



AI-Powered Predictive Women Safety Ecosystem with Real-Time Audio Threat Detection: A Review

Shraddha Rajesh Baisane, Prof. Bangar A.P., Prof. Bhosale S.B.

Department of Computer Engineering
Jaihind College of Engineering, Pune, India

Abstract. Women's safety is still an everyday concern in public places, workplaces and homes, and most tools in use today depend on the victim pressing a button at the worst possible moment. Artificial intelligence (AI) offers another route: listening for distress in the surroundings, combining it with other signals, and forecasting where and when the risk is highest. This paper reviews 15 IEEE-published sources from 2014 to 2024 that relate to this idea. They cover IoT-based safety devices, scream and distress detection from audio, speech emotion and fear recognition, multimodal wearable systems, crime hotspot prediction, and edge and privacy-preserving deployment. The review finds that the audio models report accuracies between 92% and 100% on the datasets they were tested on, while the one multimodal wearable study that tested fear recognition on unseen subjects reached 63.61%. Crime-prediction methods are mature at city level but are rarely linked to women-specific incidents or to live sensing. Most studies rely on small, public or acted data, say little about false alarms, response delay and battery use, and give limited attention to privacy, accents and noisy streets. The paper contributes a comparison of eleven representative works, a conceptual seven-layer architecture that joins on-device audio detection with predictive risk analytics, and a prioritised list of research gaps. The architecture is conceptual and has not been tested experimentally.

Keywords: Women safety, audio threat detection, scream detection, speech emotion recognition, predictive analytics, Internet of Things, edge computing, federated learning

I. Introduction

Feeling safe while walking home, travelling to work or simply staying at home is something many women cannot take for granted. Harassment, molestation, rape and domestic violence are the threats that researchers in this area keep returning to [1]. The usual technical answers are helpline numbers, panic-button apps and wearable alarms. These work only if the woman can reach the device and press the button, which is exactly what a frightened or restrained person may not be able to do. A systematic review of IoT-based safety devices reaches a similar conclusion: systems that raise the alert automatically, with less human interaction and better accuracy, are still needed [1].

Sound is an obvious place to look for such an automatic trigger. Attacks, struggles and calls for help are almost always noisy, and a microphone is cheap, small and already present in every smartphone. Sen et al. built a software-only prototype on this idea, in which threatening sounds around a person are detected and registered contacts are alerted without any extra hardware



[2]. Audio monitoring also needs no line of sight and can be less intrusive than camera-based monitoring [3]. Still, recognising a scream is only one piece of the problem.

The word predictive in the title is just as important. Risk is not spread evenly across places and times of the day, and crime-pattern research has produced mature ways of finding hotspots and forecasting incident counts [4]–[6]. A system that listens only after something has gone wrong is reactive. A system that also knows which streets and hours are riskier can become more sensitive there, suggest safer routes and prepare responders earlier. In the IEEE literature reviewed here, these two threads, real-time audio detection and predictive crime analytics, are developed almost independently of each other.

The main contributions are as follows. The paper brings together sources that are usually read separately, namely IoT safety devices, audio detection, emotion recognition and crime prediction. It compares them in a common table and points out where reported results cannot be compared. It also turns the gaps found into a conceptual architecture and a short research agenda. The paper is a review, so it presents no new experiments.

This paper therefore reviews the work that a predictive, audio-driven women safety ecosystem would have to build on. Its aims are to (i) summarise approaches to IoT safety devices, audio distress detection, speech emotion recognition and crime prediction, (ii) compare representative studies, (iii) identify the challenges that block real-world use, and (iv) propose a conceptual architecture and a research agenda. Section II explains how the sources were selected, Section III reviews the literature, Section IV compares the studies, Section V lists challenges and gaps, Section VI proposes directions and an architecture, and Section VII concludes.

II. REVIEW METHODOLOGY AND LITERATURE SELECTION CRITERIA

1. Search Strategy

Searches were run through a general web search engine and then followed through to IEEE Xplore records, DOAJ entries, publisher pages, DOI records and institutional repositories. The search terms combined women safety, IoT wearable, scream detection, distress detection, sound event detection, speech emotion recognition, fear recognition, crime hotspot prediction, crime forecasting, edge intelligence and federated learning. Work from 2019 to 2024 was preferred. Older papers were kept only when they introduced a dataset or method that later studies still depend on.

2. Selection Criteria

Following the seminar guidelines, only IEEE-published papers were eligible: IEEE journals and magazines, IEEE/ACM transactions, the Proceedings of the IEEE and IEEE conference proceedings. A source was included if (a) it dealt directly with women's safety or distress detection, or supplied a method, dataset or survey that such a system depends on, and (b) its bibliographic record could be confirmed through a DOI or publisher page. Preprints, theses, product pages, student project reports and venues outside IEEE were excluded, even where they looked relevant. This keeps the reference list consistent, but it also means that some recent women-safety work published elsewhere is not covered here.

3. Outcome and Limitations

Fifteen sources were kept, published between 2014 and 2024. Four are primary studies (three on audio or wearable threat detection and one on crime-zone prediction), eight are surveys or tutorials, two are datasets and one is a family of pretrained models. Searches in Scopus and Web of Science were not logged, so the review is narrative rather than systematic and no hit counts are reported. For several sources the abstract and metadata were the main material available, so statements about them are limited to what those records say. Performance figures are quoted as published and were not re-checked.



III. Literature Review

1. IoT-Based Women Safety Devices

Farooq et al. reviewed IoT devices for women's safety published between 2016 and 2022 and organised them into a taxonomy of techniques, features, wearables and sensors [1]. Pulse-rate and pressure sensors were the most common way of sensing a woman's condition, while GPS, GSM modules and Raspberry Pi boards were typically used to send the alert. Where machine learning was used, it was fairly light: logistic regression, hidden Markov models and decision trees. The authors point out that combining several kinds of sensors tends to give better accuracy and precision than relying on a single input [1]. In short, the first generation of devices is mostly about reliable location sharing and button-driven alerts, with a little intelligence added around the edges.

2. Audio-Based Distress and Scream Detection

Sen et al. developed a prototype that listens to the sounds around a person and, when a threat is detected, notifies registered contacts by e-mail, SMS and WhatsApp using only the smartphone [2]. They explored an audio dataset from Kaggle using exploratory data analysis and then trained LSTM and CNN models, reporting 96.6% accuracy in classifying audio events [2]. Varsha et al. followed a hybrid route. A sound-detection module feeds an SVM and CNN pipeline that classifies a clip at three levels before labelling it as distress or no distress, with accuracies of 93%, 100% and 92% at the three levels [3]. A perfect score at one level most likely reflects an easy sub-task, so these numbers are better read as stage-wise results than as the reliability of the whole chain.

Both studies sit on top of a wider body of work on sound event detection. Mesáros et al. explain that the task is to say what is happening and when, that the number of possible sound classes is practically unlimited, and that events are often far from the microphone and distorted by the room [7]. Two large resources help with these difficulties. AudioSet provides a hierarchical ontology and human-labelled 10-second YouTube clips [8], and the PANNs family of convolutional networks, trained on it, can be transferred to other audio tasks with strong results [9]. Neither is specific to distress, but together they offer a way to start from a pretrained model when labelled scream data is scarce.

3. Speech Emotion and Fear Recognition

A scream is not the only sign of danger, because a voice can also carry fear and stress. Khalil et al. reviewed deep learning for speech emotion recognition (SER), covering the databases used, the emotions extracted and the limitations of the field [10]. Latif et al. went further and surveyed deep representation learning for SER, including multi-task learning, domain adaptation and unsupervised learning, which matter when labelled emotional speech is limited [11].

Much of the benchmark data is acted. CREMA-D, for example, holds 7,442 clips from 91 actors expressing six emotions, fear among them, and its human raters identified the intended emotion from audio alone only 40.9% of the time [12]. If people find it hard to judge emotion from a voice without context, an automatic detector that works from voice alone should be expected to stay uncertain too.

4. Multimodal Wearable Systems

Bindi is the most complete system among the reviewed papers. It is an end-to-end wearable that tries to recognise violent situations by detecting fear-related emotions instead of waiting for a panic button. It uses auditory and physiological sensors in wearable devices, hierarchical multisensor fusion and an edge–fog–cloud IoT architecture [13]. The authors compared late-



fusion strategies for combining the speech and physiological pipelines, and the best one reached 63.61% overall fear classification accuracy in a subject-independent setting [13]. This is far below the 92–100% seen in audio-event classification, but the tasks are different. Recognising fear in people the system has never met is much harder than recognising a loud scream in a curated clip.

5. Predictive Safety Analytics

Butt et al. surveyed 49 studies on spatio-temporal crime hotspot detection and prediction. Clustering was widely used to find hotspots, time-series and deep learning methods supported trend prediction, and adding spatial and temporal information made predictions more accurate and reliable [4]. In a follow-up study they used HDBSCAN to find hotspots and SARIMA to forecast incident counts in each dense zone, evaluated on ten years of New York City crime data [6]. Mandalapu et al. examined more than 150 articles on machine learning and deep learning for crime prediction and listed the datasets, approaches and open gaps [5].

These works show that forecasting where risk concentrates is feasible. However, as described in their abstracts, none of the three focuses on women-specific incidents, and none connects its forecasts to live sensing on a person's device.

6. Edge Intelligence and Privacy

A microphone that listens all day raises two practical questions: where should the model run, and where does the audio go? Zhou et al. survey edge intelligence, which means running AI close to where data is produced instead of in a distant cloud, and they name audio surveillance among its motivating applications [14]. Running detection on the phone or wearable reduces delay and keeps working when connectivity is poor, which is the approach taken by the smartphone-only prototype of Sen et al. [2].

To improve models without collecting everyone's recordings, federated learning trains a shared model across devices while the data stays local. Li et al. identify statistical and system heterogeneity, communication cost and privacy as its main challenges [15]. None of the audio safety studies reviewed here mentions federated training in its abstract, even though recordings of a woman's surroundings are about as sensitive as personal data can be.

7. Summary of the Literature

Three patterns stand out. First, audio detection results are strong on curated data [2], [3] but weaker once the task becomes recognising fear in unseen speakers [13]. Second, predictive crime analytics is well developed at city scale [4]–[6] but disconnected from personal devices. Third, edge and privacy-preserving techniques are well described in general [14], [15] but are not yet reported in women-safety audio systems.

IV. Comparative Analysis of Existing Studies

Table I compares the reviewed works that matter most for this topic. The reported figures cannot be ranked against each other, because the tasks, datasets and evaluation methods differ, and for some entries only the abstract was available.

1. What the Studies Show

Reading across the rows shows a few things. The three detection studies [2], [3], [13] report only accuracy-type figures. None of their abstracts mentions false alarms per day, missed alarms at a fixed false-alarm rate, response delay or battery use, and these are the numbers that decide whether a user would keep such a device switched on. The only study tested on unseen subjects reports 63.61% [13], and the 40.9% human-rater figure for audio-only emotion in CREMA-D [12] hints at the reason:



the signal itself is ambiguous. On the prediction side, [4]–[6] show mature methods, but all of them work on general crime records.

2. Scope and Layering of the Studies

The studies also differ in how much of a safety system they cover. The IoT review [1] looks at hardware and alert channels. The audio studies [2], [3] work on sound alone, and only Bindi [13] fuses speech with body signals. The survey papers [4], [5], [10], [11] map their fields well but do not build anything for women's safety, and the two datasets [8], [12] were not designed for distress. Each study therefore addresses one part of the problem.

The table also shows a useful layering. Foundation resources [8], [9], [12] supply data and pretrained models, the detection studies supply task-specific pipelines, and the prediction studies supply risk context. No row combines all three layers, and this is the gap that the architecture in Section VI is meant to address.

Table 1: Comparison of Representative Studies Relevant to a Predictive Women Safety Ecosystem

Study	Focus	Method / data	Key reported result	Limitation
Farooq et al., 2023 [1]	IoT safety devices (systematic review)	Papers from 2016–2022; taxonomy of sensors, wearables and ML	Pulse-rate and pressure sensors most common; GPS, GSM and Raspberry Pi used for alerts; LR, HMM, decision trees	Calls for auto-activation and better accuracy; review only
Sen et al., 2024 [2]	Live threat detection from sound	Exploratory analysis, then LSTM and CNN on a Kaggle audio dataset; smartphone alerts	96.6% accuracy in classifying audio events	Public dataset; field and noise tests not stated in the abstract
Varsha et al., 2023 [3]	Scream-based distress detection	Hybrid SVM + CNN with three classification levels	93%, 100% and 92% at the three levels	Stage-wise scores, not end-to-end; no deployment reported
Miranda Calero et al., 2022 [13]	Fear detection in a wearable system	Speech + physiological sensing; late fusion; edge–fog–cloud	63.61% fear accuracy, subject-independent	Modest accuracy; needs several body-worn sensors
Khalil et al., 2019 [10]	Speech emotion recognition (review)	Deep learning models, databases and emotions surveyed	Maps techniques and their limitations	General SER, not safety-specific
Latif et al., 2023 [11]	Representation learning for SER (survey)	Multi-task, domain adaptation, unsupervised learning	First comprehensive survey of the topic	Not safety-specific; no deployment guidance
Cao et al., 2014 [12]	Emotional speech dataset	7,442 clips, 91 actors, six emotions	Raters identified the intended emotion from audio alone 40.9% of the time	Acted speech, not real distress
Gemmeke et al., 2017 [8]; Kong et al., 2020 [9]	Large-scale audio data and pre-trained models	AudioSet clips and ontology; CNNs trained on AudioSet and transferred to six tasks	State-of-the-art results on several audio tasks	Not distress-specific; needs adaptation



Study	Focus	Method / data	Key reported result	Limitation
Butt et al., 2020 [4]	Crime hotspot prediction (review)	49 studies; clustering, time series, deep learning	Spatial and temporal data improve prediction	General crime, not women-specific
Butt et al., 2021 [6]	City crime-zone forecasting	HDBSCAN hotspots + SARIMA per zone; NYC 2008–2017	Forecasts incident counts per dense zone	One city; no link to live sensing
Mandalapu et al., 2023 [5]	Crime prediction (review)	150+ articles on ML and DL; datasets	Summarises algorithms, datasets and gaps	General crime focus; review only

3. Comparison with the Proposed Ecosystem

Putting the reviewed studies side by side makes it easier to see where a predictive women safety system stands today. Table III compares four kinds of approach found in the literature with the conceptual ecosystem proposed in this paper. The entries for the existing approaches are taken from the sources discussed above, and the last column describes a design goal, not a tested result.

Table 2: Existing Approaches Compared with the Proposed Ecosystem

Aspect	Button-driven IoT devices [1]	Audio-only detection [2], [3]	Multimodal wearable [13]	City crime prediction [4]–[6]	Proposed ecosystem (conceptual)
How an alert starts	Mostly the user; some auto-activation sought	Detected sound	Detected fear from speech and body signals	No personal alert	Detected sound and context, with a cancel window
Signals used	Pulse rate, pressure, GPS, GSM	Audio only	Audio and physiological sensors	Historical crime records	Audio, optional motion and heart rate, location, risk map
Predictive use of risk	None reported	None reported	None reported	Hotspots and forecasts	Risk map used as a prior for alert thresholds
Where it runs	Device with GSM or Raspberry Pi	Smartphone [2]	Edge–fog–cloud	Analysis server	Edge for sensing and alerts; cloud for forecasting
Privacy measures	Not the focus	Not discussed in abstracts	Not discussed in abstracts	Not applicable	On-device processing, federated updates, consent controls
Evidence so far	Review of devices	96.6%; 93/100/92%	63.61% on unseen subjects	Mature at city level	None; needs implementation and field tests

4. Discussion

Three themes come out of this comparison. The first is the data problem. The detection studies use public or acted audio, and the larger datasets that could support them were built for general audio events [8] or acted emotion [12]. A model that does



well on such data has shown that it can recognise the sounds in that data, which is not the same as recognising a real emergency. The second theme is the metric problem. Accuracy is easy to report, but a safety device is judged by how often it cries wolf and how often it stays silent, and neither number is reported in the abstracts reviewed. The third theme is integration. Each reviewed study handles one layer of the problem well, but a woman using such a system would need all the layers to work together: sensing, deciding, alerting, learning from feedback and protecting privacy.

It is also worth noting how the results relate to each other. The high audio scores [2], [3] and the low fear-recognition score [13] do not contradict one another, because they measure different things. Classifying a distinct sound class in a prepared dataset is a narrower task than judging a person's emotional state from speech and body signals in people the system was not trained on. A fair reading is that sound-based triggers such as screams are a realistic starting point, whereas emotion-based triggers need more data and stronger evaluation before they can be trusted on their own. This is our interpretation of the published figures and not a conclusion drawn by the original authors.

V. Research Challenges and Identified Gaps

The studies above show real progress, but moving from a laboratory result to a system a woman can trust on a late-night street needs much more. The main obstacles are the following.

- Scarce realistic data: The audio studies work with a public Kaggle dataset [2] or with data that the abstract does not describe in detail [3]. The large AudioSet is not centred on distress [8], and the best-known emotional speech corpus is acted [12]. Genuine recordings of assaults are, understandably, very hard to collect and ethically sensitive. Without them it is unclear how well any model recognises real fear, as opposed to a staged scream.
- Accuracy that may not carry over: High scores on curated clips sit uneasily beside the 63.61% obtained on unseen subjects [13] and the 40.9% human recognition rate for audio-only emotion [12]. Real streets add distance, reverberation and overlapping sounds, which sound event detection research lists as core difficulties [7].
- False and missed alarms go unreported: A scream at a football match, a child at play or a loud film can all trigger a false alarm. If there are too many, users will switch the system off. If the detector is too cautious, it stays silent when it matters. Because these two errors cost so differently, recall at a fixed false-alarm rate says more than overall accuracy, yet the reviewed detection studies report only accuracy-style figures [2], [3], [13].
- Latency, power and on-device limits: Wearables have small batteries. Edge deployment reduces delay and cloud dependence [14], but none of the reviewed primary studies reports energy use or end-to-end alert delay in its abstract. The computational cost of audio networks also varies widely [9], so the choice of model matters for a device worn all day.
- Privacy, consent and misuse: Continuous listening also records bystanders. Federated learning [15] and on-device processing [14] are promising answers, but the audio safety studies reviewed do not discuss them in their abstracts. There is also a risk that a protective tool could be turned into a way of monitoring women, which calls for user-controlled consent and clear rules on how long data is kept.
- Language, accent and speaker variation: Emotional speech corpora differ in speakers, languages and recording conditions, and models trained on one often fall short on another. This is why domain adaptation features so strongly in SER surveys [10], [11]. A system for a multilingual country has to cope with many languages, dialects and ways of calling for help.
- Prediction is disconnected from sensing, and its data can be biased: City-scale forecasting [4]–[6] and personal audio detection [2], [3], [13] have not been joined. Crime records also contain only what was reported, and harassment of women is widely under-reported, so hotspot maps may miss exactly the incidents of interest. This point is our own reasoning and not a finding of the reviewed papers.
- No field validation or response chain: Farooq et al. call for systems with auto-activation, less human interaction and better accuracy [1], yet none of the reviewed audio systems describes a field trial with real users. Technology alone does not

help if nobody responds. How alerts reach trusted contacts, volunteers or the police, and how quickly they act, remain open design and policy questions.

These gaps are linked to one another. A shortage of realistic data makes it hard to judge accuracy outside the lab, and without realistic tests the false-alarm and delay figures that matter most cannot be measured. Privacy choices, such as keeping audio on the device, in turn limit how much data can be collected for training, so data collection and privacy design have to be planned together.

VI. Proposed Research Directions

1. Proposed Conceptual Architecture

Fig. 1 shows a conceptual seven-layer architecture drawn from the gaps above. It has not been implemented or tested.

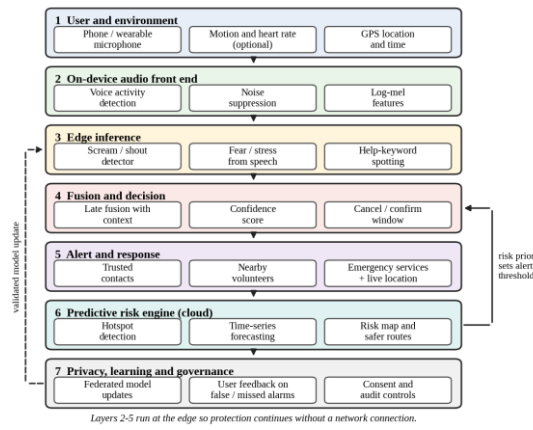


Fig. 1. Proposed conceptual architecture for a predictive women safety ecosystem with real-time audio threat detection (conceptual, not experimentally validated).

Layer 1 collects signals from the person's phone or wearable: microphone audio, optionally motion and heart-rate readings, and location and time, in line with the multi-sensor advice in [1] and [13].

Layer 2 runs on the device and prepares the audio with voice activity detection, noise suppression and log-mel features, so that raw recordings do not have to be stored or uploaded.

Layer 3 performs edge inference with three light models: a scream and shout detector, a fear or stress recogniser for speech, and a spotter for help-related keywords. These can start from pretrained audio networks [8], [9] and be adapted with distress-specific data [2], [3], [13].

Layer 4 fuses the model outputs with motion, physiological and location context using late fusion, as explored in [13], and produces a confidence score. A short cancel window lets the user stop a false alarm before anyone is notified.

Layer 5 sends tiered alerts to trusted contacts, nearby volunteers and emergency services together with the live location.



Layer 6 is the predictive part. In the cloud, hotspot detection and time-series forecasting [4], [6] produce a risk map by place and hour. The map feeds back to Layer 4 as a prior, so thresholds can be a little more sensitive where and when risk is higher, and it also supports safer-route suggestions. Layer 7 covers learning and governance: federated model updates [15], user feedback on false and missed alarms, and consent and audit controls. Layers 2 to 5 run at the edge so that protection continues when the network drops [14].

2. Priority Directions and Evaluation Practice

Table II maps each gap from Section V to a direction that we think deserves priority. Evaluation practice also needs attention. Data should be split by speaker and by recording session so that the same person never appears in both the training and test sets. Papers should report recall at a fixed false-alarm rate per day, the end-to-end time from sound to alert, and battery drain on a real device. Tests should cover several languages, accents and noisy places, and publications should explain how consent was obtained for any recorded audio.

Table 3: Research Gaps and Proposed Directions

Gap	Proposed direction	Basis
Scarce realistic data	Consented, privacy-safe collection of distress-like audio from role-play and simulated scenes in several languages; transfer from pretrained models	[2], [3], [8], [9], [12]
Accuracy outside the lab	Speaker-independent testing in noisy streets, homes and transport; robustness to distance and reverberation	[7], [12], [13]
False and missed alarms unreported	Report recall at a fixed false-alarm rate, alert delay and cancel-window outcomes	[2], [3], [13]
Latency and power	Compact edge models; energy and delay profiling on real wearables and phones	[9], [14]
Privacy and misuse	On-device processing, federated learning and user-controlled data retention	[14], [15]
Language and accent variation	Cross-corpus and domain-adaptation methods; multilingual evaluation	[10], [11]
Prediction disconnected from sensing	Use hotspot and forecast output as a prior for detector thresholds; study women-specific incident data with attention to reporting bias	[4]–[6]
No field validation	Pilot studies with real users and response partners; shared evaluation protocols	[1]

The gaps are not equally urgent. Our view is that three of them come first. Realistic, consented data and honest evaluation (false and missed alarms, speaker-independent tests) are needed before any accuracy claim can be trusted, and privacy design should be settled before a device that listens all day is tested with real users. Predictive integration and large field trials can follow once a reliable edge detector exists. This ordering is a suggestion based on the review and has not been tested.

A staged plan would also keep the work manageable for a small team. A first stage could build the on-device audio detector from a pretrained network [9] and test it on consented role-play recordings. A second stage could add the fusion and cancel-window logic and measure delay and battery use on a real phone. A third stage could connect the hotspot and forecasting layer



[4], [6] and run a small pilot with volunteers and trusted contacts. Each stage would produce numbers that the reviewed papers do not yet report.

3. Ethical and Practical Considerations

A system that listens to a woman's surroundings needs clear rules about consent and use. Audio can capture other people, so a design should record as little as possible, keep raw audio on the device where it can, and let the user switch monitoring off at any time. The cancel window in Layer 4 is one way to give the user control. Retention periods and audit logs in Layer 7 are another. These points come from our own reasoning about the risks raised in the reviewed literature [14], [15], and they would need to be checked with legal and ethical guidance.

The response side is equally important. An alert that reaches a trusted contact who cannot act is of little use. Any real deployment would need agreement with the people and services who receive the alerts, and a clear plan for what happens in the first minutes after an alert. The reviewed technical papers do not cover this part, and it is one reason why field trials with real users are listed as a priority in Table II.

4. Limitations of This Review

This review has several limits that readers should keep in mind. Only IEEE-published sources were used, so relevant work from other publishers is missing. The search was narrative and not a logged systematic search. For several sources the abstract and metadata were the main material available, so the comparison rests on what those records state. The performance figures were taken as published and were not re-checked, and they come from different tasks and datasets. The architecture in Fig. 1 is a proposal based on the reviewed literature and has not been built or tested.

VII. Conclusion and Future Scope

This review looked at 15 IEEE-published sources that together cover IoT safety devices, audio distress detection, speech emotion recognition, crime prediction and edge deployment. Audio models can classify scream and distress sounds with reported accuracies of 92–100% on the datasets used, but recognising fear in unseen speakers is much harder, and the one multimodal wearable that tested this reported 63.61% [13]. Crime prediction methods are mature at city scale, yet they remain disconnected from personal sensing. The evidence is limited by small and public datasets, accuracy-only reporting, and little attention to privacy, language variation and field validation. The seven-layer architecture proposed here is a starting point and not a validated design.

The natural next steps are to build and test the edge layers of the architecture on a low-power wearable or smartphone, to collect a consented multilingual distress-audio dataset, to measure false-alarm and delay budgets with volunteers, and to link the predictive layer to women-specific, bias-aware incident data. A further step is to widen this review beyond IEEE venues with a systematic search protocol, and to study the social and legal side of the problem: who responds, how quickly, and how users' data is protected.

References

1. M. S. Farooq, A. Masooma, U. Omer, R. Tehseen, S. A. M. Gilani, and Z. Atal, "The role of IoT in woman's safety: A systematic literature review," *IEEE Access*, vol. 11, pp. 69807–69825, 2023, doi: 10.1109/ACCESS.2023.3252903.
2. A. Sen et al., "Live event detection for people's safety using NLP and deep learning," *IEEE Access*, vol. 12, pp. 6455–6472, 2024, doi: 10.1109/ACCESS.2023.3349097.



3. M. Varsha, Y. R. Aithal, S. Fathima, and S. Sen, "Distress detection using a hybrid SVM–CNN classifier," in Proc. Int. Conf. Comput. Intell. Inf. Security Commun. Appl. (CIISCA), Bengaluru, India, Jun. 2023, pp. 224–229, doi: 10.1109/CIISCA59740.2023.00051.
4. U. M. Butt et al., "Spatio-temporal crime hotspot detection and prediction: A systematic literature review," IEEE Access, vol. 8, pp. 166553–166574, 2020, doi: 10.1109/ACCESS.2020.3022808.
5. V. Mandalapu, L. Elluri, P. Vyas, and N. Roy, "Crime prediction using machine learning and deep learning: A systematic review and future directions," IEEE Access, vol. 11, pp. 60153–60170, 2023, doi: 10.1109/ACCESS.2023.3286344.
6. U. M. Butt et al., "Spatio-temporal crime predictions by leveraging artificial intelligence for citizens security in smart cities," IEEE Access, vol. 9, pp. 47516–47529, 2021, doi: 10.1109/ACCESS.2021.3068306.
7. A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," IEEE Signal Process. Mag., vol. 38, no. 5, pp. 67–83, Sep. 2021, doi: 10.1109/MSP.2021.3090678.
8. J. F. Gemmeke et al., "Audio Set: An ontology and human-labeled dataset for audio events," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), New Orleans, LA, USA, Mar. 2017, pp. 776–780, doi: 10.1109/ICASSP.2017.7952261.
9. Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," IEEE/ACM Trans. Audio, Speech, Language Process., vol. 28, pp. 2880–2894, 2020, doi: 10.1109/TASLP.2020.3030497.
10. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," IEEE Access, vol. 7, pp. 117327–117345, 2019, doi: 10.1109/ACCESS.2019.2936124.
11. S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Survey of deep representation learning for speech emotion recognition," IEEE Trans. Affect. Comput., vol. 14, no. 2, pp. 1634–1654, Apr.–Jun. 2023, doi: 10.1109/TAFFC.2021.3114365.
12. H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," IEEE Trans. Affect. Comput., vol. 5, no. 4, pp. 377–390, Oct.–Dec. 2014, doi: 10.1109/TAFFC.2014.2336244.
13. J. A. Miranda Calero et al., "Bindi: Affective Internet of Things to combat gender-based violence," IEEE Internet Things J., vol. 9, no. 21, pp. 21174–21193, Nov. 2022, doi: 10.1109/JIOT.2022.3177256.
14. Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," Proc. IEEE, vol. 107, no. 8, pp. 1738–1762, Aug. 2019, doi: 10.1109/JPROC.2019.2918951.
15. T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Process. Mag., vol. 37, no. 3, pp. 50–60, May 2020, doi: 10.1109/MSP.2020.2975749.