2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

# Towards Distributed Robotic Casualty Assessment using Multimodal, Non-Contact Perception and Probabilistic Inference

Zachary Bortoff[1], Srijal Shekhar Poojari[2], Kleio Baxevani[1], Joshua Gaus[3], Christopher Titus[3], Ahmed Ashry[1], and Derek A. Paley[4]

***Abstract*— Mass-casualty incidents demand rapid and accurate triage, but the scale and acuity of injuries often overwhelm available medical personnel. To address this, we present a system that enables ground and aerial robots to localize and assess casualties using non-contact sensors, including color and thermal cameras, millimeter wave radar, and microphones. Injury and vital sign measurements from modality-specific classifiers are fused using a probabilistic model that captures correlations between injury states and supports distributed, asynchronous evidence accumulation. We validate the system through a series of timed mass-casualty field experiments using custom-built drones and Boston Dynamics Spot ground robots customized for robotic medical triage, demonstrating reliable estimation of casualty states and robustness to noisy conditions and sensor drop out.**

## I. INTRODUCTION

Prioritizing casualties based on injury severity is essential to saving lives during mass casualty incidents (MCIs), where the number and acuteness of casualties leave medical personnel overwhelmed [1]. The potential to scale triage operations by providing medics with a casualty's location and injury severity is made possible due to advances in non-contact vitals sensors, image and audio classifiers, and ground and aerial robots.

However, deploying robotic platforms in real-world triage settings introduces challenges. Injuries may manifest across different sensing modalities, necessitating the integration and fusion of multiple sensor streams. Environmental noise such as smoke obscuring thermal signatures, poor lighting degrading color camera imagery, and loud ambient sounds masking casualty vocalizations significantly reduces sensor data quality. At the same time, failing to estimate injuries accurately risks not prioritizing life-saving care appropriately.

Addressing these challenges requires principled models that integrate heterogeneous data sources and provide calibrated estimates of injury likelihood and triage urgency. The


Fig. 1: (Top) Three customized Boston Dynamics Spots and one custom drone performing medical triage of casualties in simulated mass-casualty incident. (Bottom) Spot and drone injury assessments of same casualty.

System Competition Track of the DARPA Triage Challenge (DTC) aims to promote the development of robotic platforms that can assist emergency medical services in quickly and accurately performing triage of casualties during MCIs [2]. This paper presents a system for assessing casualties developed by our team, RoboScout DTC, for participation in the second year of the three-year Challenge.

Recent work has demonstrated the feasibility of deploying deep learning models on aerial and ground robots for contact-free injury classification and vital sign estimation using color and infrared cameras, microphones, and millimeter-wave (mmWave) radar. Carrion et al. proposes an object detection model for detecting wounds in color imagery for use in medical treatment facility (MTF) settings [3]. West et al. shows that simulated blood is visible from a thermal image taken by a drone, but do not propose a method of automatically detecting blood [4]. Saeed et al. compares two machine-learning models for human scream detection deployed on a mobile ground robot for use in fire emergencies [5]. Zhao et al. and Chen et al. each propose algorithms for using a mmWave radar sensor deployed on a robot to estimate heart rate [6], [7]. While these approaches enable medical assessments by robotic platforms, the challenges

*This work was supported by Defense Advanced Research Projects Agency (DARPA) Grant No. HR00112420304. Approved for public release; distribution is unlimited.

[1]Zachary Bortoff, Kleio Baxevani, and Ahmed Ashry are with the Department of Aerospace Engineering, University of Maryland, College Park, MD, USA. {zbortoff, kleio, aashry}@umd.edu

[2]Srijal Poojari is with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA. srijal@umd.edu

[3]Joshua Gaus and Christopher Titus are with University of Maryland UAS Research and Operations Center, California, MD, USA. {jgaus, ctitus}@umd.edu

[4]Derek A. Paley is with the Department of Aerospace Engineering and the Institute for Systems Research, University of Maryland, College Park, MD, USA. dpaley@umd.edu

posed by realistic MCI conditions have been largely ignored and each operates on an individual sensor stream, precluding comprehensive assessment of the casualty.

Multimodal sensor fusion has been explored to estimate a patient's condition using wearable medical senors or within controlled environments, like MTFs. Shi et al. and Krygier et al. propose fusion frameworks for estimating physiological signatures of soldiers using a network of wearable sensors [8], [9]. Warnecke et al. uses a piezoelectric sensor along with two accelerometers attached to a seatbelt along with an in-vehicle camera to estimate respiratory rate within a car [10]. Yang et al. fuses multiple indoor radars for activity and vitals monitoring [11]. Choi et al. applies speech recognition and face movement models to video and audio data to assess degree of consciousness in patients within an MTF [12]. While these systems seek to fuse multiple sensor streams to provide more holistic assessment of a patient's injuries, they are not deployable in MCIs, where the environment is not controlled and casualties may not be wearing medical sensors.

A growing body of work in probabilistic modeling offers Bayesian techniques for fusing the outputs of multiple classifiers. Kim et al. models classifier outputs as mutually independent and multinomially distributed over each row of the classifiers' confusion matrices [13]. Nazabal et al. introduces a model for fusing mutually independent probabilistic classifiers that output categorical distributions instead of only discrete class labels [14]. Pirs et al. and Trick et al. explicitly model correlations between probabilistic classifiers by defining complex hierarchical graphical models [15], [16]. These late-stage fusion methods offer two practical benefits: (1) they support asynchronous and distributed evidence accumulation, so classifier outputs collected independently by a UAV, UGV, or leave-behind sensor can be incrementally fused via Bayesian update, and (2) they are extensible, since new sensor modalities can be integrated by simply defining another classifier and an additional observation model without modifying the other model parameters. The second property is particularly critical in MCI settings, where no unified data sets exist containing all relevant modalities, but individual classifiers can still be trained on modality-specific data sets.

We train modality-specific classifiers to detect injuries and estimate vitals. Then, we fuse the classifier outputs using a dynamic Bayesian network model to infer the most likely distribution over eleven casualty health states deemed relevant for prioritizing medical care. This two-step fusion pipeline is evaluated for overall accuracy and reliability in several timed, mock mass-casualty field experiments in which three customized Boston Dynamics Spot robots and two custom drones assess the injuries of dozens of human actors and trauma manikins. The contributions of this paper are as follows:

1) Modality-specific classifiers for injury detection and vitals estimation using the following non-contact sensors: color and thermal cameras, mmWave radar, and a microphone;
2) A principled method of fusing the classifier outputs using Bayesian inference; and
3) Experimental validation of our system in several mock mass-casualty incidents.

The rest of the paper is organized as follows. Section II provides a brief description of the DARPA Triage Challenge. Section III presents the modality-specific classifier training and a probabilistic fusion model. Section IV describes the performance of the proposed approach in field experiments. Section V concludes with final remarks and future work.

## II. DARPA Triage Challenge

The DARPA Triage Challenge Event 2 requires estimating a set of injury states outlined in the Massive Hemorrhage, Airway, Respiration, Circulation, and Hypothermia (MARCH) protocol for each casualty in an MCI. Table I summarizes the full set of estimated variables and their respective state spaces.

TABLE I: Injury States and State Spaces

| **Injury Variable** | **State Space** |
|---|---|
| Severe Hemorrhage | {0=absent, 1=present} |
| Respiratory Distress | {0=absent, 1=present} |
| heart rate | $\mathbb{R}^+$ |
| respiratory rate | $\mathbb{R}^+$ |
| head trauma | {0=normal, 1=wound} |
| torso trauma | {0=normal, 1=wound} |
| upper extremity trauma | {0=normal, 1=wound, 2=amputation} |
| lower extremity trauma | {0=normal, 1=wound, 2=amputation} |
| motor alertness | {0=normal, 1=abnormal, 2=absent} |
| ocular alertness | {0=open, 1=closed, 2=not testable} |
| verbal alertness | {0=normal, 1=abnormal, 2=absent} |

## III. CLASSIFIER TRAINING AND BAYESIAN INFERENCE FOR CASUALTY ASSESSMENT

This section presents the modality specific classifiers that interpret raw sensor data and a probabilistic model that fuses classifier outputs to infer latent injury states.

### A. Modality-Specific Injury Data Sets & Classifiers

Fig. 2: Sample aerial and ground images in custom color and thermal data sets.

*1) Color & Thermal Cameras:* In color images, injuries can manifest as blood-soaked or torn clothing, hands pressed against wounds, pooling blood, etc.. Manifestations of this kind motivate our approach to detect severe hemorrhage, respiratory distress, head, torso, upper and lower extremity traumas, and ocular alertness by training dedicated color image classifiers.

Color image classifier development was supported by the collection and annotation of 1693 color images, yielding 2672 color labeled instances of casualties in diverse poses,

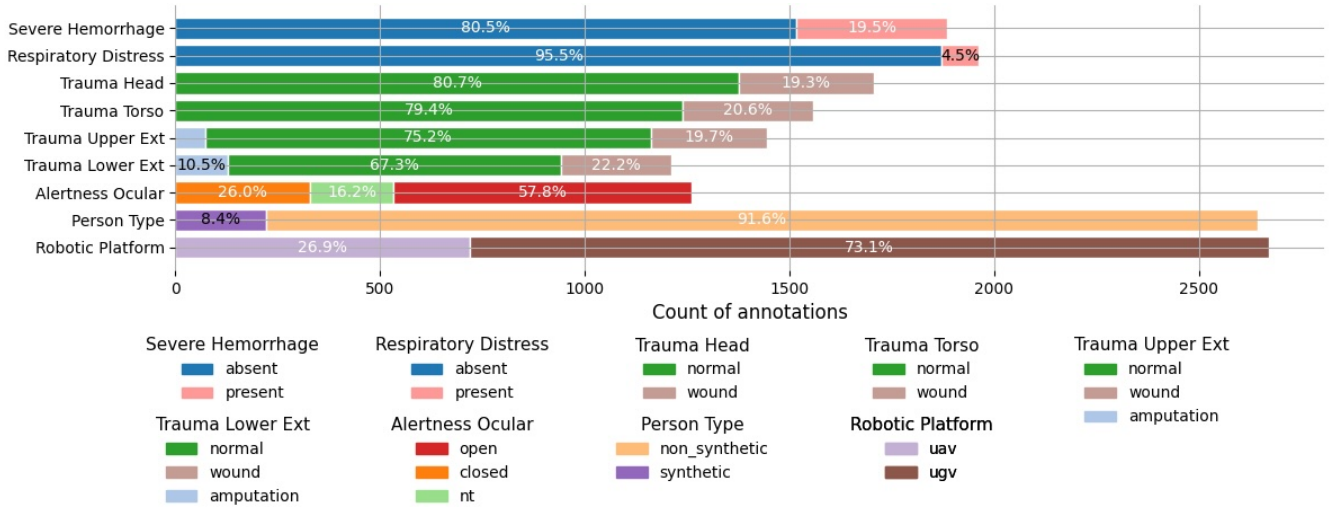


Fig. 3: Label distribution by injury class for custom color data set.

environments, and injuries. The authors labeled the data following guidelines provided by DARPA. Figure 2 depicts representative images from color and thermal cameras from both aerial and ground perspectives.

Figure 3 presents the distributions of labels by class for the custom color image data set. These distributions show that certain classes, particularly respiratory distress and upper and lower extremity trauma, exhibit significant class imbalances. To mitigate the effects of these imbalances during training, we employed a focal loss function with large focusing parameters to emphasize minority-class samples. Additionally, to further expand coverage of underrepresented class labels, we generated synthetic injury samples using Stable Diffusion [17]. Figure 4 compares a real color image and a synthetically generated version produced by Stable Diffusion guided by text prompts describing specific injury patterns.

To quantify the impact of synthetic data, we performed an ablation to compare the best performing fine-tuned model, CLIP@336px, trained with and without the synthetic injury samples, evaluating balanced accuracy averaged across all injury patterns; balanced accuracy improved from 71% to 77%. These results suggest that high-quality synthetic training data can induce marginal classifier performance gains for minimal relative effort compared to manual data collection and annotation. We expect that training on more high-quality synthetic samples would improve the model even more, but with diminishing marginal benefit.

Table II compares the balanced accuracy, weighted F1 score, and expected calibration error (ECE) averaged across all injury classes of two model types for injury classification in color images: fine-tuned supervised classifiers and zero-shot vision-language models (VLMs). Balanced accuracy, which is computed as the average of the diagonal components of the normalized confusion matrix, was chosen because the distribution of labels in the test data set is not uniform. Weighted F1 score provides a representative summary of performance, because high scores are only achieved when both precision and recall are high. Similarly, ECE was chosen because it is expected that well calibrated classifiers will perform better than poorly calibrated classifiers when using the classifier outputs in a probabilistic fusion scheme, as we do in this paper. CLIP@336px and InternVL3-8b both achieved highest balanced accuracy averaged across all injury classes while meeting constraints on model size and inference latency for deployment on our hardware. Incorporating a zero-shot model like InternVL3 enhances robustness by mitigating potential biases introduced during fine-tuning of CLIP.

Fig. 4: Comparison of original color image and synthetic image augmented with head, torso, & upper extremity trauma.

TABLE II: Color Image Classifier Balanced Accuracy and ECE ($n = 257$)

| Model | Fine-Tuned | Bal. Acc. | Weighted F1 | ECE |
|---|---|---|---|---|
| ViT | ✓ | 0.648 | 0.708 | 0.097 |
| CLIP | ✓ | 0.713 | 0.745 | 0.094 |
| **CLIP@336px** | ✓ | **0.771** | 0.776 | 0.102 |
| InternVL3-8b | ✗ | 0.833 | 0.844 | — |
| QWEN2.5-7b | ✗ | 0.823 | 0.832 | — |

The injury state most salient in color and thermal video, as opposed to still images, is motor alertness, which can manifest as walking, standing, sitting unsupported (normal), twitching, sitting supported (abnormal), and lying with no

limb movement (absent). Manifestations of this kind motivate our approach of using the You Only Look Once (YOLO) pose estimator, operating on thermal and color image streams, to extract keypoints of a casualty's joints [18]. To assess motor alertness in a fixed-length video clip using keypoints, we compute a scalar statistics, "average pose displacement". Since the keypoints are extracted in pixel coordinates, this scalar statistic is likewise measured in pixel units. Assuming all YOLO-derived keypoint pixel coordinates are corrupted by additive, zero-mean Gaussian noise, we find that the average pose displacement approximately follows a Gaussian distributions conditioned on the true class label. Using a corpus of fixed-duration annotated video segments labeled as 'normal', 'abnormal', and 'absent', we estimated the parameters of the conditional distributions. To prevent camera movement from artificially inflating this statistic, we only take data when the robot is stationary.

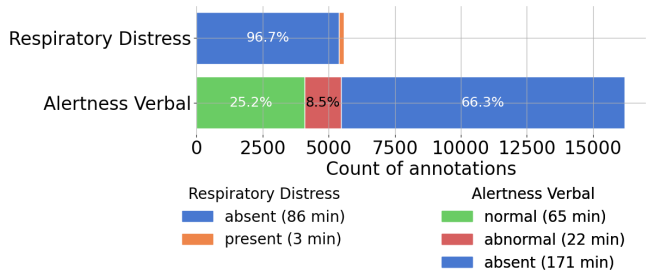


Fig. 5: Label distribution by injury class for custom audio data set.

*2) Audio:* In audio, injuries manifest through vocal and respiratory cues, including speaking, coughing, gasping, wheezing, or complete silence. Manifestations of this kind motivate our approach to detect respiratory distress and verbal alertness by using pre-trained Audio Spectrogram Transformer (AST), an audio classifier [19]. Although we currently deploy the AST without fine-tuning, we plan to fine-tune the model on this domain to improve future performance once sufficient data is collected. The data set used to validate this model includes Google's AudioSet, which contains a wide range of verbal and non-verbal vocalizations relevant to injury assessment [20]. Figure 5 depicts the distributions of labels by class of 65848 audio samples between 0.96 and 9.6 seconds long. Table III shows the balanced accuracy, weighted F1 score, and ECE of AST for each injury class on this data set.

TABLE III: Audio Classifier Performance ($n = 65848$)

| Injury Class | Bal. Acc. | Weighted F1 | ECE |
|---|---|---|---|
| Respiratory Distress | 0.579 | 0.983 | 0.012 |
| Alertness Verbal | 0.522 | 0.756 | 0.168 |

*3) mmWave Radar:* We use IWR6843 frequency modulated continuous wave (FMCW) mmWave radar from Texas Instruments to estimate heart and respiratory rates up to 5 meters from the casualty. The sensor directly measures vitals signs onboard by extracting micro-Doppler phase shifts from the reflected radar chirps to measure the millimeter-scale chest displacements of respiration and cardiac activity, and then provides these vitals measurements to a companion computer [21]. We assume these radar-derived vitals measurements follow zero-mean Gaussian distributions conditioned on the true physiological rate. The results of computing the maximum likelihood estimates (MLE) of the covariance of these conditional distributions on a training data set may be found in a Table IV.

TABLE IV: mmWave radar Vitals Covariance

| Injury State | Num. Seconds | Meas. Cov. (95% CI) |
|---|---|---|
| Heart Rate | 540 | 8.07 (6.94, 9.63) |
| Respiratory Rate | 540 | 6.76 (5.82, 8.07) |

### B. Probabilistic Fusion Model

This section introduces a probabilistic framework for doing late-stage fusion of multimodal, contact-free sensors.

*1) Dynamic Bayesian Network:* We model the temporal evolution of casualty injury states using a dynamic Bayesian network (DBN) learned from a data set of casualty injuries. At each time slice, the DBN represents the joint distribution over the latent injury states, with each node corresponding to a latent variable defined in Table I. To capture dependencies among the injury variables within a slice, we learn the structure of the Bayesian network to maximize the Bayesian Dirichlet equivalent uniform score, which balances model fit with structural complexity [22]. Once the slice-level structure is fixed, we estimate the model parameters using maximum likelihood estimation. Each injury variable at time $t$ depends not only on its parents within the same slice, but also on its own state at the previous time step $t-1$, enabling the model to capture persistence of injuries. Figure 6 illustrates the learned slice-level structure of the network.

To demonstrate the importance of modeling correlations between injury states, we compare the structured network to a baseline model in which all latent injury variables are assumed mutually independent, with empirical prior distributions estimated from the same dataset. The structured model achieves an average likelihood approximately 3.45 times higher than the independent model per casualty on a held-out validation data set, demonstrating that injury patterns exhibit significant statistical dependencies which are captured by the learned network. For clarity, we refer to the two models as the Structured Bayesian Network (SBN) and the Independent Latent Network (ILN), respectively.

We computed prior distributions over heart and respiratory rate by fitting Gamma distributions to the same data set of casualty injuries. To validate these priors, we compared the estimated Gamma distributions to the empirical distributions of the same variables in a held-out validation dataset. This validation was performed graphically by overlaying the fitted distributions and empirical histograms, as shown in Figure 7, allowing visual assessment of the goodness-of-fit.

*2) Observed Variable Models:* We model each observed variable, i.e., the classifier outputs, YOLO-derived statistics,

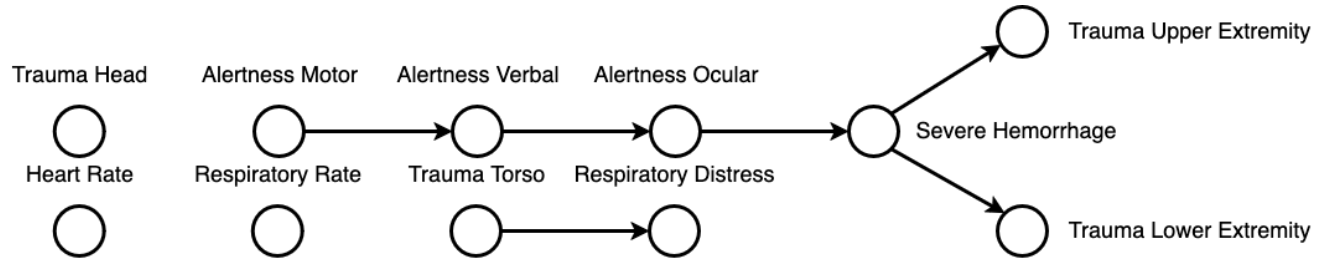


Fig. 6: Structured Bayesian Network.

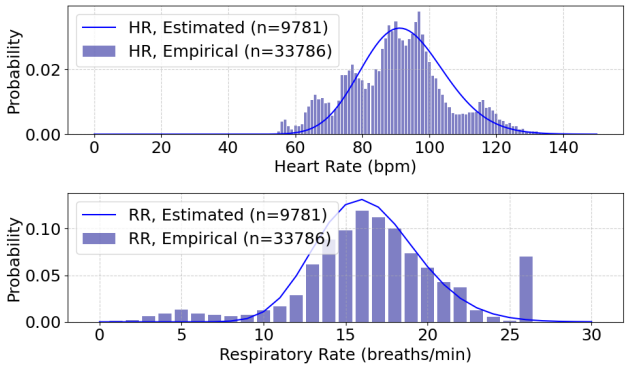


Fig. 7: Estimated heart and respiratory rate distributions overlaid on validation data set histograms.

and the mmWave radar measurements, as one of two classes of conditional distributions conditioned on the true latent injury state. The first is the Independent Fusion Model (IFM), a hierarchical Bayesian model used in prior work to fuse the outputs of $K$ independent classifiers [16].

Let $X$ be a latent injury variable with a categorically distributed prior. Let $y^1, \dots y^K$ be observed classifier outputs for $K$ classifiers where $y^k \in \Delta^{J-1}$, the probability simplex over $J$ classes and $k = 1, \dots, K$. Conditioned on $X = x$, each classifier output $y^k$ is drawn from a Dirichlet distribution with parameters $\alpha^k$, encoding the skill of the classifier at labeling samples of that class.

$$X \sim \text{Categorical}(p), \quad p \in \Delta^{J-1}$$
$$Y^k \mid X = x \sim \text{Dirichlet}(\alpha^k), \quad k = 1, ..., K$$

For the second class of models, observed values $y$ are assumed to follow Gaussian distributions with known mean and covariance conditioned on the true latent injury state. It models continuous-valued sensor outputs such as heart and respiratory rates from mmWave radar as well as motor alertness statistics.

$$X \sim \text{Categorical}(p), \quad p \in \Delta^{J-1}$$
$$Y \mid X = x \sim \mathcal{N}(\mu, \sigma^2)$$

Parameters of both models were estimated using maximum likelihood estimation on modality-specific training datasets outlined in Section III-A.

Computing the posterior distribution of any latent injury variable $\{X_i\}_{i=1}^{12}$ proceeds in two steps: prediction and update. At some fixed time-interval, the prediction step is performed for each latent injury node in the DBN:

$$p(X_{i,t}|Y_{1:t}) = \sum_{X_{t-1}} p(X_t|X_{t-1})p(X_{t-1}|Y_{1:t-1}) \quad (1)$$

where $P(X_t|X_{t-1})$ represents a transition matrix encoding how injuries change over the course of the robotic assessment. For example, the likelihood that a person's verbal alertness is absent increases in the absence of classifier observations indicating the person is speaking. Correspondingly, in the transition matrix for verbal alertness, the absent state is absorbing. Whenever an observation is received from any robotic platform, the posterior distribution of every latent injury variable changes according to the update step:

$$p(X_{i,t}|Y_{1:t}) \propto p(Y_t|X_{i,t})p(X_{i,t}|Y_{1:t-1}) \quad (2)$$

where $P(X_t|Y_{1:t-1})$ factorizes into a contribution from the posterior distributions over the parents of node $i$ and a contribution from the posterior distribution of node $i$ from the last time step $t-1$. This same update step applies regardless of how long it has been since the previously received measurement.

The proposed late-stage sensor fusion framework offers several practical benefits over early-stage fusion paradigms. First, the probabilistic model supports distributed and asynchronous evidence accumulation: observations collected independently by different platforms, such as a UAV and a UGV at different times, can be incrementally fused via Bayesian updates. This contrasts to early-stage fusion approaches, which rely on temporally aligned or co-located sensor inputs. Second, the late-fusion architecture offers strong flexibility and extensibility. New sensing modalities or sensor platforms can be incorporated into the system by defining appropriate observation models, without modifying the latent variable model or requiring that all sensors be active simultaneously. This property is particularly important in the MCI context, where no unified data sets exist containing all relevant modalities, but individual models can still be trained independently on modality-specific data sets.

## IV. MOCK MASS-CASUALTY FIELD EXPERIMENTS

This section evaluates the reliability and robustness of our automated triage system in several timed, mock mass-casualty field experiments.

### A. Experimental Setup

We conducted two experimental trials at a testing site under both daytime and nighttime conditions. Each trial followed a standard procedure. Prior to the start of the run, all casualties were given wearable heart and respiratory rate monitors, placed at predefined locations, and instructed to exhibit their assigned injuries. Robots, including three customized Spots and two custom aerial drones, were staged in a designated launch area. When the 20-minute timer began, human operators navigated the robots to each casualty. Upon arrival, the operator triggered the robot's autonomous assessment routine, during which the system collected multimodal observations and transmitted them to a base station for probabilistic inference. Using both the SBN and ILN models, the base station estimated the most likely injury pattern and submitted a structured casualty report to a mock DARPA server, where it was scored in real time. Robots continued assessing casualties until the 20-minute window elapsed, at which point the trial was terminated.

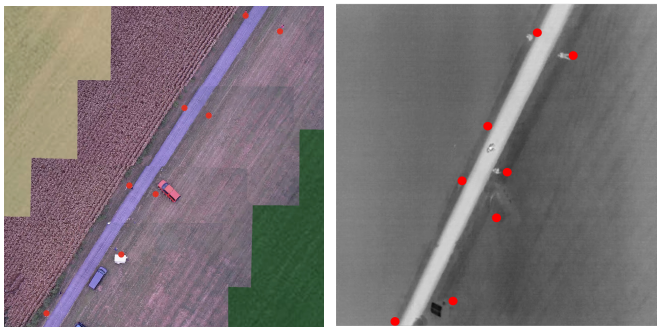

Fig. 8: Color and thermal aerial images of the test site overlaid with ground truth casualty locations (red).

### B. Customized Hardware Platforms

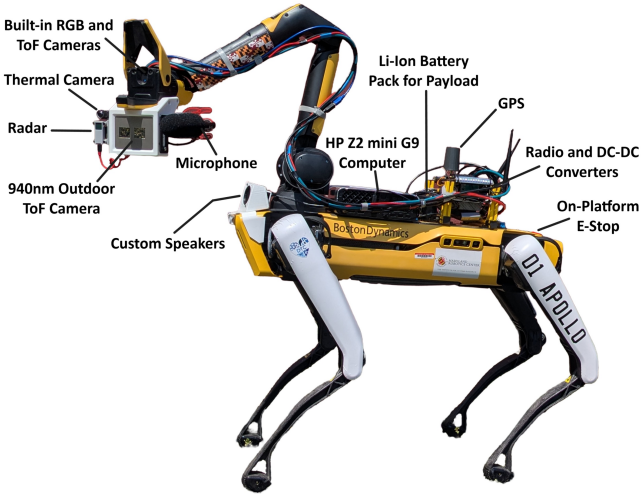


Fig. 9: Boston Dynamics Spot with customized hardware setup capable of performing medical triage.

*1) Spot:* We equipped a Boston Dynamics Spot robot with a custom sensor and computing payload, as shown in Figure 9. The payload supplies up to 280 watts to an HP Z2 mini G9 workstation, which includes an Intel i9-14900K CPU, an NVIDIA RTX 4000 Ada Generation GPU, and 64GB of RAM. The Spot arm is fitted with a 4K color camera and a FLIR Boson thermal camera for person detection and injury classification for daytime and nighttime operations, respectively, a Basler blaze-101 Time-of-Flight (ToF) camera for casualty localization, a RØDE NTG VideoMic for listening to verbal and nonverbal vocalizations, and a Texas Instruments IWR6843 mmWave radar for vitals estimation. A custom speaker enables the robot to converse with the casualty. For ego localization, we use an Emlid M2 multiband GPS. These integrated components enhance Spot's autonomous capabilities for field operations.

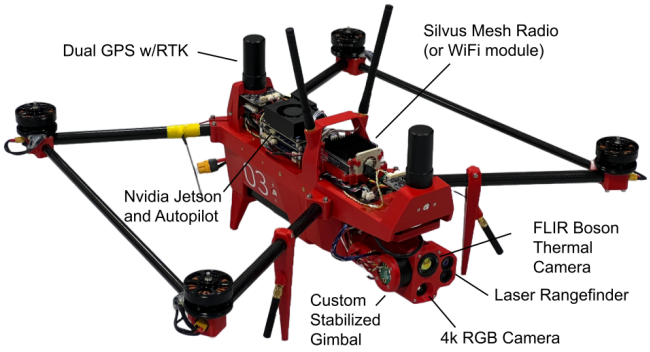


Fig. 10: Custom-made Chimera D UAV capable of performing medical triage. Propellors removed in image.

*2) Chimera:* The Chimera UAV is a custom research test platform, designed and built by University of Maryland, UAS Research and Operations Center. It is a mostly 3D printed design to allow significant customization for any specific project. The D model shown in Figure 10 is 2.6kg flight ready. It includes a custom brushless stabilized 2-axis gimbal with 4K color and 640x512 thermal cameras and a laser rangefinder attached. It can use either a WiFi module or Silvus Mesh radio for network connectivity. Two ARK RTK GPS units are used to provide system location, heading via GPS, and RTK capability for increased accuracy. Flight endurance is approximately 45 minutes on a 25.2V, 10Ah Li-Ion battery. The aircraft uses PX4 firmware on its autopilot, which can communicate directly with the onboard NVIDIA Jetson Orin NX 16GB SOC to enable autonomous behavior.

### C. Software Architecture

The software architecture is built on the Robot Operating System 2 (ROS2) [23]. All system components, including UGVs, UAVs, Operations Center computers communicate over a ROS2 network.

Dedicated ROS2 wrapper nodes running on the Spot's companion computer publish measurements from each sensor. The models described in Section III-A operate on those sensor streams to generate classifier outputs. For every sensing event, the Spot publishes a custom "Observation" message containing the classifier or sensor output and the estimated location of the casualty at the time of measurement; in this way, every observation message contains a position

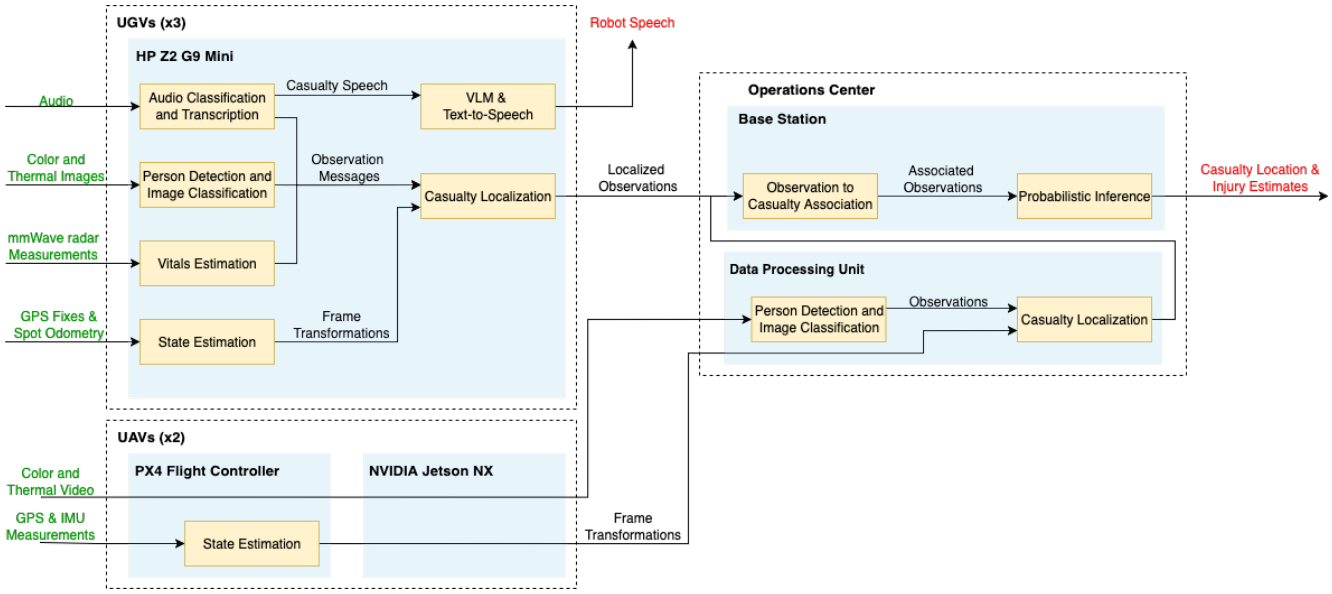


Fig. 11: Software architecture diagram illustrating flow of raw sensor data from UGVs and UAVs through onboard classifiers to probabilistic inference engine on base station. The diagram depicts software modules (yellow boxes), compute resources (blue boxes), input signals (green text), internal signals (black text), and output signals (red text).

estimate of the casualty in the world frame. In addition to injury-specific classifiers, each Spot deploys a VLM, a text-to-speech model, and a speech-to-text model, enabling verbal interaction with casualties.

The Chimera UAV's PX4 telemetry data is sent over the air to the ground control station (GCS) via MAVLink where it is translated by MAVROS into ROS2 messages. Color and thermal camera imagery are streamed using GStreamer.These sensors streams are operated on by the classifiers to generate localized Observation messages.

The base station computer subscribes to the localized Observation messages being published by each robotic platform. If any localized observation is not within 2 meters of a known casualty, the system assumes the observation corresponds with a new casualty whose estimated position equals that of the localized observation. If a localized observation is within 2 meters of one or more known casualties, the localized observation gets associated with the closest known casualty. A Kalman filter updates the associated casualty's location estimate using the location component of the localized observation message. The classifier output or sensor measurement component of the localized observation message is fed into the DBN. The resulting injury estimates are made available to downstream software components. Figure 11 depicts this architecture.

### *D. Results from Mock Mass-Casualty Incidents*

Table V reports the average accuracy per casualty of the proposed system under various ablation conditions across two test scenarios conducted during the day and at night, where no visible spectrum illumination was used. To evaluate system performance, each estimated casualty report, which consists of a casualty's estimated injuries and location, was first associated with a ground truth report. Estimated reports

TABLE V: Average Classification Accuracy (%)

| Ablation | Test 1 – Day (SBN, ILN) | Test 2 – Night (SBN, ILN) | Avg. Accuracy (SBN, ILN) |
|---|---|---|---|
| All modalities | (70.0, 62.9) | (65.0, 64.3) | (**67.5**, 63.6) |
| Only UGVs | (70.0, 62.9) | (65.0, 64.3) | (**67.5**, 63.6) |
| Only UAVs | (58.3, 58.3) | (58.0, 58.0) | (**58.2**, **58.2**) |
| No camera | (62.1, 62.1) | (65.0, 64.3) | (**63.6**, 63.2) |
| No radar | (70.0, 62.9) | (63.6, 62.9) | (**66.8**, 62.9) |
| No audio | (62.9, 60.0) | (59.4, 59.4) | (**61.2**, 59.7) |

located more than four meters from any ground truth casualty were discarded. In cases where multiple casualty reports fell within four meters of the same ground truth casualty, associations were resolved according to the official DARPA rules to maximize points. For each association, we computed the proportion of correctly classified fields, averaging over all associations. This ensures that reports are accurate with respect to both location and injuries. The reason for reporting the accuracy between the ground truth injury label and a MAP classification of our probabilistic estimate of that injury label is to be consistent with DARPA scoring, which requires providing classifications, not probability distributions.

Results show that the system can localize and estimate injuries reasonably well. Across all tests, all casualties were localized within the four meter requirement, and the proportion of correct injury assessments is well above random guessing with well-informed priors, but injury assessment models have significant room for improvement. The results show that both the SBN outperforms the ILN both during the day and at night. We believe this is a consequence of the SBN network encoding more information about correlated injury states. Correspondingly, we find that the SBN network is more robust to sensor dropout than the ILN netowrk, as evidenced by the SBN reporting higher average accuracy under various ablations.

These findings highlight the system's robustness to missing modalities and sensor platforms, its ability to generalize across environments, and the impact of using a structured Bayesian network over an independent latent network. The ablations show that the network relies much less on the UAVs than the UGVs for assessment. This is a direct consequence of the UAVs only having a camera for doing assessment, whereas the UGV additionally has a microphone and radar. The UAVs also have a larger safety stand-off distance from the casualty compared to the UGV, precluding detailed imagery of their injuries. Unsurprisingly, the camera ablation has no consequence for the average accuracy at night, since no camera data is available due to the absence of visible light during testing. While these results are promising, more field experiments must be carried out to further tune the DBN model parameters in order to improve the average accuracy per casualty and to validate the system in more diverse conditions.

## V. CONCLUSION

This paper presented a robotic system for casualty assessment in mass-casualty incidents using multimodal, non-contact sensing and probabilistic inference. Results demonstrate that the proposed approach is both robust to sensor noise and reliable in producing accurate injury estimates under operational constraints.

In ongoing and future work, we are also exploring active sensing strategies that use intermediate assessment results to inform how robots should autonomously position their sensors, allowing for adaptive data collection that improves diagnostic accuracy and reduces operator burden.

## DISCLAIMER

The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

## ACKNOWLEDGMENT

Training data for this work was collected by the US Army Telemedicine and Advanced Technology Research Center (TATRC) and co-owned with DARPA.

We thank Alexandra Mangel, Atharv Marathe, Jeremy Kuznetsov, and Arian Moradi, from the University of Maryland (UMD), College Park; Tony Christiani from C-STARS Baltimore at R. Adams Cowley Shock Trauma Center; and Joshua Schmucki, Grant Williams, and Rob Neuner from the UMD UAS Research and Operations Center.

## REFERENCES

[1] J. Mistovich, K. Karren, and B. Hafen, *Prehospital Emergency Care*, 10th ed. Pearson, 2013.

[2] D. A. R. P. A. (DARPA), "DARPA Triage Challenge — About The Challenge," 2025, last accessed 1 August 2025. [Online]. Available: https://triagechallenge.darpa.mil/about

[3] H. Carrión, M. Jafari, M. D. Bagood, H. Y. Yang, R. R. Isseroff, and M. Gomez, "Automatic wound detection and size estimation using deep learning algorithms," *PLoS Computational Biology*, vol. 18, no. 3, p. e1009852, 2022.

[4] C. West, B. Kaus, S. O. Sullivan, H. Schneider, and O. Seifert, "Using infrared cameras in drones to detect bleeding events," *BMC Emergency Medicine*, vol. 23, no. 1, p. 142, 2023.

[5] F. S. Saeed, A. A. Bashit, V. Viswanathan, and D. Valles, "An initial machine learning-based victim's scream detection analysis for burning sites," *Applied Sciences*, vol. 11, no. 18, p. 8425, 2021.

[6] P. Zhao, C. Lu, B. Wang, C. Chen, L. Xie, M. Wang, N. Trigoni, and A. Markham, "Heart rate sensing with a robot mounted mmwave radar," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 2812–2818.

[7] Y. Chen, J. Yuan, and J. Tang, "A high precision vital signs detection method based on millimeter wave radar," *Scientific Reports*, vol. 14, p. 25535, 2024.

[8] H. Shi, H. Zhao, Y. Liu, W. Gao, and S.-C. Dou, "Systematic analysis of a military wearable device based on a multi-level fusion framework: Research directions," *Sensors*, vol. 19, no. 12, p. 2651, 2019.

[9] J. Krygier, P. Lubkowski, K. Maslanka, A. P. Dobrowolski, T. Mrozek, W. Znaniecki, and P. Oskwarek, "Smart medical evacuation support system for the military," *Sensors*, vol. 24, no. 14, p. 4581, 2024.

[10] J. M. Warnecke, J. Lasenby, and T. M. Deserno, "Robust in-vehicle respiratory rate detection using multimodal signal fusion," *Scientific Reports*, vol. 13, p. 20435, 2023.

[11] X. Yang, X. Zhang, Y. Ding, and L. Zhang, "Indoor activity and vital sign monitoring for moving people with multiple radar data fusion," *Remote Sensing*, vol. 13, no. 18, p. 3791, 2021.

[12] D. H. Choi, K. J. Hong, S. D. Shin, S. Kim, M. Chung, K. H. Kim, K. J. Song, M. Cho, D. Yoon, and J. Lee, "Measurement of level of consciousness by AVPU scale assessment system based on automated video and speech recognition technology," *The American Journal of Emergency Medicine*, vol. 74, pp. 112–118, 2023.

[13] H. C. Kim and Z. Ghahramani, "Bayesian classifier combination," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. Proceedings of Machine Learning Research, vol. 22. PMLR, 2012, pp. 619–627.

[14] A. Nazabal, P. Garcia-Moreno, A. Artes-Rodriguez, and Z. Ghahramani, "Human activity recognition by combining a small number of classifiers," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 5, pp. 1342–1351, 2016.

[15] G. Pirš and E. Štrumbelj, "Bayesian combination of probabilistic classifiers using multivariate normal mixtures," *Journal of Machine Learning Research*, vol. 20, no. 51, pp. 1–18, 2019.

[16] S. Trick and C. Rothkopf, "Bayesian classifier fusion with an explicit model of correlation," in *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. Proceedings of Machine Learning Research, vol. 151. PMLR, 2022, pp. 2282–2310.

[17] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[18] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[19] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proceedings of Interspeech*, 2021, pp. 571–575.

[20] J. Gemmeke, D. Ellis, D. Freedman, A. J. W. Lawrence, R. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[21] Texas Instruments, "Building a multipatient contactless vital signs sensor for at-home use with mmwave," Texas Instruments, Dallas, TX, USA, Tech. Rep. SLYT836, 2021. [Online]. Available: https://www.ti.com/lit/an/slyt836/slyt836.pdf?ts=1768903199998

[22] N. Kitson, A. Constantinou, Z. Guo, Y. Liu, and K. Chobtham, "A survey of Bayesian Network structure learning," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8721–8814, 2023.

[23] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot operating system 2: Design, architecture, and uses in the wild," *Science Robotics*, vol. 7, no. 66, p. eabm6074, 2022. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.abm6074