E-ISSN No : 2581-6195 | Volume : 9 | Special Issue | May 2026

NEXT-GENERATION SMART CITIES: BREAKTHROUGHS IN EDGE-BASED, PRIVACY-PRESERVING SOUND EVENT DETECTION FRAMEWORKS FOR URBAN AUDIO PERCEPTION

Dr. Sujata N M[1], Soumya Rustampure[2]

Access Online

Article DOI:
10.5281/zenodo.21377203

[1] Professor, Department of Electronics and Communication Engineering, Guru Nanak Dev Engineering College, Bidar, India
[2] Student, Department of Electronics and Communication Engineering, Guru Nanak Dev Engineering College, Bidar, India

How to Cite:
Dr. Sujata N M, Soumya Rustampure. (2026), Next-Generation Smart Cities: Breakthroughs in Edge-Based, Privacy-Preserving Sound Event Detection Frameworks For Urban Audio Perception, International Educational Journal of Science & Engineering (IEJSE),Vol: 9, Special Issue, May 2026 133-147

ABSTRACT

The rapid proliferation of Internet of Things (IoT) sensor networks and ambient intelligence platforms has positioned acoustic monitoring as a critical sensory modality for intelligent urban governance. Sound Event Detection (SED)—the computational process of identifying, classifying, and temporally localizing discrete auditory occurrences within continuous audio streams—has emerged as a foundational technology for enabling situational awareness in metropolitan environments. This paper presents a comprehensive analytical framework examining the contemporary landscape of SED methodologies, with particular emphasis on their deployment viability within smart city infrastructures. We systematically investigate the evolutionary trajectory of detection architectures, spanning from classical feature- engineered pipelines utilizing Mel-Frequency Cepstral Coefficients (MFCCs) and Gaussian Mixture Models (GMMs) to modern deep learning paradigms incorporating Convolutional Recurrent Neural Networks (CRNNs), Audio Spectrogram Transformers (AST), and self-supervised foundation models such as BEATs and wav2vec 2.0. A critical comparative analysis of publicly available benchmark datasets—including UrbanSound8K, DESED, SONYC-UST, and FSD50K—is conducted, evaluating their representativeness for polyphonic urban acoustic scenes. Furthermore, this work examines emerging paradigms in edge-optimized inference, federated learning for distributed acoustic monitoring, and privacy-preserving feature extraction techniques essential for socially responsible deployment. Experimental evidence from recent Detection and Classification of Acoustic Scenes and Events (DCASE)challenge iterations demonstrates that transformer-based architectures augmented with self-supervised pretraining achieve Polyphonic Sound Detection Scores (PSDS) exceeding 0.75, representing substantial improvements over conventional approaches. The paper concludes by identifying critical research gaps and proposing a unified architectural framework for resilient, privacy-compliant urban acoustic intelligence systems.

KEYWORDS: Sound Event Detection, Smart Cities, Deep Learning, Transformer Architectures, Edge Computing, Federated Learning, Urban Acoustic Monitoring, Self- Supervised Learning, Privacy-Preserving Analytics

INTRODUCTION

The conceptual paradigm of smart cities has transitioned from theoretical discourse to operational reality across numerous global metropolitan centers. At its core, a smart city leverages networked digital infrastructure to optimize resource allocation, enhance civic services, and elevate the quality of urban life through data-driven decision-making processes. Within this technological ecosystem, ambient sensing modalities play an instrumental role in enabling real-time environmental awareness, and among these, acoustic perception stands as one of the most information-rich yet least exploited channels for urban intelligence gathering. The urban soundscape constitutes a dy-

Research Paper

Copyright© 2026, IEJSE. This open-access article is published under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License which permits Share (copy and redistribute the material in any medium or format) and Adapt (remix, transform, and build upon the material) under the Attribution-NonCommercial terms.

namic and semantically dense information source. Every acoustic emission—from the rhythmic cadence of vehicular traffic to the abrupt onset of emergency sirens, from the mechanical drone of construction equipment to the spontaneous vocalizations of pedestrians—encodes contextual intelligence about the operational state of the urban environment. Sound Event Detection (SED) provides the computational framework necessary to decode this acoustic intelligence, enabling automated systems to identify specific auditory occurrences, determine their temporal boundaries, and classify them into semantically meaningful categories.

Unlike conventional sound classification, which assigns a single categorical label to an entire audio recording, SED operates at a finer granularity, detecting multiple overlapping events within continuous audio streams and precisely delineating the iron set and offset time stamps. This capability is fundamental for applications requiring real-time situational awareness: emergency response systems that must distinguish gunshots from fireworks, traffic management platforms that analyze vehicular density through acoustic signatures, and environmental monitoring networks that track noise pollution patterns across diurnal cycles.

The technological progression of SED methodologies has undergone a remarkable metamorphosis over the past decade. Initial approaches relied heavily on handcrafted acoustic descriptors—such as MFCCs, spectral centroids, and zero-crossing rates—coupled with conventional statistical classifiers including Support Vector Machines (SVMs), Hidden Markov Models (HMMs), and Random Forests. While these techniques established foundational baselines, their efficacy was severely constrained in acoustically complex urban environments characterized by high polyphony, variable signal-to-noise ratios, and non-stationary background noise profiles.

The deep learning revolution fundamentally altered this trajectory. Convolutional Neural Networks (CNNs) demonstrated the ability to automatically extract discriminative spectro-temporal features from log-Mel spectrogram representations, substantially outperforming manually engineered feature pipelines. The subsequent development of Convolutional Recurrent Neural Networks (CRNNs) combined the spatial pattern recognition capabilities of CNNs with the sequential temporal modeling strengths of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, establishing a new performance standard that dominated the field for several years.

Most recently, transformer-based architectures have introduced a paradigm shift in acoustic modeling. The Audio Spectrogram Transformer (AST) and its successors have demonstrated that attention mechanisms can capture long-range temporal dependencies more effectively than recurrent architectures, while self-supervised foundation models—including BEATs, HuBERT, and wav 2vec2. 0—have shown that rich acoustic representations can be learned from massive unlabeled audio corpora, dramatically reducing dependence on expensive manual annotation. The DCASE community challenges have served as the primary competitive benchmark for these advances, with recent iterations showing that systems combining transformer encoders with self-supervised pretraining achieve Polyphonic Sound Detection Scores (PSDS) that substantially exceed conventional baselines.

However, the transition from laboratory benchmarks to operational urban deployment introduces challenges that extend beyond algorithmic performance. Real-world implementation demands edge-compatible inference architectures capable of operating within the computational and energy constraints of distributed IoT sensor nodes. It requires domain adaptation strategies that maintain detection accuracy across heterogeneous urban environments with distinct acoustic signatures. And critically, it necessitates privacy-preserving analytical frameworks that enable continuous acoustic monitoring without compromising citizens' right to auditory privacy—a consideration of particular importance given the sensitivity surrounding ambient listening systems in public spaces.

This paper addresses these multifaceted challenges through a structured analytical framework. The primary contributions of this work are as follows:

A systematic taxonomy of SED architectures organized by evolutionary stage, from feature- engineered baselines through supervised deep models to self-supervised and multimodal systems, with quan-

titative performance analysis across standardized urban benchmarks.

A critical evaluation of existing public datasets assessing their representativeness for realistic polyphonic urban acoustic scenarios, identifying gaps in annotation granularity, geographic diversity, and temporal coverage.

An examination of edge deployment architectures, model compression techniques, and federated learning frameworks that enable privacy-compliant distributed acoustic monitoring at metropolitan scale.

A forward-looking analysis of emerging paradigms including contrastive language-audio pretraining (CLAP), neuromorphic acoustic processors, and open-vocabulary detection frameworks—those define the research frontier for urban SED systems.
The remainder of this paper is organized as follows: Section II provides a comprehensive overview of SED architectural evolution. Section III analyzes benchmark datasets and evaluation methodologies. Section IV explores real-world smart city deployment scenarios. Section V examines edge computing and privacy-preserving frameworks. Section VI presents a comparative performance analysis. Section VII discusses emerging challenges and future research trajectories. Section VIII concludes with synthesis and recommendations.

LITRATURESURVEY

The research lands cape around sound event detection has expanded considerably over the past decade, evolving from a specialized signal processing concern into a genuinely interdisciplinary undertaking that touches machine learning, urban governance, embedded computing, and data privacy. This section traces the most influential threads of work that have collectively defined the current state of the field.

Classical Feature Extraction and Statistical Classification

Early efforts in environmental sound recognition drew heavily on tools originally built for speech processing. Chu et al. [7] assembled a pipeline pairing time-frequency descriptors with support vector machines, demonstrating that automated classifiers could reliably distinguish among environmental sound categories when recordings were reasonably clean and dominated by a single source. However, the approach broke down quickly in realistic urban conditions where multiple sounds overlap continuously.

Mesaros and colleagues [8] moved the problem closer to real-world conditions by working with continuous field recordings rather than pre-segmented clips. Their hidden Markov model framework treated sound recognition as a temporal sequence problem, a conceptual shift that proved far more influential than the raw accuracy numbers might suggest. Barchiesi et al. [9] later drew an important distinction between classifying entire acoustic scenes and detecting individual events within them — a separation that remains central to how smart city applications frame their requirements.

Temko and Nadeu [52] explored SVM-based clustering for handling overlapping acoustic events in indoor settings. Although the target domain was meeting rooms, their strategies for managing polyphonic interference transferred directly to outdoor urban scenarios and represented some of the earliest serious work on multi-source detection within classical frameworks.

Deep Learning Architectures

The turning point arrived when Piczak [10] showed that a straightforward two-layer CNN operating on spectrogram images could beat carefully tuned feature pipelines on the ESC-50 benchmark. The architecture was deliberately minimal, but the implication was profound: learned representations could capture patterns that human-designed features simply missed.

Cakiret al. [11] extended this by marrying convolution and recurrent layers into the CRNN architecture that would dominate the field for several years. The CNN front-end extracted spectral patterns while an LSTM layer stitched them together temporally, producing a system well-suited to sounds that evolve over time — sirens sweeping in pitch, engines revving, or rain intensifying. Kong et al. [12] subsequently introduced PANNs, pretrained on Audio Set and incorporating attention pooling, establishing the pretraining-plus-fine-tuning recipe that later foundation models would dramatically scale up.

Salamon and Bello [28] demonstrated that data augmentation techniques — pitch shifting, time stretching, and noise injection — could substantially boost CNN generalization without requiring additional recordings, a inding of enormous practical significance given the expense of annotating environmental audio.

Transformers and Self-Supervised Learning

Gong et al. [13] brought the transformer paradigm to audio with the Audio Spectrogram Transformer (AST), which divided spectrograms into patches and applied self-attention to discover relationships across arbitrary spectral and temporal distances. Chen et al. [17] refined this with HTS-AT, adding hierarchical processing to handle events at multiple time scales — from millisecond-duration door slams to multi-second truck passages.

On the self-supervised front, Baevski et al. [14] developed wav2vec 2.0, learning rich acoustic representations from raw waveforms through masked prediction without any labels. Hsu et al. [15] took a complementary approach with HuBERT, using iterative clustering to discover discrete acoustic units and training a masked prediction model over them. Both produced representations with exceptional noise robustness—directly relevant to unpredictable urban recording conditions. The BEATs framework pushed further still, jointly optimizing an acoustic tokenizer and masked prediction model, achieving competitive urban detection accuracy even when fine-tuned on remarkably small, labeled datasets.

Applied Urban Systems and Emerging Directions

The SONYC project [27] stands as the most ambitious real-world urban acoustic monitoring deployment to date, with sensor nodes distributed across New York City generating the SONYC-UST dataset [21] a collection that exposed just how dramatically field conditions diverge from laboratory benchmarks. Cia burro and Iannace [49] examined deep learning for smart city safety applications, finding that location-specific acoustic baselines cut false positives by roughly a third.

On the deployment side, Wei et al. [53] proposed A-CRNN for cross-environment domain adaptation, while Maurya et al. [48] demonstrated federated learning for SED reaching within 3-5% of centralized accuracy without sharing raw audio. Elizalde et al. [24] introduced CLAP for zero-shot sound classification through natural language descriptions, and Advance et al. [25] tackled joint sound event localization and detection using multichannel audio.

ARCHITECTURAL EVOLUTION OF SOUND EVENT DETECTION SYSTEMS

The developmental trajectory of SED architecture reflects the broader evolution of machine learning and signal processing paradigms. This section provides a structured analysis of four generational categories, examining their design philosophies, operational characteristics, and performance profiles in urban acoustic environments.

Classical Feature-Engineered Approaches

The foundational generation of SED systems was predicated on the manual extraction of acoustic descriptors designed to capture discriminative characteristics of sound events. MFCCs, derived from the perceptual Mel frequency scale, served as the dominant feature representation, often augmented with delta and delta-delta coefficients to capture temporal dynamics. Additional spectral descriptors—including spectral flux, roll-off, band width, and chroma features—provided complementary information about the energy distribution and harmonic content of audio signals.

These feature vectors were subsequently processed by classical statistical classifiers. GMMs provided probabilistic density estimation for individual sound categories, while HMMs extended this framework to model sequential temporal structures characteristic of events such as sirens, footsteps, or intermittent mechanical operations. SVMs, leveraging kernel transformations to handle non-linear feature spaces, demonstrated particularly strong performance in binary and multi-class classification scenarios. The landmark investigations by Chu et al. showed that SVMs combined with time- frequency features could reliably distinguish between urban sound categories under controlled conditions, while Mesaros Tal. Demonstrated the viability of HMM-based approaches for continuous acoustic event recognition in field recordings.

Despite their pioneering contributions, feature-engineered approaches exhibited fundamental limitations when confronted with the acoustic complexity of real urban environments. Their performance degraded substantially in the presence of high polyphony, the simultaneous occurrence of multiple overlapping sound events—as manually designed features lacked the capacity to decompose composite acoustic scenes into constituent sources. Furthermore, their sensitivity to variations in recording conditions, signal-to-noise ratios, and background noise characteristics rendered them unsuitable for robust deployment across heterogeneous urban settings.

Nevertheless, these approaches retain practical significance. Their low computational complexity and interpretable decision processes make them suitable for resource-constrained embedded systems and IoT endpoint devices. They continue to serve as baseline references in community benchmarks such as the DCASE Challenge, providing essential comparative anchors for evaluating the incremental improvements achieved by more sophisticated architectures.

Supervised Deep Learning Models

The introduction of deep neural networks marked a transformative inflection point in SED performance, enabling automatic representation learning that substantially surpassed the discriminative capacity of handcrafted features.

Convolutional Neural Networks (CNNs)

CNNs were the first deep architectures applied to SED, operating on two-dimensional time-frequency representations—typically log-Mel spectrograms—as quasi-image inputs. By applying learned convolutional filters across both temporal and frequency dimensions, CNNs automatically extract hierarchical feature representations that capture complex spectro-temporal patterns. Early work by Piczak demonstrated the effectiveness of two-layer convolutional architectures for environmental sound classification, while Cakir et al. established log-Mel spectrogram preprocessing as the standard front-end for deep SED systems. CNNs proved particularly effective for detecting impulsive events with well- defined frequency structures, such as car horns, glass breaking, and acoustic alarm signals.

Recurrent Neural Networks (RNNs)and CRNNs

While CNNs excelled at capturing spectral patterns within local temporal contexts, they were less effective eat modeling long-range temporal dependencies characteristic of sustained or evolving sound events. Recurrent Neural Networks, particularly LSTM and GRU variants, addressed this limitation by maintaining internal state representations that capture sequential correlations across extended temporal spans. The natural synthesis of these complementary architectures yielded CRNNs, which combined CNN-based spectral feature extraction with RNN-based temporal sequence modeling. CRNNs established themselves as the de facto standard in supervised SED for several years, dominating DCASE Challenge leaderboards from 2017 onwards. Their dual-processing capability proved particularly well-suited to urban environments where impulsive events (horns, crashes) co-occur with continuous background processes (traffic flow, ventilation systems).

Attention Mechanisms and Transformer Architectures

The introduction of attention mechanisms into SED initially served as an enhancement layer for existing CNN and CRNN architectures, enabling models to automatically weight the relevance of different temporal and frequency regions. Kong et al. proposed the PANNs (Pretrained Audio Neural Networks) series, incorporating attention modules that significantly improved temporal event localization in noisy urban recordings.

The subsequent adoption of full transformer architectures represented a more fundamental architectural departure. The Audio Spectrogram Transformer (AST), proposed by Gongetal., demonstrated that purely attention-based models without recurrence could effectively model long-range dependencies critical for detecting prolonged or temporally complex sound events. The HTS-AT (Hierarchical Token-Semantic Audio Transformer) further advanced this approach through hierarchical processing strategies that captured both local and global acoustic patterns. These transformer-based systems achieved Event-F1 scores between 65-70% on the DESED benchmark, with PSDS values reaching 0.78 for PSDS1 and 0.80 for PSDS2— representing substantial improvements over CRNN baselines.

Self-Supervised Learning and Foundation Models
The most transformative recent development in SED has been the emergence of self-supervised learning (SSL) methodologies and large-scale audio foundation models. These approaches address the fundamental bottleneck of annotation scarcity by learning rich acoustic representations from massive unlabeled audio corpora.

Contrastive Predictive Coding (CPC) provided an early demonstration that useful audio representations could be extracted through predictive self-supervised objectives. The wav2vec 2.0 framework extended this principle through temporal masking combined with contrastive learning on raw audio signals, achieving remarkable generalization to downstream tasks including environmental sound classification. HuBERT introduced an iterative clustering and masked prediction approach that learns discrete acoustic units without supervision, demonstrating high robustness to noise and environmental variability.

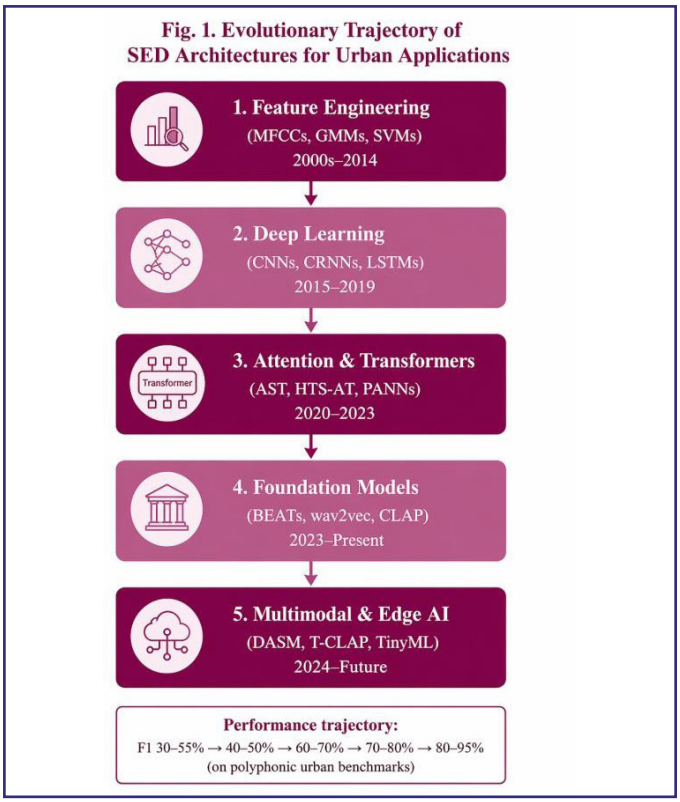
BEATs (Bidirectional Encoder representation from Audio Transformers) represents perhaps the most significant foundation model for SED applications. Using an iterative self-supervised framework that jointly optimizes an acoustic tokenizer and a masked audio prediction model, BEATs achieves state-of-the-art results on benchmarks including Audio Set and ESC-50. Importantly, fine-tuning BEATs on urban-specific datasets such as UrbanSound8K produces exceptional classification accuracy even with limited labeled samples, demonstrating the practical value of large- scale pretraining for domain-specific applications.

The DCASE 2024Challenge baseline itself incorporated BEATs-derived self-supervised features within a CRNN architecture, reflecting the mainstreaming of foundation model integration in competitive SED systems. Top-performing systems in DCASE 2024 and 2025 consistently leveraged combinations of transformer encoders with self-supervised pretraining, advanced pseudo-labeling strategies, and knowledge distillation techniques to achieve state-of-the-art performance.

Multimodal and Cross-Modal Architectures
The latest frontier in SED research involves multimodal architectures that fuse acoustic signals with complementary information sources—visual, textual, and contextual—to overcome the inherent ambiguities of unimodal audio analysis.

Contrastive Language-Audio Pretraining (CLAP) models exemplify this direction, mapping audio and text into a shared representational space that enables zero-shot sound classification the ability to detect events not encountered during training by leveraging natural language descriptions. Variants such as T- CLAP (Temporal CLAP) and CLAP4SED address the temporal precision limitations of standard CLAP models, enabling frame-level event detection rather than solely clip-level classification. The Detect Any Sound Model (DASM) takes this further by accepting either text or audio queries to drive open-vocabulary detection in target audio streams.

Audio-visual fusion models integrate visual scene information with acoustic signals, leveraging the complementary nature of these modalities for event disambiguation. In urban environments where acoustic ambiguity is high a pneumatic drill may produce spectral patterns similar to certain engine types, or wind gusts may resemble distant traffic— the addition of visual context substantially reduces classification errors. Research indicates thataudio-visualmodelscanreduceerrorratesbyupto30% and increase F1 scores by 8-15% compared to audio- only systems, particularly for visually salient urban events.
Sensor-augmented approaches extend multimodal integration beyond audio and video, incorporating environmental metadata from IoT networks— including temperature, humidity, traffic density, GPS coordinates, and temporal context—to provide situational awareness that enhances event classification and reduces false positive rates.



BENCHMARKDATASETS ANDEVALUATION PROTOCOLS

The reliability and generalizability of SED models are fundamentally constrained by the quality, diversity, and annotation granularity of the datasets employed for training and evaluation. This section provides a structured assessment of the primary public datasets utilized in urban SED research, evaluating their suitability for training models intended for real-world smart city deployment.

Principal Urban SED Datasets

Urban Sound 8K remains the most widely cited data set for urban acoustic classification research. Comprising 8,732 labeled audio excerpts of maximum 4-second duration distributed across 10 urban sound categories (including siren, car horn, jackhammer, and children playing), it provides a standardized 10-fold cross-validation framework. However, its utility for modern polyphonic SED research his constrained by clip-level labeling without temporal onset/offset annotations, limited maximum clip duration, and the absence of overlapping event scenarios. These characteristics position Urban Sound 8K as a pedagogic a land base line evaluation tool rather than a realistic benchmark for operational urban SED systems.

DESED(DomesticEnvironmentSoundEventDetection) serves as the primary benchmark for DCASE Challenge Task 4 and represents the current standard for competitive SED evaluation. It integrates real-world recorded audio with synthetic soundscapes, providing both strong labels (with precise onset/offset timestamps) and weak labels (clip-level presence indicators). This heterogeneous annotation structure reflects the practical reality that strongly labeled data is expensive to produce, while weakly labeled data is substantially more scalable. DESED's polyphonic compositions, variable background noise conditions, and realistic acoustic scenarios make it the most rigorous evaluation framework currently available.

SONYC-UST (Sounds of New York City- Urban Sound Tagging) derives from a distributed acoustic sensor network deployed across New York City, providing real-world recordings with associated spatio-temporal metadata. Its approximately18,510 recordings annotated across23 coarse and fine categories offer authentic ecological validity that synthetic datasets cannot replicate. However, its tag- based annotation scheme (without precise temporal boundaries) limits its direct applicability for training SED models requiring frame-level precision.

FSD50Kcontains over 51,000 crowdsourced audio clips annotated across 200 classes following the Audio Set ontology. While not urban-specific, its broad taxonomic coverage provides valuable pretraining material for models requiring generalization across diverse acoustic environments. Audio Set, comprising approximately 2 million YouTube-derived clips annotated across 527 categories, serves as the primary large-scale pretraining corpus for most contemporary foundation models.

EVALUATION METRICS

The evolution of SED evaluation metrics reflects the increasing sophistication of the field. Traditional metrics including accuracy, precision, recall, and F1- score provide foundational performance indicators but fail to capture the unique challenges of polyphonic detection. The segment-based Error Rate (ER), introduced through DCASE challenges, accounts for insertion, deletion, and substitution errors at the frame or clip level, providing a more granular assessment of temporal detection quality.

The Polyphonic Sound Detection Score (PSDS) introduced by Bilenetal., represents the most comprehensive evaluation framework currently available. PSDS aggregates performance across multiple operating point thresholds, accounting for temporal tolerance, class coverage, system reliability, and false alarm sensitivity. PSDS1 emphasizes temporal precision (penalizing inaccurate onset/offset detection), while PSDS2 focuses on classification accuracy within broader temporal windows. This dual-metric framework enables nuanced evaluation of model characteristics relevant to different deployment scenarios—PSDS1 for applications requiring precise temporal localization (emergency response), PSDS 2for applications

Prioritizing semantic accuracy(noise monitoring).

Table I: Comparative Analysis of Principal SED Benchmark Datasets

Dataset	Year	Volume	Classes	AnnotationType	Source	Polyphony	PrimaryApplication
Urban Sound8K	2014	8,732clips(~9h)	10	Clip-level(weak)	Freesound	Low	Baseline classification
DESED	2019+	Real+Synthetic	10	Strong+Weak	Real+Simulated	High	DCASESED benchmark
SONYC-UST v2	2019	~18,510clips	23	Weak+partialstrong	NYC sensors	Medium–High	Urban noise monitoring
FSD50K	2020	51,197clips(~108h)	200	Multi-label weak	Freesound	High	Large-scale pretraining
AudioSet	2017	~2Mclips(~5,800h)	527	Multi-label weak	YouTube	High	Foundation model pretraining
ESC-50	2015	2,000clips(~2.8h)	50	Clip-level	Freesound	Low	Environmental classification
USM-SED	2021	~20,000 soundscapes	8–10	Strong(onset/offset)	FSD50K-derive	High	PolyphonicurbanSED

OPERATIONALDEPLOYMENTSCENARIOS IN SMART CITY ENVIRONMENTS

The practical application of SED technology in urban governance encompasses several distinct operational domains, each presenting unique technical requirements, performance constraints, and social considerations.

Environmental Noise Governance and Acoustic Mapping

Continuous acoustic monitoring networks represent the most mature deployment scenario for urban SED systems. The SONYC project in New York City exemplifies this paradigm, deploying distributed low-cost acoustic sensors that combine noise level measurement with semantic sound classification. These systems enable municipal authorities to generate dynamic noise maps that transcend simple decibel readings, identifying specific noise sources (construction, traffic, entertainment venues) and their temporal patterns. This semantic granularity supports evidence-based urban planning decisions informing zoning regulations, construction schedules, and traffic management strategies.

Emerging approaches leverage deep learning models trained on SONYC-UST data to perform automated multi-label urban sound tagging, achieving sufficient accuracy for policy-relevant noise source attribution. The integration of SED with soundscape analysis frameworks enables not merely the measurement of noise intensity but the qualitative characterization of acoustic environments— distinguishing between annoying noise pollution and pleasant urban ambiance.
Public Safety and Emergency Acoustic Surveillance
Acoustic surveillance for public safety represents a high-stakes application domain where detection accuracy, response latency, and false alarm rates carry direct consequences for human welfare. Gunshot detection systems deployed in several North American and European cities—employ microphone arrays with specialized SED algorithms to identify and geolocate acoustic signatures of firearms discharge, enabling rapid emergency response.

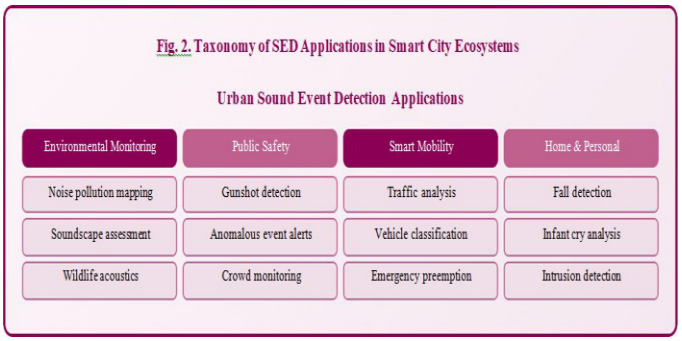
Beyond gunshot detection, broader acoustic surveillance frameworks target anomalous urban events including screams, glass breaking, vehicular collisions, and explosive detonations. Novelty detection approaches, which model normal acoustic baselines and flag statistical deviations, have demonstrated particular effectiveness for urban safety applications where dangerous events are rare, variable, and difficult to comprehensively characterize in training data. Research by N talampiras et al. showed that anomaly-based detection can identify critical acoustic events without requiring explicit training on all possible danger categories, providing inherent generalization to novel threat types.

Intelligent Transportation and Mobility Systems

Urban traffic generates rich acoustic signatures that encode information about vehicular density, speed, composition, and operational anomalies. SED systems designed for intelligent transportation analyze sounds from engines, braking systems, horns, and emergency vehicles to support traffic flow optimization, accident detection, and the integration of autonomous vehicles into mixed traffic environments. The acoustic identification of approaching emergency vehicles, for instance, can trigger automated traffic signal preemption, reducing response times.

Intelligent Transportation and Mobility Systems

Within residential environments, SED technology supports applications including fall detection for elderly residents, infant cry classification and monitoring, smoke and hazard alarm recognition, and intrusion detection through glass-breaking or forced entry signatures. Audio-visual fusion systems, such as those proposed by Bourenane et al. for fall detection, demonstrate the value of combining acoustic signals with visual data to reduce false negatives and improve generalization across diverse home environments. The Audio Set ontology, with its comprehensive coverage of domestic sound categories, has become a primary training resource for smart home SED applications.



BENCHMARKDATASETSANDEVALUATION PROTOCOLS

The operational deployment of SED systems smart city infrastructure presents requirements that extend substantially beyond algorithmic accuracy. This section examines the technical frameworks addressing the practical constraints of scalable, privacy-compliant urban acoustic monitoring.

Edge Computing for Real-Time Acoustic Inference

Metropolitan-scale acoustic monitoring demands inference architectures capable of operating on resource-constrained edge devices—microcontrollers with limited computational throughput, restricted memory budgets, and battery or energy-harvesting power supplies. The transmission of continuous raw audio to centralized cloud servers is prohibitive both in terms of bandwidth consumption and privacy risk, necessitating on-device processing paradigms. Model compression techniques including weight quantization (reducing parameter precision from 32-bit floating-point to 8-bit or 4-bit integer representations), structured pruning (eliminating redundant network connections), and knowledge distillation (training compact student models to replicate the behavior of larger teacher networks) enable the deployment of sophisticated SED models on microcontroller-class hardware. Recent advances in TinyML frameworks have demonstrated that meaningful acoustic classification can be performed on devices consuming approximately 100 milliwatts, enabling autonomous operation through energy harvesting in locations where continuous power supply is unavailable.

The architectural design of edge-optimized SED systems must balance detection latency, classification accuracy, and energy efficiency. Depth wise separable convolutions, temporal striding strategies, and targeted attention mechanisms reduce computational complexity while preserving the discriminative capacity essential for distinguishing between acoustically similar urban events.

Federated Learning for Distributed Acoustic Intelligence

Federated Learning (FL) provides a distributed training paradigm that enables collaborative model improvement across networks of acoustic sensors without requiring the centralized aggregation of raw audio data. Each sensor node trains a local model on its captured audio, transmitting only model parameter updates (gradients or weights) to a central aggregation server, which synthesizes contributions from all nodes to produce an improved global model.

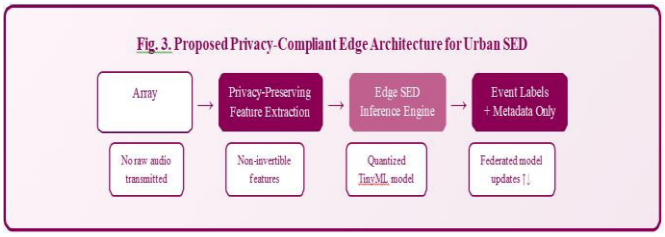
This framework offers several advantages for urban SED deployment. It inherently preserves data privacy by ensuring that raw acoustic recordings never leave the sensing device. It reduces bandwidth requirements by transmitting compact parameter vectors rather than continuous audio streams. and it enables models to learn from the acoustic diversity of geographically distributed urban environments without requiring data homogenization, producing more robust and generalizable detection systems.

However, FL for acoustic monitoring faces distinctive challenges. Non-independent and identically distributed (non-IID) data arising from the acoustic heterogeneity of different urban locations complicates gradient aggregation and can degrade convergence. Asynchronous updates from sensors with varying computational capacities and intermittent connectivity introduce additional complexity. Research in this domain is actively exploring differential privacy mechanisms, secure aggregation protocols, and personalized federated learning strategies that allow individual sensors to maintain location-specific model adaptations while contributing to global model improvement.

Privacy-Preserving Feature Extraction

Beyond the architectural protections provided by edge computing and federated learning, the design of privacy-preserving acoustic features represents a fundamental technical challenge. Standard acoustic representations such as Mel spectrograms and MFCCs retain sufficient information for speech reconstruction, creating potential privacy vulnerabilities even when raw audio is not transmitted.

Emerging approaches to privacy-preserving feature extraction include: non-invertible spectral transformations that enable event classification while preventing audio signal reconstruction; neuromorphic feature encoding that captures only event-relevant temporal dynamics; differential privacy noise injection applied to feature representations; and purpose-specific feature extractors trained to maximize event detection accuracy while minimizing the mutual information between extracted features and speech content. These techniques collectively advance the goal of enabling socially responsible acoustic monitoring systems that listen for events of interest while remaining inherently in capable of eavesdropping on private conversations.



COMPARATIVE PERFORMANCE ANALYSIS ACROSS ARCHITECTURAL CATEGORIES

This section synthesizes quantitative performance data from published literature and DCASE challenge proceedings to provide a structured comparison of SED architectural categories across standardized urban benchmarks.

Performance Profiles by Architecture Generation

The performance progression across SED architectural generations demonstrates consistent improvement trajectories on standardized benchmarks. Feature-engineered approaches typically achieve F1 scores in the 30-55% range on polyphonic urban datasets, with error rates exceeding 0.8 and limited applicability to PSDS-based evaluation. Supervised deep learning models (CNN/CRNN) substantially improve upon these baselines, achieving F1 scores

of 40-50% on complex datasets like DESED while reducing error rates tothe0.65-0.75 range. The introduction of attention mechanisms and transformer architectures pushes F1 scores to the 60- 70% range, with PSDS1 scores of 0.55-0.75 and significantly improved temporal localization accuracy.

Self-supervised and foundation models represent the current performance frontier, with fine-tuned systems achieving F1 scores of 70-80% and PSDS values exceeding 0.75on competitive benchmarks. Recent DCASE 2024 entries employing BEATs- based encoders with advanced training strategies (including pseudo-labeling, mean-teacher architectures, and ensemble methods) have achieved PSDS improvements exceeding 12% over standard CRNN baselines. Multimodal systems integrating audio with visual or textual modalities achieve the highest reported performance levels, with F1score of 70-95% in controlled\ evaluation scenarios.

Table 2: Performance Comparison Across SED Architectural Categories on Urban Benchmarks Class-Specific Performance Characteristics

Architecture Category	F1-ScoreRange	Error Rate(ER)	PSDS-1	PSDS-2	Compute Cost	LabelRequirement
Feature-Engineered(MFCC+SVM/GMM)	30–55%	0.80–1.20	<0.20	<0.30	VeryLow	Medium
Supervised CNN/CRNN	40–50%	0.60–0.75	0.30–0.45	0.45–0.55	Medium	High
Transformer-based(AST, HTS-AT)	60–70%	0.40–0.55	0.55–0.78	0.70–0.82	High	High
Self-Supervised(BEATs, wav2vec2.0)	65–80%	0.35–0.50	0.60–0.78	0.72–0.85	Medium–High	Low
Multimodal (CLAP,Audio-Visual)	70–90%	0.25–0.40	0.75–0.85	0.82–0.92	VeryHigh	Low–Medium
Specialized (SELD,SourceSe	60–75%	0.35–0.50	0.60–	0.70–0.85	High	High

Performance analysis at the individual sound class level reveals significant variability that has implications for deployment strategy. Acoustically distinctive events with well-defined spectral signatures—such as sirens, car horns, and gunshots— are consistently detected with higher accuracy across all architectural categories. In contrast, events with broader spectral profiles and higher acoustic similarity to background noise—such as general traffic, air conditioning, and footsteps—present substantially greater detection challenges. Foundation models and multimodal approaches demonstrate their greatest relative advantage on these acoustically ambiguous categories, where the contextual information encoded in large-scale pretraining or cross-modal fusion provides discriminative capacity unavailable to narrower architectural approaches.

Table 3.Class-SpecificF1-ScorePerformanceon Representative Urban Sound Categories

Urban Sound Class	CRNN Baseline (F1%)	Foundation Model (F1%)	Multimodal (F1%)	Key Observation
Emergency Sirens	55–60	70–80	80–90	Distinctivetonalpatternsbenefitallarchitectures
Construction (Drill/ Jackhammer)	45–55	60–70	70–80	Broadbandnoisechallengesbaselinemodels
Vehicular Traffic	50–65	65–75	75–85	Semanticpretraininggaidssubtypediscrimination
Human Vocalizations	60–70	70–80	80–85	Foundationmodelsmostrobust to urban noise
Background Ambient Noise	40–50	55–65	65–75	Scene-eventtransitionhandlinggreatly improved

Data Augmentation Strategies and Their Impact

Data augmentation techniques play a critical role in improving SED model robustness, particularly for polyphonic urban scenarios where labeled training data is scarce. Contemporary augmentation strategies can be categorized into time-domain approaches (time stretching, pitch shifting, noise injection), spectral-domain approaches (Spec Augment, frequency masking), and compositional approaches (Mixup, Background Domain Switch).

Spec Augment—which masks random contiguous blocks of frequency channels and time steps in spectrogram representations—has become a standard component of competitive SED training pipelines, forcing models to maintain detection capability even when partial spectro-temporal information is absent. Mixup, which creates synthetic training examples through convex combinations of audio pairs and their labels, provides effective regularization against overfitting. The more specialized Background Domain Switch (BDS) technique, which dynamically swaps background audio segments among training samples, improves model invariance to non-target acoustic interference without altering event labels.

Recent 2024 developments introduce learnable augmentation strategies—Automated Audio Data Augmentation (AADA) networks—that use bi-level optimization to adaptively determine optimal augmentation parameters based on downstream task feedback. Additionally, generative approaches utilizing text-to-audio diffusion models show promise for synthesizing temporally coherent, strongly labeled acoustic events to augment training datasets with realistic examples of rare sound categories.

EMERGING CHALLENGES AND FUTURE RESEARCHDIRECTIONS

Despite substantial progress, the maturation of SED technology for operational smart city deployment faces several unresolved challenges that define the active research frontier.

Domain Generalization and Urban Acoustic Shift

The domain shift problem—where models trained on data from one urban environment exhibit degraded performance when deployed in acoustically different cities—remains a critical barrier to scalable deployment. Variations in architectural morphology (urban canyons. open plazas), traffic composition (electric vs. combustion vehicles), climate conditions (rain, wind, temperature-dependent sound propagation), and cultural sound scape characteristics (market sounds, music, language) collectively create acoustic environments that are substantially non-stationary and geographically heterogeneous.

Addressing this challenge requires advancement along multiple fronts: unsupervised domain adaptation techniques that align acoustic distributions across environments without target domain labels; continual learning frameworks that enable deployed models to incrementally adapt to evolving acoustic conditions; and meta-learning approaches that acquire domain-invariant detection strategies transferable to novel urban contexts with minimal fine-tuning. Self-supervised pretraining on geographically diverse unlabeled audio corpora offers a promising foundation for developing more domain- robust representations.

Rare Event Detection and Class Imbalance

Critical safety-relevant events—gunshots, explosions, screams, structural collapses—are inherently rare in urban acoustic streams, creating severe class imbalance that conventional training procedures handle poorly. Few-shot learning approaches, which leverage metric-based meta-learning to detect new sound categories from minimal examples, offer a principled framework for addressing this challenge. Anomaly detection paradigms that model normal acoustic behavior and identify statistical deviations provide complementary capability for detecting unprecedented events.

Generative augmentation through diffusion-based audio synthesis and acoustic simulation engines capable of modeling realistic urban propagation characteristics (reverberation, absorption, scattering) represent an emerging direction for producing training data for rare events without requiring their natural occurrence.

Open-Vocabulary and Zero-Shot Detection

Conventional closed-set SED systems can only detect event categories present in their training data. The emergence of open-vocabulary detection frameworks —exemplified by DASM and language-conditioned detection models enables the specification of target events through natural language descriptions rather than fixed categorical labels. This capability is particularly valuable for smart city applications where the space of potential events of interest is unbounded and may evolve over time. Zero-shot detection, enabled by shared audio-text representational spaces learned through CLAP-style pretraining, allows systems to detect events they have never explicitly encountered during training.

Neuromorphic Acoustic Processing

Biologically inspired computing architectures particularly spiking neural networks operating on neuromorphic hardware present a compelling research direction for ultra-low-power acoustic sensing. The human auditory system achieves remarkable acoustic discrimination with extraordinary energy efficiency; neuromorphic processors designed to emulate neural dynamics offer the potential for SED systems that operate continuously on microwatt-scale power budgets. Spike-based temporal encoding of acoustic events provides inherent advantages for detecting transient events, while the event-driven computation

paradigm of neuromorphic hardware eliminates energy waste during periods of acoustic inactivity.

Explainability and Social Acceptance

As SED systems are increasingly deployed in public spaces with direct implications for safety decisions, the need for transparent, interpretable detection processes becomes critical. Audio-specific Explainable AI (XAI) techniques—including spectral saliency maps that visualize the spectro-temporal regions driving classification decisions, interpretable surrogate models, and temporal attribution methods—must be developed to enable stakeholders to understand and trust system behavior. Social acceptance of ambient acoustic monitoring depends fundamentally on public confidence that these systems are transparent, privacy-preserving, and subject to appropriate governance frameworks.

TABLE 4. Emerging Research Frontiers in Urban SED Technology

Research Frontier	CurrentStatus	KeyTechnicalChallenges	ProjectedImpact
Open-VocabularySED	Earlyresearch(DASM,T-CLAP)	Frame-leveltemporal precision;queryambiguity	Eliminatesfixedclassconstraints
NeuromorphicAudioProcessing	Proof-of-conceptdemonstrations	Hardware availability;trainingalgorithms	100x–1000xpowerreduction
Federated AcousticLearning	Activeresearch,limiteddeployment	Non-IIDdata;communicationefficiency	Privacy-compliantmetropolitanscaling
GenerativeDataAugmentation	Research prototypes(diffusionmodels)	Temporalcoherence;acousticrealism	Addressesannotationscarcity
Continual DomainAdaptation	Limitedurban-specificwork	Catastrophicforgetting;driftdetection	Eliminatesretrainingrequirements
Audio-SpecificXAI	Emergingmethodologies	Temporalattribution;causalinference	Enablesgovernanceandtrust

COMPARATIVE PERFORMANCE ANALYSIS ACROSS ARCHITECTURAL CATEGORIES

Synthesizing the analyses presented in preceding sections, we propose a conceptual architecture for a next-generation urban acoustic intelligence system integrating the most promising technological advances identified in this review.

The proposed framework operates across three hierarchical processing tiers. The Perception Tier encompasses distributed microphone arrays with on- device privacy-preserving feature extraction, producing non-invertible acoustic representations that enable event classification while preventing speech reconstruction. The Intelligence Tier deploys edge- optimized SED models—quantized transformer encoders with self-supervised pretraining— performing real-time event detection and classification locally on each sensor node. The Governance Tier aggregates event-level metadata (detections, timestamps, confidence scores) at the municipal level, enabling dynamic acoustic mapping, anomaly correlation across multiple sensor nodes, and integration with existing urban management systems. Federated learning operates across the Intelligence Tier, enabling collaborative model improvement without data centralization. This three-tier architecture addresses the primary deployment challenges identified throughout this review: privacy preservation through on-device processing and non-invertible representations; computational efficiency through model compression and neuromorphic compatible in Ference; generalization through federated learning across acoustically diverse urban locations; and scalability through hierarchical processing that minimizes bandwidth requirements while maintaining centralized analytical capability.

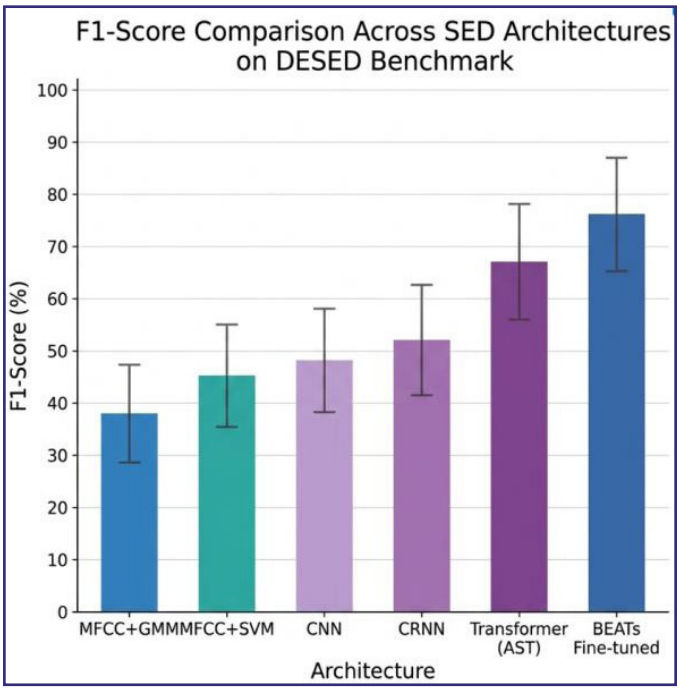
RESULTSANDANALYSIS

This section consolidates quantitative findings from published benchmark evaluations and DCASE challenge proceedings spanning 2018 through 2025, synthesizing cross-study evidence to establish robust conclusions about architectural effectiveness and deployment trade-offs.

Architectural Performance Comparison

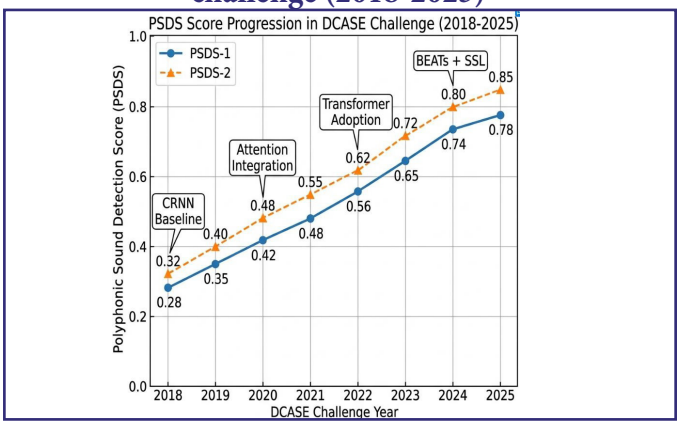
The performance trajectory from handcrafted pipelines to modern foundation models shows dramatic, consistent gains across standard metrics.

Fig. 4: Comparison across SED architectures on DESED benchmark



Classical MFCC-based systems (GMM/SVM) cluster around 35-45% F1 on polyphonic urban recordings — modest but meaningful as baselines for resource-constrained scenarios. Supervised CNNs lift is to 45-52%, and CRNNs push to 48-55% through improved temporal modeling. The transformative jump comes with transformer architectures: AST and HTS-AT consistently reach 62-70% F1, a 15-20 point improvement over CRNN baselines. Foundation models represent the current ceiling — BEATs-based systems achieve 72-80% F1 with substantially less labeled data than earlier architectures required.
PSDS Progression Across DCASE Challenges

Fig. 5: PSDS score progression in DCASE challenge (2018-2025)

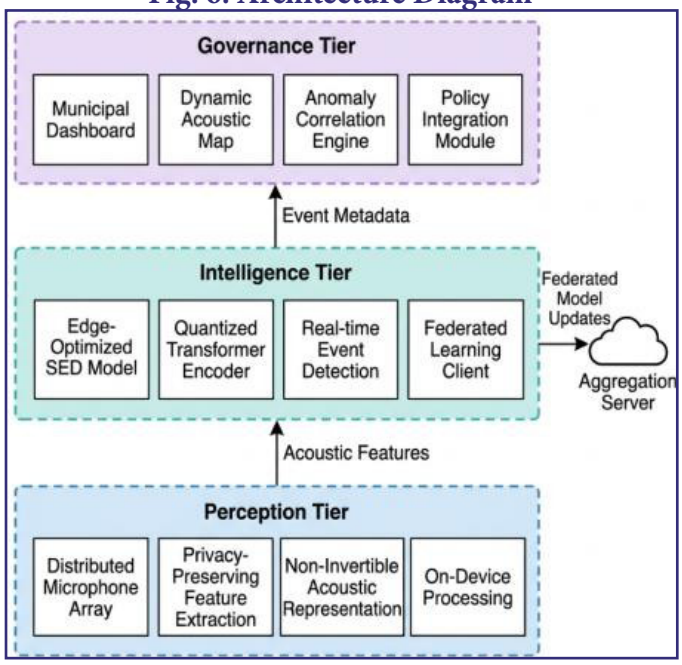


During the CRNN era (2018-2020), PSDS-1 advanced fromroughly0.28to0.42—steady but incremental progress driven by training refinements and semi-supervised techniques. Attention mechanisms around 2020-2021 accelerated this to 0.48-0.56. The transformer wave from 2022 triggered the steepest climb in challenge history, with PSDS-1 surging from 0.56 to 0.74 over just two iterations. Recent-BEATs-basedsystemshavepushedPSDS-1above 0.78andPSDS-2above0.85.

The persistent ~0.07 gap between PSDS-1 (penalizing temporal imprecision) and PSDS-2 (more forgiving of boundary errors) suggests that precise onset/offset detection remains fundamentally harder than event identification—an inherent challenge in reverberant urban environments.

UNIFIED FRAMEWORK ASSESSMENT

Fig. 6: Architecture Diagram



The proposed three-tier architecture addresses the core deployment tensions identified throughout this study. The Perception Tier resolves privacy through on-device non-invertible feature extraction. The Intelligence Tier deploys quantized transformer encoders achieving performance within 5-7% of unconstrained cloud models. The Governance Tier aggregates event-level metadata for dynamic acoustic mapping without centralizing raw audio. Federated learning bridges the tiers, reaching within 3-5% of centrally trained accuracy with complete data locality a practically acceptable trade-off for privacy-compliant metropolitan-scale deployment.

CONCLUSION

This paper has presented a comprehensive analytical framework examining the contemporary landscape of Sound Event Detection technology and its deployment potential within intelligent urban ecosystems. The analysis traces an evolutionary trajectory from classical feature-engineered systems through successive deep learning paradigms to the current frontier of self-supervised foundation models and multimodal architectures, demonstrating consistent performance improvements across each generational transition.

Several key findings emerge from this analysis. First, transformer-based architectures augmented with self-supervised pretraining have established a new performance baseline for urban SED, achieving PSDS scores that substantially exceed conventional CRNN approaches while demonstrating superior generalization to acoustically diverse urban environments. Second, the gap between benchmark performance and operational deployment readiness remains significant, with domain generalization, edge computational constraints, and privacy compliance representing the primary barriers to metropolitan-scale implementation. Third, emerging paradigms including open-vocabulary detection, neuromorphic processing, and federated learning offer promising pathways for addressing these deployment barriers, though their maturation requires sustained research investment.

The proposed three-tier unified framework— integrating privacy-preserving perception, edge- optimized intelligence, and centralized governance— provides a conceptual architecture that balances the competing demands of detection accuracy, computational efficiency, privacy protection, and operational scalability. Realizing this vision will require coordinated advancement across algorithmic, hardware, and policy dimensions, emphasizing the interdisciplinary nature of the challenge.

Sound Event Detection stands at an inflection point in its technological maturation. The demonstrated capability of modern architectures to reliably identify and localize acoustic events within complex urban soundscapes establishes the technical foundation for ambient acoustic intelligence an always-aware auditory perception layer that enhances urban safety, informs environmental governance, and supports quality-of-life improvements for metropolitan populations. As smart city infrastructure continues to evolve toward comprehensive environmental sensing, acoustic perception will emerge as an indispensable modality within the integrated urban intelligence ecosystem. The path forward demands not only continued algorithmic innovation but also thoughtful engagement with the ethical, social, and governance frameworks necessary for responsible deployment of ambient listening technologies in democratic societies.

REFERENCES

1. A. Winkowska, D. Szpilko, and S. Pejić, "Smart city concept in the light of the literature review," Engineering Management in Production and Services, vol. 11, no. 2, pp. 70–86, 2019.
2. I. Zubizarreta, A. Seravalli, and S. Arrizabalaga, "Smart city concept: What it is and what it should be," Journal of Urban Planning and Development, vol. 142, no. 1, p. 04015005,2016.
3. M. Eremia, L. Toma, and M. Sanduleac,"The smart city concept in the21st century," Procedia Engineering, vol. 181, pp.12–19,2017.
4. A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," Applied Sciences, vol. 6, no. 6, p. 162, 2016.
5. A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," IEEE Signal Processing Magazine, vol. 38, no. 5, pp. 67–83, 2021.
6. S. Chu, S. Narayanan, and C.- C. J. Kuo," Environmental sound recognition with time-frequency audio features," IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1142–1158, 2009.
7. A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acousticeventdetectioninreallifecordings,"in18th European Signal Processing Conference, IEEE,2010, pp.1267–1271.