 (myocardial infarction localization)
[50] 22	Transformer	Wearable target	–	–	Intra	Acc 99.32; N: Pr 99.75, Se 99.84, F1 99.79; S: Pr 93.86, Se 83.00, F1 88.10; V: Pr 91.58, Se 94.47, F1 93.00
[52] 24	Transformer	GAP9	1,940,000 †	6649	Intra	Acc 98.97 (8-bit, deployed)/99.05 (FP32)/98.36 (noisy)/98.65 (balanced mix)
[54] 22	Spiking	Artix-7	–	–	Inter	Acc 92.07 (DS1/DS2); Acc 98.26, Se 94.75, Pr 84.69, F1 89.09 (random 70:30)
[56] 22	Spiking	ASIC 28 nm	–	–	Patient	Acc 93.67→97.36, Sp 55.54→82.30, F1 96.78→98.66 after on-chip adaptation

8. Reporting Practice and Proposed Conventions

The embedded corpus exhibits a systematic hierarchy of reporting, as depicted in Figure 9a, with the declaration of the evaluation protocol by classifier family shown in Figure 9b. Each entry specifies its platform and reports at least one classification metric. Overall accuracy is presented in twenty-three of the thirty entries, while the remaining seven provide sensitivity, specificity, or class-wise F1 scores, reflecting a selection of summary metric and not a lack of evaluation. Thirteen entries disclose an operation count, and twelve report a parameter count or footprint. Peak activation memory (the limiting factor on a fixed datapath) and a prevalence-insensitive metric are the two elements identified for future reporting within the corpus. Three further items follow from Sections 3 and 5. The

precision–recall area, a predictive value at a declared prevalence and a patient-level result are items for future reporting across the corpus, one entry separates the non-volatile from the volatile part of its footprint [36], and the four entries that report a latency report the sampling rate, window length and inference period separately from it.

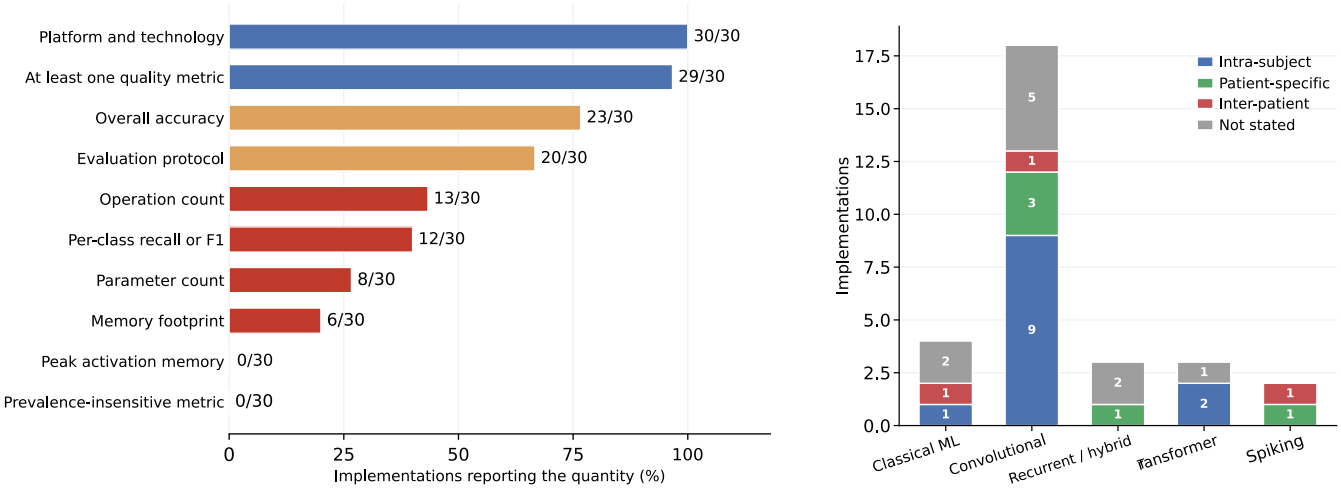


Figure 9. Reporting practice across the embedded corpus. (**a**) Fraction of implementations reporting each quantity. (**b**) Declaration of the evaluation protocol broken down by classifier family.

Figure 9b illustrates instances in which the protocol declaration is omitted. Twenty of thirty entries specify a partition. Convolutional entries indicate this specification in most cases, a practice shaped by two decades of scrutiny from the arrhythmia community [5,7,12]. Of the recurrent entries, two out of three omit the declaration, as does one of three transformer entries. Among the six designs from the circuits literature, only two declare a partition, demonstrating that omission occurs even in venues where hardware cost is reported with the greatest precision, consistent with communities in which random splitting is standard practice. Reporting templates for edge machine learning are available [97,98], but the surveyed implementations precede their adoption. The present analysis identifies ten reporting items not included in current templates, consistent with the items that the current scientific statement on artificial intelligence in clinical electrophysiology asks of a clinical ECG model, among them the training and test populations, the partition between them and the performance at the patient level [67]. They are summarized in Table 6.

Table 6. Required items for the protocol-to-platform comparison developed in this review, with the reason each is needed.

Item	Reason
Partitioning strategy, stated unambiguously	Classification quality is uninterpretable without it, and cross-study comparison is invalid when the protocols differ.
Inter-patient performance for any device claiming screening applicability	Screening encounters individuals who were never enrolled, so this is the only indicator that describes the intended use.
Parameter count with the arithmetic width used	Together, these give the weight footprint, which the residency check compares against device memory.
Peak activation memory	On a fixed datapath this is the binding constraint, and it cannot be inferred from the parameter count.
Smallest device on which the design fits	Utilization as a percentage of a device the authors selected conveys nothing about the cost floor.

Table 6. *Cont.*

Item	Reason
Per-class recall and a prevalence-insensitive metric alongside accuracy	Aggregate accuracy conceals the collapse of minority-class recall that evaluation on unseen subjects exposes.
Precision–recall area, and predictive values at a declared prevalence	The receiver operating characteristic is insensitive to prevalence, and a positive predictive value reported at the prevalence of a research database does not transfer to a screening population, Equation (1).
Per-record sensitivity and specificity for any device claiming screening applicability	Beat-level values pooled across records do not distinguish nine subjects detected in ten from nine beats in ten within every subject, and the screening decision is taken per subject.
The five terms of the deployment footprint, Equation (7), each assigned to the storage that holds it	The weight set is a lower bound on the footprint, flash and static memory are different budgets on a microcontroller, and block memory is shared by all five terms on a reconfigurable device.
Platform resources consumed in the units of Table 3, together with the nominal budget of the part	A utilization percentage without the budget of the part, or a footprint without the memory class it occupies, cannot be carried to another device.
Sampling rate, window length and inference period reported with the latency	A latency is interpretable only against the deadline of the operating mode, Equation (8), and the same latency can imply a utilization of one percent or of forty.

9. Research Directions for Complexity Reduction Under Inter-Patient Constraints

Portable ECG classification for screening necessitates inter-patient evaluation, and Section 6 demonstrated that increasing the parameter count alone is an inefficient approach. The central issue is thus how to minimize complexity while maintaining classification quality on unseen subjects. Six directions are presented, ordered from the most immediately actionable to the most exploratory.

9.1. Optimization Targeted at Minority-Class Recall

The interventions that enhance inter-patient capability without increasing capacity operate on the objective and the class distribution. Context-aware weighting [83], informed resampling [60], generative augmentation of minority beats [61], and class-constrained embedding regularization [84] each improve the screening metrics at a fixed parameter count, and their evaluation on embedded platforms under an inter-patient partition constitutes an accessible and cost-effective experiment. Operating points should be selected based on minority class recall and prevalence-insensitive metrics [62], since a compression that preserves accuracy but sacrifices supraventricular recall may satisfy the technical benchmark while failing the clinical one.

9.2. Representation Improvement in Place of Capacity Growth

The strongest inter-patient results are achieved by architectures that modify the input representation. Denoising autoencoder embeddings combined with RR features enhance minority-class performance under inter-patient partitioning with lower capacity compared to convolutional baselines [85], and low-dimensional time and frequency embeddings address the same challenge [75]. Interval features provide normalization across subjects, as ratios of consecutive RR intervals eliminate a substantial portion of the variability that would otherwise necessitate additional parameters. A systematic analysis of which handcrafted descriptors achieve the greatest reduction in across-subject variance per byte of memory would establish a valuable design guideline.

9.3. Efficient Attention as a Path to Inter-Patient Capability at Low Cost

Attention captures inter-beat context that fixed convolutional receptive fields do not address, which is critical for distinguishing rhythm from morphology abnormalities across subjects. The primary challenges involve the quadratic score matrix, as well as fixed-point implementations of softmax and layer normalization. Light convolutional attention eliminates 72% of the parameters of the attention block while achieving 99.32% accuracy under an intra-patient partition [50], refined and gated attention decrease computational requirements and accelerate convergence [79], and optimized encoders currently operate at 0.09 mJ per inference on microcontrollers [52]. Integrating these advances and assessing them under an inter-patient partition, with both peak activation memory and parameter count reported, represents the most promising unexploited combination identified in this analysis.

9.4. On-Device Adaptation as a Substitute for Capacity

In contexts where enrollment is feasible, subject-specific adaptation transforms the inter-patient problem into a patient-specific one and eliminates the capacity floor. On-chip learning implemented in custom silicon within 8 kB increased specificity from 55.54% to 82.30% [56], and personalized transfer learning delivered comparable improvements in related domains [99]. The outstanding questions concern the required volume of labeled data per individual, the memory requirements of the adaptation state, and the practicality of weakly supervised adaptation in the absence of clinician labels. A screening device capable of adapting after a brief supervised enrollment constitutes an intermediate paradigm that remains unexplored.

9.5. Cost Reduction Directed at the Inter-Patient Objective

Knowledge distillation, pruning, and architecture search are represented in the corpus with well-characterized compression factors, each focused on accuracy under the partition that the teacher, the pruning criterion, and the search objective inherit from the training set [49,87,88,96]. The published capacity-versus-quality curves [92,94] illustrate the impact of these methods, and the reduction sequence presented by Rawal et al., which documents quality at each stage from an unconstrained architecture to a multiplier-free implementation [45], serves as a reference model. The sweep that converts the generalization-tax hypothesis into a measured coefficient is specified as follows. A single family of models spanning two orders of magnitude of parameters, trained and evaluated under the DS1/DS2 partition with the five AAMI classes, reporting per-class recall, a prevalence-insensitive metric, the parameter count, peak activation memory and the deployed arithmetic width at every point, gives the capacity at which the minority-class quality of the intra-subject regime is recovered. The experiment is inexpensive relative to a hardware implementation. Event-driven computation belongs here too. The spiking designs reach the microwatt regime, and one evaluated under an inter-patient partition loses as much as conventional networks [54,56], so the encoding does not itself sacrifice generalization. A translation of synaptic operations into a comparable cost measure would allow them to enter the analysis in Section 6, and a cascade in which a cheap stage resolves the normal beats is untested under an inter-patient partition.

9.6. Standardization of the Missing Protocol

No inter-session protocol is currently established for arrhythmia classification, despite the fact that electrode placement varies across applications and morphology undergoes drift over extended periods of wear. Notably, electromyography has employed an inter-session standard for more than a decade [59]. Data suitable for such a protocol are already available

in long-term ambulatory ECG databases, and the protocol most closely aligned with the use of wearable patches stands to derive the greatest benefit from these resources. The inter-session case is an instance of dataset shift within one subject, and the practices the clinical literature proposes for shift between populations, periodic re-evaluation on data from the deployment condition and explicit reporting of the validation conditions [24], apply to it directly.

10. Conclusions

The evaluation protocol of an embedded ECG classifier has been argued to be a hardware design parameter. Four protocols are in use, they differ in the variability the estimate accounts for, and the difference propagates through the capacity the model needs to the memory it occupies and to the class of device that can host it. The capacity part of that chain is stated as a hypothesis, the generalization tax, and what the literature measures, a loss of quality at constant capacity, is separated from what the hypothesis asserts, an additional capacity at constant quality.

Three findings substantiate this argument. The primary evidence is derived from paired analyses. Five studies evaluated a single model under multiple partitions, maintaining constant database and preprocessing conditions, and replicated the observed performance loss under controlled circumstances. The magnitude of this effect varies by metric, ranging from a 6.2 percentage point decrease in accuracy to a macro-averaged F1 reduction from 0.81 to 0.35 in a model whose accuracy remains near 0.98 [27,40,54,64,93]. Across studies, reported quality is differentiated by protocol and only weakly by complexity, with a descriptive difference of 5.4 percentage points between the intra-subject and inter-patient group means over twelve and two works that differ in database, class set, and preprocessing. This finding aligns with the 8.2-point separation identified over eleven paired studies by an independent systematic review [4]. On-chip residency functions as a threshold (benign until a device boundary is crossed, and then immediately costly). The on-chip memory of the thirty implementations ranges from 2.22 kB to 186.2 kB [29,42]. The intra-subject group covers the entire complexity spectrum, as this protocol imposes no capacity floor, whereas, within the inter-patient group, quality does not increase with capacity; this observation prompts the hypothesis to be formulated as a floor to be cleared, a form suggested by the two inter-patient entries and one that an inter-patient capacity sweep will settle. The interventions that improve inter-patient capability act on the objective, the representation, and the class distribution, and incur minimal memory cost.

Three principal results are established. The first is a corpus comprising thirty embedded implementations drawn from 2622 records, stratified by protocol and assessed for risk of bias across five domains. Of these, twenty-two report on subjects whose beats were available during training. The second is an analysis of reporting practices, in which peak activation memory and a prevalence-insensitive metric, a predictive value at a declared prevalence, a patient-level result, the separation of the footprint by storage type and the acquisition parameters that make a latency interpretable are identified as the items whose adoption would make cross-study comparison exact, with the protocol omission reaching the circuits venues where execution cost is measured most precisely. The third is the first confrontation between deployed footprints and the on-chip memory of named devices over 1.3 kB to 186.2 kB, carried out on the twelve implementations that report a footprint and anchored to the inter-patient regime by the one implementation that reports both a capacity quantity and a record-level result [40]. The paired evidence quantifies a loss of quality at constant capacity, and the capacity-versus-quality curves published for embedded ECG classification were obtained under intra-subject partitions [92,94], so the mechanism and the direction of the association are established here and the coefficient is assigned to the capacity sweep under

a fixed inter-patient partition specified among the research directions. The conclusions hold for beat-level classification under inter-patient evaluation, a necessary condition for a screening device, and the patient-level validation that completes the argument is the next step. Devices intended for screening in settings without specialist support must be dimensioned for the individuals they will actually encounter, and the evidence assembled here gives the first estimate of what that requires.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/electronics15215046/s1. Table S1: PRISMA 2020 checklist. Table S2: PRISMA 2020 for Abstracts checklist. Table S3: Screening record. Table S4: Exclusion reasons at full-text reading. Table S5: Data dictionary of the extraction sheet. File S1: The extraction sheet of the embedded implementation corpus in comma-separated format. File S2: The extraction sheet of the algorithmic corpus.

Funding: This research was supported by the Secretariat of Science, Humanities, Technology and Innovation (SECIHTI).

Data Availability Statement: No new data were created in this study. The extraction sheet that supports every table and figure is provided as Supplementary Materials in comma-separated format.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AAMI	Association for the Advancement of Medical Instrumentation
AF	Atrial fibrillation
ASIC	Application-specific integrated circuit
DS1, DS2	Training and test record sets of the inter-patient partition
ECG	Electrocardiogram
F	Fusion beat, AAMI class
FP32	32-bit floating-point arithmetic
FPGA	Field-programmable gate array
INT4, INT8	4-bit and 8-bit integer arithmetic
LIF	Leaky integrate-and-fire
MAC	Multiply–accumulate operation
MCC	Matthews correlation coefficient
MIT-BIH	Massachusetts Institute of Technology, Beth Israel Hospital arrhythmia database
N	Non-ectopic beat, AAMI class
PPV	Positive predictive value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
Q	Unknown beat, AAMI class
RR	Interval between two consecutive R peaks
S	Supraventricular ectopic beat, AAMI class
SMOTE	Synthetic minority oversampling technique
SNN	Spiking neural network
SoC	System on chip
SRAM	Static random-access memory
SynOps	Synaptic operations
TCN	Temporal convolutional network
TinyML	Machine learning on resource-constrained embedded devices
V	Ventricular ectopic beat, AAMI class
ViT	Vision transformer

References

1. Chong, B.; Jayabaskaran, J.; Jauhari, S.M.; Chan, S.P.; Goh, R.; Kueh, M.T.W.; Li, H.; Chin, Y.H.; Kong, G.; Anand, V.V.; et al. Global burden of cardiovascular diseases: Projections from 2025 to 2050. *Eur. J. Prev. Cardiol.* **2025**, *32*, 1001–1015. [CrossRef]
2. Islam, M.S.; Kalmady, S.V.; Hindle, A.; Sandhu, R.; Sun, W.; Sepehrvand, N.; Greiner, R.; Kaul, P. Diagnostic and Prognostic Electrocardiogram-Based Models for Rapid Clinical Applications. *Can. J. Cardiol.* **2024**, *40*, 1788–1803. [CrossRef]
3. Saeb, S.; Lonini, L.; Jayaraman, A.; Mohr, D.C.; Kording, K.P. The Need to Approximate the Use-Case in Clinical Machine Learning. *GigaScience* **2017**, *6*, gix019. [CrossRef]
4. Xiao, Q.; Lee, K.; Mokhtar, S.A.; Ismail, I.; Pauzi, A.L.b.M.; Zhang, Q.; Lim, P.Y. Deep Learning-Based ECG Arrhythmia Classification: A Systematic Review. *Appl. Sci.* **2023**, *13*, 4964. [CrossRef]
5. de Chazal, P.; O'Dwyer, M.; Reilly, R.B. Automatic Classification of Heartbeats Using ECG Morphology and Heartbeat Interval Features. *IEEE Trans. Biomed. Eng.* **2004**, *51*, 1196–1206. [CrossRef]
6. Association for the Advancement of Medical Instrumentation. *ANSI/AAMI EC57: Testing and Reporting Performance Results of Cardiac Rhythm and ST Segment Measurement Algorithms*; Technical Report; Association for the Advancement of Medical Instrumentation: Arlington, VA, USA, 2012.
7. Silva, G.A.L.; Silva, P.H.L.; Moreira, G.J.P.; Freitas, V.L.S.; Gertrudes, J.C.; Luz, E.J.S. A Systematic Review of ECG Arrhythmia Classification: Embedded Feasibility, Adherence to Standards, and Fair Evaluation. *Res. Biomed. Eng.* **2026**. [CrossRef]
8. Ansari, Y.; Mourad, O.; Qaraqe, K.; Serpedin, E. Deep Learning for ECG Arrhythmia Detection and Classification: An Overview of Progress for Period 2017–2023. *Front. Physiol.* **2023**, *14*, 1246746. [CrossRef]
9. Gupta, U.; Paluru, N.; Nankani, D.; Kulkarni, K.; Awasthi, N. A comprehensive review on efficient artificial intelligence models for classification of abnormal cardiac rhythms using electrocardiograms. *Heliyon* **2024**, *10*, e26787. [CrossRef]
10. Moody, G.B.; Mark, R.G. The Impact of the MIT-BIH Arrhythmia Database. *IEEE Eng. Med. Biol. Mag.* **2001**, *20*, 45–50. [CrossRef]
11. Wei, Y.; Zhou, J.; Wang, Y.; Liu, Y.; Liu, Q.; Luo, J.; Wang, C.; Ren, F.; Huang, L. A review of algorithm & hardware design for AI-based biomedical applications. *IEEE Trans. Biomed. Circuits Syst.* **2020**, *14*, 145–163. [CrossRef]
12. Luz, E.J.d.S.; Schwartz, W.R.; Cámara-Chávez, G.; Menotti, D. ECG-Based Heartbeat Classification for Arrhythmia Detection: A Survey. *Comput. Methods Programs Biomed.* **2016**, *127*, 144–164. [CrossRef]
13. Horowitz, M. 1.1 Computing's Energy Problem (and What We Can Do About It). In *Proceedings of the 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*; IEEE: San Francisco, CA, USA, 2014; pp. 10–14. [CrossRef]
14. Chen, Y.H.; Krishna, T.; Emer, J.S.; Sze, V. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. *IEEE J. Solid-State Circuits* **2017**, *52*, 127–138. [CrossRef]
15. Sze, V.; Chen, Y.H.; Emer, J.; Suleiman, A.; Zhang, Z. Hardware for Machine Learning: Challenges and Opportunities. *In Proceedings of the 2017 IEEE Custom Integrated Circuits Conference (CICC)*; IEEE: Austin, TX, USA, 2017; pp. 1–8. [CrossRef]
16. Ansari, M.Y.; Qaraqe, M.; Charafeddine, F.; Serpedin, E.; Righetti, R.; Qaraqe, K. A Survey of Transformers and Large Language Models for ECG Diagnosis: Advances, Challenges, and Future Directions. *Artif. Intell. Rev.* **2025**, *58*, 261. [CrossRef]
17. Ghimire, D.; Kil, D.; Kim, S.h. A Survey on Efficient Convolutional Neural Networks and Hardware Acceleration. *Electronics* **2022**, *11*, 945. [CrossRef]
18. Liu, F.; Li, H.; Hu, W.; He, Y. Review of Neural Network Model Acceleration Techniques Based on FPGA Platforms. *Neurocomputing* **2024**, *610*, 128511. [CrossRef]
19. Xie, Y.; Oniga, S. A Comprehensive Review of Hardware Acceleration Techniques and Convolutional Neural Networks for EEG Signals. *Sensors* **2024**, *24*, 5813. [CrossRef]
20. Keivanimehr, A.R.; Akbari, M. TinyML and Edge Intelligence Applications in Cardiovascular Disease: A Survey. *Comput. Biol. Med.* **2025**, *186*, 109653. [CrossRef]
21. Oloko-Oba, M.; Esenogho, E.; Aruleba, K. From Biosignals to Bedside: A Review of Real-Time Edge Machine Learning for Wearable Health Monitoring. *Bioengineering* **2026**, *13*, 559. [CrossRef]
22. Ebrahimi, Z.; Loni, M.; Daneshtalab, M.; Gharehbaghi, A. A Review on Deep Learning Methods for ECG Arrhythmia Classification. *Expert Syst. Appl. X* **2020**, *7*, 100033. [CrossRef]
23. Kapoor, S.; Narayanan, A. Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. *Patterns* **2023**, *4*, 100804. [CrossRef]
24. Finlayson, S.G.; Subbaswamy, A.; Singh, K.; Bowers, J.; Kupke, A.; Zittrain, J.; Kohane, I.S.; Saria, S. The Clinician and Dataset Shift in Artificial Intelligence. *N. Engl. J. Med.* **2021**, *385*, 283–286. [CrossRef]
25. Hong, S.; Hartnett, C.; Ruane, M.; Tagliarino, J.; Nizich, M.; Toma, M. Electrocardiogram Abnormality Classification From Summary Statistics with Patient-Level Validation and Process-Focused Evaluation. *Ann. Noninvasive Electrocardiol.* **2026**, *31*, e70228. [CrossRef]
26. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71. [CrossRef]

27. Raj, S.; Ray, K.C. A personalized arrhythmia monitoring platform. *Sci. Rep.* **2018**, *8*, 11395. [CrossRef]
28. Zhang, R.; Zhou, R.; Han, X.; Qi, H.; Wang, Y. An energy-efficient ECG classifier with on-chip learning using binarized convolutional neural network. *IEEE Trans. Biomed. Circuits Syst.* **2025**, *20*, 107–121. [CrossRef]
29. Yin, Y.; Abubakar, S.M.; Tan, S.; Shi, J.; Yang, P.; Yang, W.; Jiang, H.; Wang, Z.; Jia, W.; U, S.-P. A 2.63 μW ECG processor with adaptive arrhythmia detection and data compression for implantable cardiac monitoring device. *IEEE Trans. Biomed. Circuits Syst.* **2021**, *15*, 777–790. [CrossRef]
30. Greco, L.; Moscato, F.; Ritrovato, P.; Vento, M. Fast and Low Cost FPGA-Based Architecture for Arrhythmia Detection with CNN. *Internet Things* **2025**, *33*, 101705. [CrossRef]
31. Cai, Q.; Xu, X.; Zhao, Y.; Ying, L.; Li, Y.; Lian, Y. A 1.3 μW event-driven ANN core for cardiac arrhythmia classification in wearable sensors. *IEEE Trans. Circuits Syst. II Express Briefs* **2021**, *68*, 3123–3127. [CrossRef]
32. Mangaraj, S.; Mahapatra, K.; Ari, S. Cardiac Arrhythmia Classification System: An Optimized HLS-Based Hardware Implementation on PYNQ Platform. *Microprocess. Microsyst.* **2025**, *120*, 105225. [CrossRef]
33. Zishan, M.A.A.; Shihab, H.; Islam, S.S.; Riya, M.A.; Rahman, G.M.; Noor, J. Dense Neural Network Based Arrhythmia Classification on Low-Cost and Low-Compute Micro-Controller. *Expert Syst. Appl.* **2024**, *239*, 122560. [CrossRef]
34. Mangaraj, S.; Mahapatra, K.; Ari, S. HLS-Compiled PYNQ-Based Cardiac Arrhythmia Detection System Leveraging Quantized ECG Beat Images. *Biomed. Signal Process. Control* **2025**, *109*, 108063. [CrossRef]
35. Lu, J.; Liu, D.; Liu, Z.; Cheng, X.; Wei, L.; Zhang, C.; Zou, X.; Liu, B. Efficient Hardware Architecture of Convolutional Neural Network for ECG Classification in Wearable Healthcare Device. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2021**, *68*, 2976–2985. [CrossRef]
36. Mendoza, A.; Nguyen, C.; Cavallaro, J.R. Real-time On-device Inference for ECG Classification of Ventricular Dysfunction. In *Proceedings of the 2025 59th Asilomar Conference on Signals, Systems, and Computers*; IEEE: New York, NY, USA, 2025; pp. 356–360.
37. Wong, D.L.T.; Li, Y.; John, D.; Ho, W.K.; Heng, C.H. An Energy-Efficient ECG Ventricular Ectopic Beat Classifier Using Binarized CNN for Edge AI Devices. *IEEE Trans. Biomed. Circuits Syst.* **2022**, *16*, 222–232. [CrossRef]
38. Cao, T.; Ng, W.S.; Goh, W.L.; Gao, Y.; Huang, H.W. FPGA-Based real-time ECG classification system using quantized Inception-ResNeXt neural network and CWT approximation. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2025**, *34*, 167–178. [CrossRef]
39. Lu, J.; Liu, D.; Cheng, X.; Wei, L.; Hu, A.; Zou, X. An efficient unstructured sparse convolutional neural network accelerator for wearable ECG classification device. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2022**, *69*, 4572–4582. [CrossRef]
40. Huang, J.; Wang, C.; Grau, A.; Xiao, Y.; Luo, S.; Che, Y.; Li, Z. EdgeECG: A lightweight edge-oriented network with dual criterion pruning for real-time ECG arrhythmia classification. *Physiol. Meas.* **2026**, *47*, 055014. [CrossRef]
41. Wong, D.L.T.; Li, Y.; John, D.; Ho, W.K.; Heng, C.H. Low Complexity Binarized 2D-CNN Classifier for Wearable Edge AI Devices. *IEEE Trans. Biomed. Circuits Syst.* **2022**, *16*, 822–831. [CrossRef]
42. Wu, J.; Li, F.; Chen, Z.; Pu, Y.; Zhan, M. A Neural Network-Based ECG Classification Processor with Exploitation of Heartbeat Similarity. *IEEE Access* **2019**, *7*, 172774–172782. [CrossRef]
43. Liu, W.; Guo, Q.; Chen, S.; Chang, S.; Wang, H.; He, J.; Huang, Q. A Fully-Mapped and Energy-Efficient FPGA Accelerator for Dual-Function AI-Based Analysis of ECG. *Front. Physiol.* **2023**, *14*, 1079503. [CrossRef]
44. Saadatnejad, S.; Oveisi, M.; Hashemi, M. LSTM-Based ECG Classification for Continuous Monitoring on Personal Wearable Devices. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 515–523. [CrossRef]
45. Rawal, V.; Prajapati, P.; Darji, A. Hardware Implementation of 1D-CNN Architecture for ECG Arrhythmia Classification. *Biomed. Signal Process. Control* **2023**, *85*, 104865. [CrossRef]
46. Sekhar, K.; K, P. Efficient detection of cardiac arrhythmias using low-latency fpga-accelerated hybrid deep learning models. *Eng. Res. Express* **2025**, *7*, 025277. [CrossRef]
47. Zhang, C.; Li, J.; Guo, P.; Li, Q.; Zhang, X.; Wang, X. A Configurable Hardware-Efficient ECG Classification Inference Engine Based on CNN for Mobile Healthcare Applications. *Microelectron. J.* **2023**, *141*, 105969. [CrossRef]
48. Shan, C.; Zhao, J.; Qiu, Z.; Wei, F.; Yuan, Z. Mobi-trans: A hybrid network with attention mechanism for myocardial infarction localization. In *Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN)*; IEEE: New York, NY, USA, 2022; pp. 1–8.
49. An, X.; Shi, S.; Wang, Q.; Yu, Y.; Liu, Q. Research on a Lightweight Arrhythmia Classification Model Based on Knowledge Distillation for Wearable Single-Lead ECG Monitoring Systems. *Sensors* **2024**, *24*, 7896. [CrossRef]
50. Meng, L.; Tan, W.; Ma, J.; Wang, R.; Yin, X.; Zhang, Y. Enhancing Dynamic ECG Heartbeat Classification with Lightweight Transformer Model. *Artif. Intell. Med.* **2022**, *124*, 102236. [CrossRef]
51. Zhang, R.; Zhou, R.; Zhong, Z.; Qi, H.; Wang, Y. A Multi-Class ECG Signal Classifier Using a Binarized Depthwise Separable CNN with the Merged Convolution-Pooling Method. *Sensors* **2024**, *24*, 7207. [CrossRef]
52. Busia, P.; Scrugli, M.A.; Jung, V.J.B.; Benini, L.; Meloni, P. A tiny transformer for low-power arrhythmia classification on microcontrollers. *IEEE Trans. Biomed. Circuits Syst.* **2024**, *19*, 142–152. [CrossRef]
53. Zhang, Z.; Guan, Y.; Ye, W. An Energy-Efficient ECG Processor with Ultra-Low-Parameter Multistage Neural Network and Optimized Power-of-Two Quantization. *IEEE Trans. Biomed. Circuits Syst.* **2024**, *18*, 1296–1307. [CrossRef]

54. Xing, Y.; Zhang, L.; Hou, Z.; Li, X.; Shi, Y.; Yuan, Y.; Zhang, F.; Liang, S.; Li, Z.; Yan, L. Accurate ECG Classification Based on Spiking Neural Network and Attentional Mechanism for Real-Time Implementation on Personal Portable Devices. *Electronics* **2022**, *11*, 1889. [CrossRef]
55. Hizem, M.; Bousbia, L.; Ben Dhiab, Y.; Aoueileyine, M.O.E.; Bouallegue, R. Reliable ECG Anomaly Detection on Edge Devices for Internet of Medical Things Applications. *Sensors* **2025**, *25*, 2496. [CrossRef]
56. Mao, R.; Li, S.; Zhang, Z.; Xia, Z.; Xiao, J.; Zhu, Z.; Liu, J.; Shan, W.; Chang, L.; Zhou, J. An ultra-energy-efficient and high accuracy ECG classification processor with SNN inference assisted by on-chip ANN learning. *IEEE Trans. Biomed. Circuits Syst.* **2022**, *16*, 832–841. [CrossRef]
57. Kiranyaz, S.; Ince, T.; Gabbouj, M. Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 664–675. [CrossRef]
58. Malik, J.; Devecioglu, O.C.; Kiranyaz, S.; Ince, T.; Gabbouj, M. Real-Time Patient-Specific ECG Classification by 1D Self-Operational Neural Networks. *IEEE Trans. Biomed. Eng.* **2022**, *69*, 1788–1801. [CrossRef]
59. Atzori, M.; Gijsberts, A.; Castellini, C.; Caputo, B.; Hager, A.G.M.; Elsig, S.; Giatsidis, G.; Bassetto, F.; Müller, H. Electromyography Data for Non-Invasive Naturally-Controlled Robotic Hand Prostheses. *Sci. Data* **2014**, *1*, 140053. [CrossRef]
60. Islam, M.R.; Qaraqe, M.; Qaraqe, K.; Serpedin, E. CAT-Net: Convolution, Attention, and Transformer Based Network for Single-Lead ECG Arrhythmia Classification. *Biomed. Signal Process. Control* **2024**, *93*, 106211. [CrossRef]
61. Xia, Y.; Xu, Y.; Chen, P.; Zhang, J.; Zhang, Y. Generative Adversarial Network with Transformer Generator for Boosting ECG Classification. *Biomed. Signal Process. Control* **2023**, *80*, 104276. [CrossRef]
62. Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef]
63. Luz, E.; Menotti, D. How the Choice of Samples for Building Arrhythmia Classifiers Impact Their Performances. In *Proceedings of the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*; IEEE: Nwe York, NY, USA, 2011; pp. 4988–4991. [CrossRef]
64. Bahrami, R.; Fotouhi, A.M. Investigation of Inter-Patient, Intra-Patient, and Patient-Specific Based Training in Deep Learning for Classification of Heartbeat Arrhythmia. *Cardiovasc. Eng. Technol.* **2025**, *16*, 333–342. [CrossRef]
65. Saito, T.; Rehmsmeier, M. The Precision–Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef]
66. Hannun, A.Y.; Rajpurkar, P.; Haghpanahi, M.; Tison, G.H.; Bourn, C.; Turakhia, M.P.; Ng, A.Y. Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. *Nat. Med.* **2019**, *25*, 65–69. [CrossRef]
67. Svennberg, E.; Han, J.K.; Caiani, E.G.; Engelhardt, S.; Ernst, S.; Friedman, P.; Garcia, R.; Ghanbari, H.; Hindricks, G.; Man, S.H.; et al. State of the art of artificial intelligence in clinical electrophysiology in 2025: A scientific statement of the European Heart Rhythm Association (EHRA) of the ESC, the Heart Rhythm Society (HRS), and the ESC Working Group on E-Cardiology. *Europace* **2025**, *27*, euaf071. [CrossRef]
68. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258. [CrossRef]
69. Courbariaux, M.; Hubara, I.; Soudry, D.; El-Yaniv, R.; Bengio, Y. Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or −1. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Barcelona, Spain, 5–10 December 2016.
70. Wu, Q.; Sun, Y.; Yan, H.; Wu, X. ECG Signal Classification with Binarized Convolutional Neural Network. *Comput. Biol. Med.* **2020**, *121*, 103800. [CrossRef]
71. Hou, B.; Yang, J.; Wang, P.; Yan, R. LSTM-Based Auto-Encoder Model for ECG Arrhythmias Classification. *IEEE Trans. Instrum. Meas.* **2019**, *69*, 1232–1240. [CrossRef]
72. Zanghieri, M.; Benatti, S.; Burrello, A.; Kartsch, V.; Conti, F.; Benini, L. Robust Real-Time Embedded EMG Recognition Framework Using Temporal Convolutional Networks on a Multicore IoT Processor. *IEEE Trans. Biomed. Circuits Syst.* **2020**, *14*, 244–256. [CrossRef]
73. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017.
74. Hu, S.; Cai, W.; Gao, T.; Zhou, J.; Wang, M. Robust wave-feature adaptive heartbeat classification based on self-attention mechanism using a transformer model. *Physiol. Meas.* **2021**, *42*, 125001. [CrossRef]
75. Guan, J.; Wang, W.; Feng, P.; Wang, X.; Wang, W. Low-Dimensional Denoising Embedding Transformer for ECG Classification. In *Proceedings of the ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*; IEEE: New York, USA, 2021; pp. 1285–1289.
76. Chon, S.; Ha, K.W.; Park, S.; Jung, S. An ECG Beat Classification Method Using Multi-Kernel ResNet with Transformer. In Proceedings of the 2023 IEEE International Conference on Big Data and Smart Computing (BigComp), Jeju, Republic of Korea, 13–16 February 2023; pp. 140–144.

77. Liu, Q.; Feng, Y.; Xu, H.; Li, J.; Lin, Z.; Li, S.; Qiu, S.; Wu, X.; Ma, Y.; Xu, Y. PSC-Net: Integration of Convolutional Neural Networks and Transformers for Physiological Signal Classification. *Biomed. Signal Process. Control* **2024**, *91*, 106040. [CrossRef]
78. Din, S.; Qaraqe, M.; Mourad, O.; Qaraqe, K.; Serpedin, E. ECG-Based Cardiac Arrhythmias Detection Through Ensemble Learning and Fusion of Deep Spatial-Temporal and Long-Range Dependency Features. *Artif. Intell. Med.* **2024**, *150*, 102818. [CrossRef]
79. Tao, Y.; Xu, B.; Zhang, Y. Refined self-attention transformer model for ECG-based arrhythmia detection. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–14. [CrossRef]
80. Wu, W.; Huang, Y.; Wu, X. SRT: Improved Transformer-Based Model for Classification of 2D Heartbeat Images. *Biomed. Signal Process. Control* **2024**, *88*, 105017. [CrossRef]
81. Wang, Y.; Yang, G.; Li, S.; Li, Y.; He, L.; Liu, D. Arrhythmia Classification Algorithm Based on Multi-Head Self-Attention Mechanism. *Biomed. Signal Process. Control* **2023**, *79*, 104206. [CrossRef]
82. Ji, C.; Wang, L.; Qin, J.; Liu, L.; Han, Y.; Wang, Z. MSGformer: A Multi-Scale Grid Transformer Network for 12-Lead ECG Arrhythmia Detection. *Biomed. Signal Process. Control* **2024**, *87*, 105499. [CrossRef]
83. El-Ghaish, H.; Eldele, E. ECGTransForm: Empowering Adaptive ECG Arrhythmia Classification Framework with Bidirectional Transformer. *Biomed. Signal Process. Control* **2024**, *89*, 105714. [CrossRef]
84. Che, C.; Zhang, P.; Zhu, M.; Qu, Y.; Jin, B. Constrained Transformer Network for ECG Signal Processing and Arrhythmia Classification. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 184. [CrossRef]
85. Xia, Y.; Xiong, Y.; Wang, K. A Transformer Model Blended with CNN and Denoising Autoencoder for Inter-Patient ECG Arrhythmia Classification. *Biomed. Signal Process. Control* **2023**, *86*, 105271. [CrossRef]
86. Liu, T.; Si, Y.; Yang, W.; Huang, J.; Yu, Y.; Zhang, G.; Zhou, R. Inter-Patient Congestive Heart Failure Detection Using ECG-Convolution-Vision Transformer Network. *Sensors* **2022**, *22*, 3283. [CrossRef]
87. Liu, Y.; Liu, J.; Tian, Y.; Jin, Y.; Li, Z.; Zhao, L.; Liu, C. Pruned Lightweight Neural Networks for Arrhythmia Classification with Clinical 12-Lead ECGs. *Appl. Soft Comput.* **2024**, *154*, 111340. [CrossRef]
88. Xie, Y.L.; Lin, C.W. YOLO-ReSTinyECG: ECG-Based Lightweight Embedded AI Arrhythmia Small Object Detector with Pruning Methods. *Expert Syst. Appl.* **2025**, *263*, 125691. [CrossRef]
89. Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.; Adam, H.; Kalenichenko, D. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2018; pp. 2704–2713. [CrossRef]
90. Lin, J.; Chen, W.M.; Lin, Y.; Gan, C.; Han, S. MCUNet: Tiny Deep Learning on IoT Devices. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 11711–11722.
91. Lin, J.; Chen, W.M.; Cai, H.; Gan, C.; Han, S. Memory-efficient patch-based inference for tiny deep learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 2346–2358.
92. Berriche, L. NanoCNN: Minority-Aware Neural Architecture Search for Edge Arrhythmia Classification. *Electronics* **2026**, *15*, 1044. [CrossRef]
93. Ran, Z.; Jiang, M.; Li, Y.; Wang, Z.; Wu, Y.; Ke, W.; Xia, L. Arrhythmia Classification Based on Multi-Feature Multi-Path Parallel Deep Convolutional Neural Networks and Improved Focal Loss. *Math. Biosci. Eng.* **2024**, *21*, 5521–5535. [CrossRef]
94. Risso, M.; Burrello, A.; Conti, F.; Lamberti, L.; Chen, Y.; Benini, L.; Macii, E.; Poncino, M.; Jahier Pagliari, D. Lightweight Neural Architecture Search for Temporal Convolutional Networks at the Edge. *IEEE Trans. Comput.* **2023**, *72*, 744–758. [CrossRef]
95. Ribeiro, H.D.M.; Arnold, A.; Howard, J.P.; Shun-Shin, M.J.; Zhang, Y.; Francis, D.P.; Lim, P.B.; Whinnett, Z.; Zolgharni, M. ECG-Based Real-Time Arrhythmia Monitoring Using Quantized Deep Neural Networks: A Feasibility Study. *Comput. Biol. Med.* **2022**, *143*, 105249. [CrossRef]
96. Zeng, W.; Shan, L.; Yuan, C.; Du, S. Advancing Cardiac Diagnostics: Exceptional Accuracy in Abnormal ECG Signal Classification with Cascading Deep Learning and Explainability Analysis. *Appl. Soft Comput.* **2024**, *165*, 112056. [CrossRef]
97. Samakovlis, D.; Albini, S.; Rodríguez Álvarez, R.; Constantinescu, D.A.; Schiavone, P.D.; Peón-Quirós, M.; Atienza, D. Biomed-Bench: A Benchmark Suite of TinyML Biomedical Applications for Low-Power Wearables. *IEEE Des. Test* **2024**, *42*, 45–54. [CrossRef]
98. Banbury, C.; Reddi, V.J.; Torelli, P.; Holleman, J.; Jeffries, N.; Kiraly, C.; Montino, P.; Kanter, D.; Ahmed, S.; Pau, D.; et al. Mlperf tiny benchmark *arXiv* **2021**, arXiv:2106.07597.
99. Hu, S.; Wang, Y.; Liu, J.; Yang, C. Personalized transfer learning for single-lead ecg-based sleep apnea detection: Exploring the label mapping length and transfer strategy using hybrid transformer model. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–15. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.