I.J. Image, Graphics and Signal Processing, **2026, 3, 17-53**
Published Online on June 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2026.03.02

Methods and Tools for Identifying Human Resource Lesions in Emergency Based on Multimodal Analysis and Deep Learning

Yurii Ushenko*
Department of Physics, Shaoxing University, Shaoxing, Zhejiang Province 312000, China
Department of Computer Science, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine
E-mail: y.ushenko@chnu.edu.ua
ORCID iD: https://orcid.org/0000-0003-1767-1882
*Corresponding author

Dmytro Uhryn
Department of Computer Science of the Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine
E-mail: d.ugryn@chnu.edu.ua
ORCID iD: https://orcid.org/ 0000-0003-4858-4511

Victoria Vysotska
Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, 79013, Ukraine
Combating Cybercrime Department, Kharkiv National University of Internal Affairs, Kharkiv, 61080, Ukraine
E-mail: Victoria.A.Vysotska@lpnu.ua
ORCID iD: https://orcid.org/0000-0001-6417-3689

Lyubomyr Chyrun
Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, 79013, Ukraine
Computer Science Department, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine
Applied Mathematics Department, Ivan Franko National University of Lviv, Lviv, 79000, Ukraine
E-mail: lyubomyr.v.chyrun@lpnu.ua
ORCID iD: https://orcid.org/0000-0002-9448-1751

Zhengbing Hu
School of Computer Science and Artificial Intelligence, Hubei University of Technology, Wuhan, China
E-mail: drzbhu@gmail.com
ORCID iD: https://orcid.org/0000-0002-6140-3351

Tetiana Rekunenko
Department of Education Quality Assurance, Kharkiv National University of Internal Affairs, 27, L. Landau Avenue, 61080 Kharkiv, Ukraine
Department of Administrative Law and Process, Donetsk State University of Internal Affairs, 1, Velyka Perspektyvna Street, Kropyvnytskyi, 25000, Ukraine
E-mail: rekunenko22@ukr.net
ORCID iD: https://orcid.org/0000-0001-7668-0581

Received: January 15, 2026; Revised: February 23, 2026; Accepted: March 18, 2026; Published: June 8, 2026

Abstract: Emergencies of natural, technological, and military origin require rapid and accurate assessment of victims' conditions to support effective rescue and medical response. Traditional visual examination methods are often limited by stress, time pressure, and incomplete information, leading to delayed or inaccurate decisions. This study proposes a multimodal deep learning approach for automated identification of human resource lesions in emergency scenarios. The developed framework integrates visual, audio, and text/sensory data using convolutional neural networks, Transformer-based models, and a Transformer Cross-Attention fusion mechanism. The proposed architecture enables effective extraction and integration of heterogeneous features for lesion classification, severity estimation, and automated medical triage. Experimental evaluation was conducted on multimodal datasets containing injury images, audio recordings, and symptom descriptions. The model was trained using a combined loss function and evaluated with

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

classification, regression, and triage metrics. The results demonstrate high system performance, achieving a macro-F1 score of 0.87, validation accuracy of 86–87%, and triage accuracy above 90%, including 95% for the RED category. The regression model for severity prediction achieved an R^2 value of 0.92, while modality importance analysis confirmed the dominant contribution of visual information. The experiments also showed stable model convergence and strong generalisation ability without significant overfitting. The proposed multimodal framework confirms the effectiveness of deep learning and cross-attention mechanisms for automated lesion identification and emergency medical triage. The developed approach can be applied in decision-support systems for rescue operations, emergency medicine, and intelligent VR/AR training simulators.

Index Terms: Multimodal Analysis, Deep Learning, Lesion Identification, Emergencies, Transformer Cross-Attention, Medical Triage, Triage, Neural Networks, Decision Support

1. Introduction

Emergencies of natural, technological, and military origin require rapid assessment of victims' conditions to ensure timely medical assistance and effective rescue operations. In such environments, traditional approaches based on visual inspection and human expertise may be insufficient because rescuers often operate under time pressure, stress, limited resources, and incomplete information. These factors increase the risk of diagnostic errors and delays in decision-making.

Recent advances in artificial intelligence and deep learning have created new opportunities for automating the analysis of human conditions in emergency scenarios. In particular, multimodal analysis enables the simultaneous processing of heterogeneous data sources, including images, audio signals, text descriptions, and physiological indicators. Combining multiple modalities provides a more comprehensive representation of victims' conditions and improves the reliability of automated lesion identification systems.

Deep learning methods, such as Convolutional Neural Networks (CNNs), Vision Transformers, and Transformer-based Natural Language Processing (NLP) models, have demonstrated high efficiency in computer vision, speech analysis, and medical data interpretation. Their integration into multimodal architectures allows automated extraction of informative features and improves classification accuracy, severity estimation, and medical triage.

This study focuses on developing methods and models for automated identification of human resource lesions in emergencies through multimodal analysis and deep learning. The proposed approach integrates visual, audio, and text/sensory data via a Transformer Cross-Attention mechanism, enabling the effective fusion of heterogeneous features and the objective assessment of victims' conditions. The main objectives of the study are:

1. To analyse current approaches to multimodal analysis and artificial intelligence in emergency medicine.
2. To develop a multimodal framework for integrating visual, audio, and text/sensory data.
3. To implement deep learning methods for feature extraction and lesion classification.
4. To design a Transformer Cross-Attention mechanism for multimodal fusion.
5. To develop a severity assessment and automated medical triage model.
6. To experimentally evaluate the effectiveness of the proposed approach using quantitative metrics.

The object of the study is the process of automated identification and assessment of victims' conditions in emergencies. The study focuses on multimodal deep learning methods, Transformer-based fusion architectures, and medical triage models for lesion identification and severity prediction.

The research hypothesis is that integrating multiple modalities via Transformer Cross-Attention can significantly improve the accuracy and robustness of lesion identification compared with single-modality approaches, particularly in noisy or incomplete data typical of emergency environments.

2. Related Works

In modern science and practice of automated identification of human resource lesions in emergencies, methods of multimodal analysis and deep learning are actively investigated. An important aspect is integrating data of various types (images, audio, text/sensory data) to improve the accuracy of injury classification and severity assessment. Most modern research focuses on the processing of visual data, including images and videos of victims. In general, algorithms for classifying injuries using convolutional neural networks (CNNs) have achieved high accuracy in detecting bleeding and burns. However, an approach based only on visual data has limitations in complex or combined lesions. To increase reliability, audio and sensory data are used, which allow the patient's condition to be taken into account beyond visual information. For example, voice analysis and alarm systems allow you to assess the victim's pain and stress levels.

Multimodal analysis enables you to integrate data from multiple sources to provide a comprehensive assessment of the human condition. The use of deep learning fusion models (CNN with RNN, attention-based fusion, Transformer cross-attention) has demonstrated high efficiency in classifying injuries and assessing severity [1]. Such approaches allow:

– Recognise combined types of injuries.
– Increase triage accuracy through the integration of multiple data sources.
– Adapt to limited or incomplete data due to the ability of models to learn important traits on their own.

One modern approach is to use Transformer-based cross-attention to integrate multimodal features. Studies show that cross-attention enables the model to effectively weigh the contributions of each modality in forming an integrated representation of features. This approach is particularly effective for classifying combined lesions and evaluating complex cases in which individual modalities may be incomplete or noisy.

An analysis of existing studies has shown the advantages of the multimodal approach, including higher accuracy rates, greater reliability in triage, and the ability to classify combined injuries. The disadvantages of homogeneous models are that visual or text-only models do not provide full coverage of emergencies. The novelty of Transformer-based cross-attention approaches lies in integrating features from different modalities, enabling the model to independently assess the relevance of each modality to a specific forecast. Thus, modern research confirms the effectiveness of integrating features and using deep neural networks with cross-attention for automated lesion identification and triage of victims in emergencies. Such studies can be divided into several areas:

1. Multimodal convergence and data synthesis architectures [1-4].
2. Specificity of emergencies and data deficit [5-7].
3. Diagnosis of specific lesions: Head and musculoskeletal injuries [8-1].
4. Generative methods and image quality improvement [12-18].

The central issue of the study is the combination of different types of data (text, images, and clinical indicators). The authors in [1] present a breakthrough Transformer-based model that unifies the processing of multimodal inputs. It allows the model to determine the correlations between heterogeneous features without complex preprocessing independently. Research [2-3] lays a theoretical foundation by analysing the fusion techniques of medical images and text reports. They emphasise that integrating visual cues with the semantics of clinical descriptions critically increases the accuracy of diagnosing complex diseases. The authors in [4] propose a confidence-guided multi-level fusion method, which is particularly useful for NS, where some data may be incomplete or noisy.

In emergencies, there is often a lack of time and quality datasets to train models. The authors in [5] propose an on-scene decision model that combines NLP and ML to classify injuries under limited data. It is a key job for your "human resource" research in the field. The authors in [6] focus on emergency dispatch, using textual analysis to predict patients' condition severity before they arrive at the hospital. The authors in [7] propose the MIND framework, which uses knowledge distillation to improve clinical prognosis, enabling the creation of "lighter" and faster models for mobile devices used by rescue services. Separate works focus on the most critical lesions characteristic of disasters. The authors in [8] develop a model for predicting intracranial infections after severe TBI (traumatic brain injury) in children using multimodal data. The authors in [9] provide a comprehensive overview of the role of AI in neurorehabilitation after TBI, outlining the future of human resource recovery. The authors in [10] investigate the use of multimodal large language models (MLLMs) to detect fractures in orthopaedic traumatology – this is the most recent approach to "understanding" X-rays through dialogue with the model. The authors in [11] present DeepStroke – a system for rapid stroke screening in emergency departments based on adversarial learning. The authors in [12] with the MultiViT2 model use latent diffusion models (LDMs) to augment neuroimaging data. It allows for the "completion" of missing data fragments or for improving the quality of images obtained with legacy or portable equipment in disaster areas.

The papers [13-18] are not always directly about the identification of lesions, but highlight key technological aspects related to multimodal deep learning in medicine:

Table 1. Comparison table of key technologies.

Technology	Sources	Applied value for emergencies
Multimodal Transformers	[1-3]	Automatic synthesis of CT/MRI data and anamnesis.
NLP with ML	[5-6]	Quick triage based on symptom descriptions.
MLLM (Multimodal LLMs)	[10]	Intellectual support of the doctor through natural language.
Generative AI / Diffusion	[12]	Work in conditions of low-quality or absent diagnostic images.

The literature analysis shows a transition from simple classifiers to adaptive multimodal systems capable of operating under conditions of uncertainty. For your work, the most promising approach is a combination of Transformer-based architectures (for processing different modalities) and knowledge distillation methods (for deployment on mobile terminals used by emergency services).

Existing studies on multimodal medical analysis and emergency response systems demonstrate significant progress in integrating heterogeneous data sources; however, most approaches remain focused on narrow clinical tasks or operate under controlled conditions. The proposed pipeline differs from previous work by combining multimodal fusion, emergency-oriented lesion identification, and automated triage within a unified architecture designed for uncertain, resource-constrained environments.

Transformer-based multimodal models presented in studies [1–3] provide effective integration of medical images and textual clinical reports. These approaches achieve high diagnostic accuracy in hospital settings, particularly in radiology and disease analysis. Nevertheless, they are primarily designed for structured medical environments where high-quality imaging and complete patient records are available. In contrast, the proposed approach targets emergencies characterised by noisy signals, incomplete information, unstable acquisition conditions, and limited response time. Unlike prior studies, the developed system simultaneously processes visual, audio, and text/sensory data and explicitly focuses on rapid lesion identification and triage support in disaster environments. The confidence-guided fusion strategy proposed in [4] improves robustness under partially missing modalities. Although this approach addresses uncertainty, it primarily focuses on confidence estimation rather than on operational emergency decision-making. The proposed model extends this idea by introducing Transformer Cross-Attention not only for modality weighting but also for dynamic interaction between modalities during lesion classification and severity estimation. It enables the system to adaptively emphasise the most informative signals depending on the emergency context.

Research on emergency trauma classification using NLP and machine learning [5–6] demonstrates the usefulness of textual information for rapid triage. However, these systems are mostly limited to text-based dispatch analysis and do not exploit multimodal representations. Their performance strongly depends on the completeness and quality of textual descriptions. The proposed pipeline overcomes this limitation by integrating textual information with visual and audio evidence, significantly increasing robustness and reducing dependence on a single modality.

The MIND framework [7] introduces modality-informed knowledge distillation for lightweight multimodal prediction. Its primary contribution lies in model compression and efficient deployment. While highly relevant for mobile healthcare systems, the framework does not address multimodal lesion identification in emergency conditions or triage severity estimation. In comparison, the proposed approach provides a complete end-to-end pipeline that combines feature extraction, multimodal fusion, lesion classification, risk estimation, and medical triage within a single architecture.

Studies focused on specific injuries, such as traumatic brain injury [8–9], fracture detection [10], or stroke screening [11], demonstrate high effectiveness of AI methods in specialised diagnostic tasks. Nevertheless, these models are domain-specific and typically designed for one lesion category. They lack the flexibility needed to address combined or heterogeneous injuries commonly observed during disasters, military attacks, or large-scale emergencies. The proposed system addresses this gap by supporting multiclass lesion classification and integrated triage assessment for diverse injury categories. Generative and diffusion-based approaches [12] improve image quality and augment training datasets under limited data conditions. Although these techniques enhance model generalisation, they are auxiliary rather than decision-making systems. The proposed framework can incorporate such augmentation methods in future work, but currently contributes primarily through multimodal integration and emergency-oriented decision support.

A key limitation of many previous studies is the isolated treatment of classification and triage tasks. Existing approaches usually focus either on diagnosis accuracy or severity estimation independently. In contrast, the proposed method jointly optimises lesion classification, severity prediction, and triage categorisation using a combined loss function. It provides a more practically applicable framework for emergency medicine and rescue operations.

Another important distinction is the interpretability of modality contribution. Previous multimodal studies rarely quantify the relative importance of visual, audio, and textual information in the final prediction. The experimental analysis in the proposed work demonstrates that visual data account for approximately 50% of the predictive importance, while audio and text modalities account for 30% and 20%, respectively. This analysis provides additional insight into the behaviour of the multimodal system and supports further optimisation for real-world deployment.

Overall, the proposed pipeline advances previous research in several aspects:

– Integration of visual, audio, and text/sensory modalities within a unified emergency-oriented framework.
– Application of Transformer Cross-Attention for adaptive multimodal interaction under noisy and incomplete data conditions.
– Simultaneous optimisation of lesion classification, severity prediction, and medical triage.
– Adaptation to emergency and disaster scenarios rather than controlled clinical environments.
– Quantitative analysis of modality importance and system interpretability.
– Potential applicability to rescue systems, mobile emergency platforms, and VR/AR medical simulators.

Thus, compared with existing multimodal and cross-attention approaches, the proposed method provides a more comprehensive, adaptive, and practically oriented solution for automated lesion identification and emergency triage.

3. Methodology

Here is a generalised mathematical pipeline for automated identification of human resource injuries in emergencies based on multimodal analysis and deep learning. This approach describes the entire process – from data collection to lesion type classification and prioritising care. Automated Lesion Identification Pipeline:

1. Formation of a multimodal data set.
2. Preprocessing of data.
3. Feature Extraction.
4. Multimodal integration (Fusion).
5. Classification of the type of lesion.
6. Severity Assessment (Triage)
7. Loss function for model training.
8. Evaluation metrics.
9. Analysis of evaluation metrics.

Let the system receive information from several sources (modalities): images or videos of the victims; audio signals (screaming, breathing, voice); text or sensor data (description, medical indicators, IoT).

A multimodal dataset can be presented as: $D = \{(x_i^{(v)}, x_i^{(a)}, x_i^{(t)}, y_i)\}_{i=1}^{N}$, where $x_i^{(v)}$ – visual data (image/video), $x_i^{(a)}$ – audio data, $x_i^{(t)}$ – text or sensory data, y_i – class of lesion (bleeding, burn, trauma, shock, etc.), N – number of examples. For each modality, normalisation and preparation are performed:

– Visual data: $\widetilde{x_i^{(v)}} = \frac{x_i^{(v)} - \mu_v}{\sigma_v}$, where μ_v is the average value of pixels, σ_v is the standard deviation.

– Audio data, i.e. audio signal is translated into a spectrogram: $S(t,f) = \left|STFT\left(x_i^{(a)}\right)\right|$, where $STFT$ is a short-term Fourier transform.

– Text/touch data, in particular, conversion to vector representation:

$$x_i^{(t)} = Embedding(w_1, w_2, \dots, w_n) \text{ or } x_i^{(t)} = TFIDF(w_1, \dots, w_n).$$

For each modality, deep neural networks are used:

– Visual model: $h_i^{(v)} = f_v\left(\widetilde{x_i^{(v)}}; \theta_v\right)$, where f_v – CNN or Vision Transformer, θ_v – network parameters.

– Audio model: $h_i^{(a)} = f_a(S(t,f); \theta_a)$, where f_a is CNN/RNN/Transformer for audio.

– Text model: $h_i^{(t)} = f_t\left(x_i^{(t)}; \theta_t\right)$, where f_t is the NLP model (for example, Transformer).

Signs are combined into a common space. Concatenation of signs: $z_i = \left[h_i^{(v)} \oplus h_i^{(a)} \oplus h_i^{(t)}\right]$, where $\oplus$ is the merge operation. A more complex option is Attention fusion:

$$z_i = \sum_{m=1}^{M} \alpha_m h_i^{(m)} \text{ and } \alpha_m = \frac{e^{w^T h_i^{(m)}}}{\sum_{k=1}^{M} e^{w^T h_i^{(k)}}},$$

where M is the number of modalities, after combining the features, a classifier is used $\hat{y}_i = Softmax(Wz_i + b)$, where W – classifier weights, b is the offset. Class Probability:

$$P(y = c|z_i) = \frac{e^{w_c^T z_i}}{\sum_{k=1}^{C} e^{w_k^T z_i}},$$

where C is the number of types of lesions. To determine the priority of medical care, the risk function is introduced: $R_i = \beta_1 P_{critical} + \beta_2 HR + \beta_3 BloodLoss$, where HR – heart rate, $BloodLoss$ – blood loss assessment, β – weight factors. Classification according to the Triage system:

$$T_i = \begin{cases} Red, R_i > \tau_1, \\ Yellow, \tau_2 < R_i \leq \tau_1, \\ Green, R_i \leq \tau_2. \end{cases}$$

The model learns by minimising cross-entropy loss: $L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic} \log(\widehat{y_{ic}})$.

Precision, Recall, F1-score and ROC-AUC are used to evaluate the system:

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}, F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}, AUC = \int_0^1 TPR\left(FPR^{-1}(x)\right)dx.$$

The function can represent the complete process: $\hat{y}_i = F\left(x_i^{(v)}, x_i^{(a)}, x_i^{(t)}; \Theta\right)$, where

$$F = Classifier\left(Fusion\left(f_v\left(x^{(v)}\right), f_a\left(x^{(a)}\right), f_t\left(x^{(t)}\right)\right)\right).$$

General scheme of the pipeline:

Multimodal Data → Preprocessing → Feature Extraction (Deep Learning - CNN / Transformer) → Multimodal Fusion(Concatenation / Attention) → Lesion Type Classification → Severity Assessment (Triage) → Decision Support

Let us describe in more detail the mathematical model of multimodal integration (fusion) based on the Transformer Cross-Attention mechanism for automated identification of human resource injuries in emergencies. This approach enables the efficient integration of information from diverse sources (visual, audio, and text/sensory data) into a single representation for further classification. Pipeline multimodal integration based on Transformer Cross-Attention:

1. Representation of multimodal features.
2. Formation of a sequence of multimodal tokens.
3. Cross-Attention mechanism.
4. Multimodal Cross-Attention
5. Multi-Head Cross Attention
6. Formation of multimodal representation
7. Classification of lesions
8. Loss function
9. Evaluation metrics.
10. Analysis of evaluation metrics.

After the stage of highlighting features using depth models, we get feature vectors for each modality: $h^{(v)} \in R^{d_v}$, $h^{(a)} \in R^{d_a}$, $h^{(t)} \in R^{d_t}$, where $h^{(v)}$ – vector of image or video features (CNN / Vision Transformer), $h^{(a)}$ – audio signs (Audio CNN / Transformer), $h^{(t)}$ – textual or sensory signs, d_v, d_a, d_t – are the dimensions of the corresponding feature spaces.

For unification, all vectors are projected into a common latent space of dimension d: $\widetilde{h^{(m)}} = W_m h^{(m)} + b_m$, where $m \in \{v, a, t\}$, $W_m \in R^{d \times d_m}$. Modality vectors form a sequence: $H = \left[\widetilde{h^{(v)}}, \widetilde{h^{(a)}}, \widetilde{h^{(t)}}\right]$, where $H \in R^{M \times d}$, M is the number of modalities. To integrate information between modalities, the Transformer cross-attention mechanism is used:

– Formation of requests, keys and values: $Q = HW_Q, K = HW_K, V = HW_V$, where $W_Q, W_K, W_V \in R^{d \times d_k}$.

– Calculation of the attention matrix: $Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, where $\frac{QK^T}{\sqrt{d_k}}$ is a scalable matrix of correlations between modalities.

For each modality, attention can be calculated relative to other modalities. For example, visual data derives context from audio and text: $Z^{(v)} = Attention\left(Q^{(v)}, K^{(a,t)}, V^{(a,t)}\right)$. Similarly: $Z^{(a)} = Attention\left(Q^{(a)}, K^{(v,t)}, V^{(v,t)}\right)$, $Z^{(t)} = Attention\left(Q^{(t)}, K^{(v,a)}, V^{(v,a)}\right)$. To increase the stability of learning, multi-headed attention is used: $MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W_O$, where $head_i = Attention\left(QW_Q^i, KW_K^i, VW_V^i\right)$. Integrated views are combined:

$$z = \frac{1}{M}\sum_{m=1}^{M} Z^{(m)}, \text{ or } z = Pooling\left(Z^{(v)}, Z^{(a)}, Z^{(t)}\right),$$

where z is a generalised multimodal vector of features, the final stage is the classification of the type of lesion: $\hat{y} = Softmax(W_c z + b_c)$, where $\hat{y} \in R^C$, C is the number of classes of lesions. Training is done through cross-entropy loss:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i).$$

Full Model Function: $\hat{y} = F_{cls}\left(F_{fusion}\left(f_v\left(x^{(v)}\right), f_a\left(x^{(a)}\right), f_t\left(x^{(t)}\right)\right)\right)$,

where f_v, f_a, f_t – trait selection models, F_{fusion} – Transformer cross-attention, F_{cls} – is a classifier. Conceptual scheme:

<[Visual Data → CNN/ViT → hv], [Audio Data → AudioNet → ha], [Text/Sensor → Transformer → ht]> →
Transformer Cross-Attention → Multimodal Representation → Injury Classification → Triage Level

The creation of a multimodal dataset is a key stage in developing a system for automated identification of human resource injuries in emergencies. The main objective of this stage is to create a coherent information space that brings together data of different natures – visual, auditory, and textual or sensory – for further analysis using deep learning techniques. Let's describe in detail the model of multimodal data representation. Let the system receive information from M s of different modalities. In general, a multimodal dataset can be represented as: $D = \{(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(M)}, y_i)\}_{i=1}^{N}$, where N – number of observations (examples); $x_i^{(m)}$ – data of the third example from the m modality; M is the number of modalities; y_i is a class label corresponding to the type of lesion.

In the context of identifying lesions in emergencies, the main modalities include: visual data – images or videos of victims; audio data – sounds, voice, breathing, alarms; text or sensory data – medical indicators, text descriptions of the victim's condition, data from wearable devices. The multimodal dataset can then be formalised as: $D = \{(x_i^{(v)}, x_i^{(a)}, x_i^{(t)}, y_i)\}_{i=1}^{N}$, where $x_i^{(v)}$ – visual information; $x_i^{(a)}$ – audio signal; $x_i^{(t)}$ – text or touch parameters; y_i is the class of damage.

The next step is to form the feature space. Each modality has its own data space: $x_i^{(v)} \in R^{H \times W \times C}$, where H – image height; W – width; C is the number of channels. The audio signal is given in the form of a time sequence: $x_i^{(a)} = \{s_1, s_2, \dots, s_T\}$, where T is the length of the signal. Text or sensory data is presented in the form of a vector of features: $x_i^{(t)} \in R^{d_t}$. One of the important tasks of forming a multimodal dataset is the temporal synchronisation of various sources of information (multimodal data). To do this, the matching function is introduced: $\Phi: (x^{(v)}, x^{(a)}, x^{(t)}) \to x^{(m)}$, which forms a coherent multimodal representation: $x_i^{(m)} = \Phi(x_i^{(v)}, x_i^{(a)}, x_i^{(t)})$. The Φ function performs time synchronisation of signals, equalisation of the sample rate, and normalisation of scale.

The next step is to annotate and classify the lesions. To train the model, data labelling is required. Let there be a set of classes of lesions: $Y = \{c_1, c_2, \dots, c_K\}$, where K is the number of types of injuries. Then each example is corresponded with a label: $y_i \in Y$. For the medical triage problem, an extended representation is possible: $y_i = (c_i, s_i)$, where c_i – type of lesion; s_i – severity. Let us now describe the statistical model of a multimodal set. Let the shared distribution of multimodal data look like: $P(x^{(v)}, x^{(a)}, x^{(t)}, y)$. The task of training is to build a model: $f: (x^{(v)}, x^{(a)}, x^{(t)}) \to y$, that maximises plausibility: $\theta^* = \arg\max_\theta \prod_{i=1}^{N} P(y_i | x_i^{(v)}, x_i^{(a)}, x_i^{(t)}; \theta)$, where θ are the parameters of the model, the use of a multimodal dataset allows you to increase the accuracy of lesion identification by integrating information from different sources. Formally, this can be presented as: $I_{multi} > I_{single}$, where I is the information completeness of the data. From the point of view of information theory, the common entropy of modalities: $H(X^{(v)}, X^{(a)}, X^{(t)})$, is less than the sum of the entropies of individual modalities: $H(X^{(v)}, X^{(a)}, X^{(t)}) < H(X^{(v)}) + H(X^{(a)}) + H(X^{(t)})$, which indicates the presence of mutual information between modalities and the potential to improve forecasting accuracy.

After the completion of the formation stage, the generalised structure of the multimodal dataset is as follows:

$$D = \begin{bmatrix} x_1^{(v)} & x_1^{(a)} & x_1^{(t)} & y_1 \\ x_2^{(v)} & x_2^{(a)} & x_2^{(t)} & y_2 \\ \vdots & \vdots & \vdots & \vdots \\ x_N^{(v)} & x_N^{(a)} & x_N^{(t)} & y_N \end{bmatrix}.$$

This kit is used for further steps: preliminary data processing → highlighting features → multimodal integration → classification of types of lesions. Preprocessing of multimodal data is an important stage in building a system for automated identification of human resource injuries in emergencies. The main objective of this phase is to prepare data of various types for further analysis using deep learning models, reduce noise, eliminate inhomogeneities, and standardise the data format. Let the multimodal dataset look like: $D = \{(x_i^{(v)}, x_i^{(a)}, x_i^{(t)}, y_i)\}_{i=1}^{N}$, where $x_i^{(v)}$ – visual data; $x_i^{(a)}$ – audio signals; $x_i^{(t)}$ – text or sensory data; y_i is the class of damage.

Preprocessing involves applying separate transformations for each modality:

1. Normalisation of visual data.
2. Spatial scaling and alignment.
3. Removing noise in visual data.
4. Preprocessing of audio data.
4.1. Signal normalisation.
4.2. Fourier transform.
5. Preprocessing of text and sensory data.
6. Coordination of multimodal data.

7. Decrease in dimensionality.

Visual data (images or videos) can vary in brightness, contrast, and resolution. Therefore, the first step is to normalise the pixel values. Let: $x_i^{(v)} \in R^{H\times W\times C}$, where H – image height; W – width; C is the number of channels. Normalisation is performed according to the formula:

$$\widetilde{x_i^{(v)}} = \frac{x_i^{(v)}-\mu_v}{\sigma_v}, \quad \sigma_v = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i^{(v)}-\mu_v\right)^2},$$

where μ_v is the average value of pixels in the sample: $\mu_v = \frac{1}{N}\sum_{i=1}^{N} x_i^{(v)}$, and σ_v is the standard deviation. It stabilises the learning process of neural networks and helps avoid gradient scaling problems.

Since images can have different sizes, the conversion to a single format is performed: $x_i^{(v)} = Resize\left(x_i^{(v)}, H_0, W_0\right)$, where H_0, W_0 – are the standard image sizes. In emergencies, images can exhibit significant noise. To reduce it, smoothing is used, for example, a Gaussian filter: $x_i^{(v)} = x_i^{(v)} * G(\sigma)$, where $*$ is the convolution operation:

$$G(\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}.$$

Audio signals may contain environmental noise (such as explosions, machinery, or crowds). The audio signal is set as a time series:

$$x_i^{(a)} = \{s_1, s_2, \dots, s_T\}, \quad \widetilde{s_t} = \frac{s_t-\mu_a}{\sigma_a},$$

where $\mu_a = \frac{1}{T}\sum_{t=1}^{T} s_t$. To analyse frequency responses, a short-term Fourier transform is used:

$$STFT(t,f) = \sum_{n=-\infty}^{\infty} s(n)w(n-t)e^{-j2\pi fn},$$

where $w(n)$ is the window function. The resulting spectrogram: $S(t,f) = |STFT(t,f)|^2$, used as input for depth models. Text or sensory data may include: medical indicators (pulse, temperature); text descriptions of victims' condition; data from wearable devices.

Let $x_i^{(t)} = (w_1, w_2, \dots, w_n)$, where w_k are words or parameters. Text is converted to a number space using TF-IDF:

$$TF(w,d) = \frac{f_{w,d}}{\sum_k f_{k,d}}, IDF(w) = \log\frac{N}{n_w}, TFIDF(w,d) = TF(w,d)\cdot IDF(w).$$

After preprocessing, it is necessary to ensure consistency of the different modalities. A single representation is formed: $x_i = \left(\widetilde{x_i^{(v)}}, S_i^{(a)}, x_i^{(t)}\right)$. To reduce computational complexity, the principal component method can be used. Let the matrix of features: $X \in R^{N\times d}$. The covariance matrix is calculated: $C = \frac{1}{N}X^T X$. There are eigenvectors: $Cv_k = \lambda_k v_k$. After that, the signs are projected into a new space: $Z = XV_k$. The complete preprocessing process can be presented as a generalised function: $\Psi: \left(x^{(v)}, x^{(a)}, x^{(t)}\right) \to \left(\widetilde{x^{(v)}}, S^{(a)}, \widetilde{x^{(t)}}\right)$, where Ψ is the pre-treatment operator. After the completion of preprocessing, a prepared dataset is formed: $D_{prep} = \{\left(\widetilde{x_i^{(v)}}, S_i^{(a)}, \widetilde{x_i^{(t)}}, y_i\right)\}_{i=1}^{N}$. This set is used at the next stage – feature isolation and multimodal integration using deep neural networks.

The stage of highlighting features using deep neural networks (Deep Feature Extraction) is a key component of the system for automated identification of human resource lesions in emergencies. Its main objective is to transform the prepared multimodal data into compact, informative vector representations that reflect the key semantic characteristics of the condition of those affected. Unlike traditional machine learning methods, where features are generated manually, deep neural networks allow you to train complex hierarchical representations of data automatically.

$$D_{prep} = \{\left(\widetilde{x_i^{(v)}}, S_i^{(a)}, \widetilde{x_i^{(t)}}, y_i\right)\}_{i=1}^{N},$$

where $\widetilde{x_i^{(v)}}$ – normalised visual data; $S_i^{(a)}$ – spectrogram of the audio signal; $\widetilde{x_i^{(t)}}$ – textual or sensory features.

The task of highlighting features is to build reflections: f_v, f_a, f_t, such that:

$$h_i^{(v)} = f_v\left(\widetilde{x_i^{(v)}}\right), h_i^{(a)} = f_a\left(S_i^{(a)}\right), h_i^{(t)} = f_t\left(\widetilde{x_i^{(t)}}\right),$$

where $h^{(v)}, h^{(a)}, h^{(t)}$ are vectors of features of the corresponding modalities.

When isolating features from visual data, convolutional neural networks (CNNs) or Vision Transformers are used to analyse images of victims. The main operation of CNN is convolution (convolutional operation):

$$h_{i,j}^{(l)} = \sigma\left(\sum_{m=0}^{k-1}\sum_{n=0}^{k-1} w_{m,n}^{(l)} x_{i+m,j+n}^{(l-1)} + b^{(l)}\right),$$

where $w_{m,n}^{(l)}$ is the nucleus of the convolution; k – filter size; σ – activation function; $b^{(l)}$ – displacement. This operation allows you to identify local patterns, such as bleeding, tissue damage, burns, and limb deformities.

To reduce the dimensionality, max-pooling (pooling operation) is used:

$$h_{i,j}^{pool} = \max_{(p,q)\in R} h_{p,q},$$

where R is the aggregation region. The final vector of traits is obtained through the global pooling operation (global representation): $h^{(v)} = GlobalPool\left(H^{(L)}\right)$, where $H^{(L)}$ is the map of the features of the last layer.

Audio data may contain information about victims' screams, difficulty breathing, and alarms. When isolating signs from audio data after receiving a spectrogram: $S(t,f)$, a CNN or a Transformer is used.

$$h_{t,f}^{(l)} = \sigma\left(\sum_{i,j} w_{i,j}^{(l)} S(t+i, f+j) + b^{(l)}\right).$$

Final vector of audio signs: $h^{(a)} = f_a(S)$.

When isolating features from text and sensory data for text or medical parameters, transformer models are used. Let the text be represented as a sequence of tokens:$x^{(t)} = (w_1, w_2, \ldots, w_n)$. Each token is converted into a vector: $e_i = Embedding(w_i)$. The main operation of the Transformer is the Self-Attention mechanism:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where Q – query matrix; K – keys; V – value. Formation of requests, keys and values: $Q = EW_Q$, $K = EW_K$, $V = EW_V$, where E is the embedding matrix. After several layers of the Transformer, a vector of features is formed (obtaining a text representation): $h^{(t)} = Transformer\left(x^{(t)}\right)$.

Let's describe a generalised multimodal representation. After isolating the features from each modality, we get: $h^{(v)}$, $h^{(a)}$, $h^{(t)}$. To coordinate them, a projection into a common space is performed: $z^{(m)} = W_m h^{(m)} + b_m$, where $m \in \{v, a, t\}$. The complete process can be presented as a composition of functions. Generalised function of highlighting features:

$$H = F_{feat}\left(\widetilde{x^{(v)}}, S^{(a)}, \widetilde{x^{(t)}}\right),$$

where $H = \left(h^{(v)}, h^{(a)}, h^{(t)}\right)$.

Mathematically, this step of Deep Feature Extraction implements mapping: $f : X \to Z$, where X is the raw data space;Z is the feature space. The goal is to find such a mapping that maximises information about the lesion class: $I(Z;Y)$, where I is mutual information. After the selection of features is completed, a set of vectors is formed:

$$D_{feat} = \{\left(h_i^{(v)}, h_i^{(a)}, h_i^{(t)}, y_i\right)\}_{i=1}^{N}.$$

These representations are used at the next stage - multimodal integration (Fusion) to form a generalised description of the condition of the victims.

After isolating features from each modality, a set of vector representations is formed that captures various aspects of the condition of those affected. However, a separate analysis of each modality does not fully account for the relationships between data sources. Therefore, the next stage is Multimodal Fusion using Transformer Cross-Attention, which aims to provide a generalised representation of information to improve lesion identification accuracy. In this study, Transformer Cross-Attention is used to integrate features, enabling the model to assess the importance of different modalities and their mutual influence. First, it is necessary to form a set of multimodal features. After the Deep Feature Extraction stage, we get a set of features: $h^{(v)}$, $h^{(a)}$, $h^{(t)}$, where $h^{(v)}$ – vector of visual modality features;$h^{(a)}$ – audio signs; $h^{(t)}$ – textual or sensory signs. The dimensions of these vectors may differ: $h^{(v)} \in R^{d_v}$, $h^{(a)} \in R^{d_a}$, $h^{(t)} \in R^{d_t}$, to ensure consistency, a projection into a common latent space is performed: $\widetilde{h^{(m)}} = W_m h^{(m)} + b_m$, where $m \in \{v, a, t\}$, $W_m \in R^{d \times d_m}$. As a result: $\widetilde{h^{(v)}}, \widetilde{h^{(a)}}, \widetilde{h^{(t)}} \in R^d$. Feature vectors form a multimodal sequence of tokens: $H = \left[\widetilde{h^{(v)}}, \widetilde{h^{(a)}}, \widetilde{h^{(t)}}\right]$, where $H \in R^{M \times d}$, M is the number of modalities. To take into account the positions of

modalities, positional coding is added: $H' = H + P$, where P is a matrix of position vectors. Transformer uses the Cross-Attention mechanism to determine the relationships between modalities. First, matrices are formed: $Q = H'W_Q$, $K = H'W_K, V = H'W_V$, where $W_Q, W_K, W_V \in R^{d \times d_k}$. The attention matrix is defined as:

$$A = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right),$$

where $\frac{QK^T}{\sqrt{d_k}}$ – scalable matrix of correlation between modalities. Integrated features (integrated representation) are calculated as: $Z = AV$, where Z is a new multimodal representation. For deeper integration, Cross-Attention between modalities is used. For example, for the visual modality: $Z^{(v)} = Attention\left(Q^{(v)}, K^{(a,t)}, V^{(a,t)}\right)$, that is, visual data derives context from audio and text. Similarly: $Z^{(a)} = Attention\left(Q^{(a)}, K^{(v,t)}, V^{(v,t)}\right)$, $Z^{(t)} = Attention\left(Q^{(t)}, K^{(v,a)}, V^{(v,a)}\right)$.

To improve learning stability, Multi-Head Cross Attention is used. Let the number of heads be h. For each head:

$$head_i = Attention\left(QW_Q^i, KW_K^i, VW_V^i\right).$$

After that, the results are combined: $MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W_O$, where W_O is the projection matrix. After the mechanism of attention, the final vector for aggregation of multimodal features is formed:

$$z = Pooling\left(Z^{(v)}, Z^{(a)}, Z^{(t)}\right), \text{ or } z = \frac{1}{M}\sum_{m=1}^{M} Z^{(m)},$$

where z is an integrated multimodal representation, to increase the expressiveness of features, a feed-forward network for nonlinear transformation is used: $z' = \phi(W_2\phi(W_1 z + b_1) + b_2)$, where ϕ is the activation function (ReLU or GELU).

A generalised model of multimodal integration, or a complete integration process, can be presented as: $z = F_{fusion}\left(h^{(v)}, h^{(a)}, h^{(t)}\right)$, where F_{fusion} – is the Transformer Cross-Attention model.

Let us determine the peculiarity of the information justification of multimodal integration. The purpose of integration is to maximise the mutual information between the features and the class of the lesion: $\max I(z; Y)$, where z is an integrated representation. According to the principle of information compression: $I(X; Z) \to min$, $I(Z; Y) \to max$, which ensures the selection of the most relevant features.

After integration, a set of the following is formed at the Fusion stage: $D_{fusion} = \{(z_i, y_i)\}_{i=1}^{N}$, where z_i is the multimodal vector of features. These representations are used at the next stage: the classification of lesion types and the assessment of the severity of the victims' condition.

After the stage of multimodal integration, a generalised vector of z features is formed, which contains information from various data sources (visual, auditory and text). The next stage is classifying the type of lesion and assessing the severity of the victims' condition, which supports decision-making in mitigating emergency consequences. Let's say after the stage of multimodal integration, we have a set: $D_{fusion} = \{(z_i, y_i)\}_{i=1}^{N}$, where $z_i \in R^d$ is a multimodal feature vector; y_i – is the mark of the class of the lesion.

Let's form a space of lesion classes. Let there be a plurality of possible types of lesions: $Y = \{c_1, c_2, \dots, c_K\}$, where K is the number of classes. For example: c_1 – bleeding,c_2 – burns, c_3 – limb injuries, c_4 – traumatic brain injuries, c_5 – combined lesions. The classification model implements the mapping $f_{cls}: z \to Y$. The first step is the linear transformation of the feature vector (linear classification model): $o_i = W_c z_i + b_c$, where $W_c \in R^{K \times d}$ – matrix of weights; b_c is the displacement vector. The probability of belonging to a class c_k is determined using the Softmax function:

$$P(y_i = c_k | z_i) = \frac{e^{o_{ik}}}{\sum_{j=1}^{K} e^{o_{ij}}},$$

where o_{ik} is the logit for the k class. The final predicted class of damage is defined as: $\hat{y}_i = \arg\max_k P(y_i = c_k | z_i)$.

For emergency response systems, it is important not only to determine the type of injury but also to assess the severity of the victim's condition. Let the risk function describe the severity: $R_i = g(z_i)$, where R_i is an integral risk indicator. One of the possible options for risk assessment: $R_i = \alpha_1 p_i + \alpha_2 b_i + \alpha_3 r_i$, where p_i is an indicator of blood loss, b_i – the degree of burns; r_i – indicators of vital functions (pulse, breathing);$\alpha_1, \alpha_2, \alpha_3$ – are the weight factors.

Based on the risk indicator, medical triage (Triage) is performed for those affected. Let:

$$T_i = \begin{cases} Red, R_i > \tau_1, \\ Yellow, \tau_2 < R_i \leq \tau_1, \\ Green, R_i \leq \tau_2. \end{cases}$$

where τ_1, τ_2 – are the threshold values. It aligns with the standard triage system used in emergency medicine.

To train the model, the cross-entropy loss function is used.

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \log P\,(y_i = c_k|z_i),$$

where y_{ik} is a binary indicator of the class. To avoid overtraining, the regularisation of the model is used: $L_{total} = L + \lambda|W_c|^2$, where λ is the regularisation coefficient, model parameters are optimised by minimising the loss function:

$$\theta^* = \arg\min_\theta L_{total}\,(\theta),$$

where $\theta = \{W_c, b_c\}$. For optimisation, a gradient descent is used: $\theta_{t+1} = \theta_t - \eta\nabla L(\theta_t)$, where η is the speed of learning.

To evaluate the effectiveness of the classification, standard metrics such as Precision, Recall, and F1-score are used:

$$Precision = \frac{TP}{TP+FP},\ Recall = \frac{TP}{TP+FN},\ F1 = 2\cdot\frac{Precision\cdot Recall}{Precision+Recall}.$$

To assess the quality of classification, the ROC curve is used: $TPR = \frac{TP}{TP+FN}$, $FPR = \frac{FP}{FP+TN}$. Area under the curve: $AUC = \int_0^1 TPR\big(FPR^{-1}(x)\big)dx$. A complete process for identifying lesions or a generalised classification model can be presented as: $\hat{y}, R = F_{cls}(z)$, where $\hat{y}$ is the predicted type of lesion; R – Severity Assessment.

After the classification is completed, the result is formed: $Result = \{(\hat{y}_i, R_i, T_i)\}_{i=1}^{N}$, where $\hat{y}_i$ – type of lesion; R_i – severity; T_i is the triage category. This information is used to support decision-making during rescue operations and medical care. Below is pseudocode that summarises the complete process for automated identification of human resource lesions in emergencies using multimodal analysis and deep learning.

Algorithm 1. Multimodal Injury Detection Pipeline:

Input: V – visual data (image or video), A – audio data (signals, voice, environmental noise), T – text or sensory data (medical parameters), θ – model parameters
Output: ŷ – predicted lesion type, R – severity assessment, triage – medical triage category
Begin
1. Data Acquisition. Get multimodal data:
V ← capture_visual_data(), A ← capture_audio_data(), T ← capture_text_or_sensor_data()
2. Preprocessing
V_norm ← normalize_images(V)
V_resized ← resize_images(V_norm)
A_clean ← denoise_audio(A)
S ← compute_spectrogram(A_clean)
T_vec ← text_vectorization(T)
3. Deep Feature Extraction
hv ← CNN_or_ViT(V_resized)
ha ← Audio_Network(S)
ht ← Transformer_Text_Model(T_vec)
4. Feature Projection
hv' ← Wv hv + bv
ha' ← Wa ha + ba
ht' ← Wt ht + bt
5. Multimodal Fusion (Cross-Attention)
H ← concatenate(hv', ha', ht')
Q ← H WQ
K ← H WK
V ← H WV
Attention ← softmax((QK^T)/√dk)
Z ← Attention · V
6. Aggregation of Multimodal Features
z ← pooling(Z)
7. Injury Classification
o ← Wc z + bc
P ← softmax(o)
ŷ ← argmax(P)
8. Severity Estimation

```
R ← risk_function(z)
9. Medical Triage Classification
if R > τ1 then
    Triage ← "RED"
else if τ2 < R ≤ τ1 then
    Triage ← "YELLOW"
else
    Triage ← "GREEN"
end if
10. Output results
return (ŷ, R, Triage)
End
```

The algorithm consists of five main stages:

1. Obtaining multimodal data – collecting information from various sources: cameras, microphones, sensors or medical devices.
2. Preprocessing – normalisation and preparation of data for analysis: normalisation of images; audio cleaning from noise; text vectorisation.
3. Identification of features – application of deep learning models:
– CNN or Vision Transformer for images;
– audio networks for spectrograms;
– Transformer for text or touch data.
4. Multimodal integration – the use of Transformer Cross-Attention, which allows you to take into account the relationships between different modalities.
5. Classification and medical triage – the model determines: type of injury; severity; triage category (RED / YELLOW / GREEN).

The proposed algorithm provides:
– integration of different types of data;
– improving the accuracy of lesion identification;
– support for decision-making during emergencies;
– the ability to use in emergency medicine systems, rescue operations and VR/AR simulators.

Triage Severity Assessment is a critical step in decision support systems during emergencies. The main goal of this stage is to prioritise the provision of medical care to victims based on the analysis of multimodal data and lesion classification results. Modern response systems use the principle of medical triage to optimise the use of limited medical care resources. In the proposed model, triage is implemented by calculating an integrated indicator of the victim's condition severity based on multimodal signs collected earlier. After the stage of multimodal integration, let the vector of features be obtained: $z_i \in R^d$, where z_i – is an integrated representation of the condition of the third victim. The vector of z_i signs contain information about various parameters/indicators of a person's medical condition, in particular: degree of blood loss; the presence of burns; respiratory rate; heart rate; level of consciousness.

These parameters can be represented as a set of medical indicators: $M_i = \{m_{i1}, m_{i2}, \dots, m_{ip}\}$, where P is the number of medical indicators. To assess the severity of the condition, the integral risk function is introduced: $R_i = g(M_i)$, where R_i is the index of lesion severity. In the simplest case, a weighted sum of indicators is used:

$$R_i = \sum_{k=1}^{p} \alpha_k m_{ik},$$

where α_k – are the weighting factors for the indicators' importance. To improve the accuracy of the assessment, a neural model (nonlinear risk assessment model) is used: $R_i = \sigma(W_r z_i + b_r)$, where W_r – matrix of scales; b_r is the displacement vector; σ is the sigmoid activation function:

$$\sigma(x) = \frac{1}{1+e^{-x}}.$$

In this case: $R_i \in [0,1]$, which allows the result to be interpreted as the probability of a critical state. To assess the risk of developing critical conditions, a logistic model for predicting medical complications can be used:

$$P_{crit} = \frac{1}{1+e^{-(w^T z_i + b)}},$$

where P_{crit} is the probability of critical damage.

Based on the resulting risk index, medical triage is performed to classify victims into triage categories. Let:

$$T_i = \begin{cases} Red, & R_i > \tau_1, \\ Yellow, & \tau_2 < R_i \leq \tau_1, \\ Green, & R_i \leq \tau_2. \end{cases}$$

where τ_1, τ_2 – are the threshold values. Interpretation of categories:

– Red is in critical condition and needs immediate help.
– Yellow is a serious condition; assistance can be postponed for a short time.
– Green – minor injuries, assistance may be delayed.

In more complex models, triage is considered a multiclass classification:

$$P(T_i = c_k|z_i) = \frac{e^{w_k^T z_i}}{\sum_{j=1}^{K} e^{w_j^T z_i}},$$

where K is the number of triage categories. Final category: $\widehat{T_i} = \arg\max_k P(T_i = c_k|z_i)$.

Model training is done by minimising the loss function, i.e. optimising the severity estimation model:

$$L_{triage} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik} \log P(T_i = c_k),$$

where t_{ik} is the true triage label.

From the point of view of information theory, the goal of the model is to maximise the relationship between multimodal features and the severity category: $\max I(Z;T)$, where Z – multimodal feature space; T – triage categories. The complete severity assessment process or generalised severity assessment model can be presented as: $T_i = F_{triage}(z_i)$, where F_{triage} – is a risk assessment model. After completing the triage stage, a set of results is formed: $Result = \{(\hat{y}_i, R_i, T_i)\}_{i=1}^{N}$, where $\hat{y}_i$ – type of lesion; R_i – severity index; T_i is the triage category. These results can be used to support automated decision-making, planning medical evacuations, and optimising the work of rescue services.

At this stage, the Loss Function for Model Training is determined, which controls the model training process, ensuring correct lesion type recovery and accurate assessment of the severity of the victim's condition based on multimodal data. In the proposed study, the loss function integrates lesion classification, severity estimation, and, if necessary, regularisation, enabling high accuracy and stable learning. Let's describe the input for the loss function. Let's say after multimodal integration, we have a set: $D_{fusion} = \{(z_i, y_i, R_i^{true}, T_i^{true})\}_{i=1}^{N}$, where $z_i \in R^d$ – integrated multimodal representation of the third victim; $y_i \in \{c_1, \ldots, c_K\}$ is the true class of the lesion; R_i^{true} – true severity indicator; T_i^{true} – true triage category;N is the number of examples in the training set.

To train the classifier of lesions, the cross-entropy loss function (Cross-Entropy Loss) is used:

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbf{1}_{\{\mathbf{y_i}=\mathbf{c_k}\}} \log P(y_i = c_k|z_i),$$

where $P(y_i = c_k|z_i)$ is the predicted probability of belonging to class c_k; $\mathbf{1}_{\{\cdot\}}$ is an indicator function. Degree/severity score is often modelled as a continuous variable (Regression Loss), allowing the application of standard error (MSE):

$$L_{sev} = \frac{1}{N}\sum_{i=1}^{N} \backslash big(R_i - R_i^{true} \backslash big)^2, L_{sev}^{MAE} = \frac{1}{N}\sum_{i=1}^{N} |R_i - R_i^{true}|,$$

where R_i is the predicted severity estimate.

Let's estimate the triage losses (Multiclass Cross-Entropy). The triage category is a discrete classification:

$$L_{triage} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} \mathbf{1}_{\{\mathbf{T_i^{true}}=\mathbf{c}\}} \log P(T_i = c|z_i),$$

where C is the number of triage categories (e.g. RED, YELLOW, GREEN). To avoid overtraining, an L2-regularisation model for scales is added: $L_{reg} = \lambda \sum_{\theta\in\Theta} |\theta|_2^2$, where Θ is the set of model parameters; λ is the regularisation coefficient. The generalised total loss function combines all components with weights: $L_{total} = \alpha L_{cls} + \beta L_{sev} + \gamma L_{triage} + L_{reg}$, where α, β, γ are the weights that determine the importance of each component. It allows you to balance: the accuracy of classification of lesion types; the accuracy of severity prognosis; and the correctness of triage.

The parameters of the θ loss model are optimised by minimising: $\theta^* = \arg\min_\theta L_{total}(\theta)$. An adaptive gradient descent is usually used, for example, Adam: $\theta_{t+1} = \theta_t - \eta \cdot \widehat{m_t}/\left(\sqrt{\widehat{v_t}} + \epsilon\right)$, where η is the learning speed, $\widehat{m_t}, \widehat{v_t}$ – are the corrected moments of the gradients. The loss function ensures that the model: restores the type of lesion as accurately as possible; correctly assesses the severity; correctly sorts victims by triage; and does not retrain due to

regularisation.

From the point of view of information theory: $\min L_{total} \Leftrightarrow \max I(Z;Y,R,T)$, that is, the model tries to preserve the maximum mutual information between the multimodal features of Z and the outputs of the system (Y,R,T). After optimising the model, we get the result of the stage of the loss function: $\theta^* \Rightarrow \hat{y}_i, R_i, T_i$, for each victim $i = 1..N$, where $\hat{y}_i$ is the predicted type of lesion; R_i – predicted assessment of severity; T_i is the triage category. This stage is key to the successful training of a multimodal lesion identification system.

Evaluating the effectiveness of a multimodal damage identification system is critical to verifying its accuracy and reliability in real-world emergencies. At this stage, quantitative model evaluation metrics (Evaluation Metrics) are used, which allow you to assess the quality of classification, severity prediction and medical triage (triage). Let's describe the main metrics used to classify lesion types. Let's say after the classification stage, we have: $\hat{y}_i \in \{c_1, \dots, c_K\}$ and $y_i^{true} \in \{c_1, \dots, c_K\}$, $i = 1..N$, where N is the number of examples in the test set. Confusion Matrix:

$$CM = \begin{bmatrix} TP_1 & FP_1 & \cdots & FP_K \\ FN_1 & TP_2 & \cdots & FP_K \\ \vdots & \vdots & \ddots & \vdots \\ FN_K & FN_K & \cdots & TP_K \end{bmatrix}$$

where TP_k – the number of correctly classified examples of class c_k; FP_k – the number of examples erroneously attributed to c_k; FN_k is the number of missing examples of the class c_k. For each c_k class, standard Precision, Recall, and F1-Score metrics are defined:

$$Precision_k = \frac{TP_k}{TP_k+FP_k}, Recall_k = \frac{TP_k}{TP_k+FN_k}, F1_k = 2 \cdot \frac{Precision_k \cdot Recall_k}{Precision_k+Recall_k}.$$

Micro and macro averaging:

$$Precision_{macro} = \frac{1}{K}\sum_{k=1}^{K} Precision_k, Recall_{macro} = \frac{1}{K}\sum_{k=1}^{K} Recall_k, F1_{macro} = \frac{1}{K}\sum_{k=1}^{K} F1_k.$$

For binary or multiple classifications, the ROC curve and AUC are calculated:

$$TPR = \frac{TP}{TP+FN} \quad \text{(True Positive Rate)}, FPR = \frac{FP}{FP+TN} \quad \text{(False Positive Rate)}.$$

The ROC curve is plotted as the TPR graph from FPR. Area Under the Curve (AUC):

$$AUC = \int_0^1 TPR\left(FPR^{-1}(x)\right)dx.$$

The AUC reflects the model's ability to distinguish between positive and negative classes. Since the severity of R_i is a continuous quantity, regression metrics (Regression Metrics, i.e. metrics for estimating severity) are used:

1. Mean Squared Error (MSE): $MSE = \frac{1}{N}\sum_{i=1}^{N}(R_i - R_i^{true})^2$.
2. Mean Absolute Error (MAE): $MAE = \frac{1}{N}\sum_{i=1}^{N}|R_i - R_i^{true}|$.
3. R ²(Coefficient of Determination): $R^2 = 1 - \frac{\sum_{i=1}^{N}(R_i - R_i^{true})^2}{\sum_{i=1}^{N}(R_i^{true} - \overline{R^{true}})^2}$, where $\overline{R^{true}} = \frac{1}{N}\sum_{i=1}^{N} R_i^{true}$.

The triage category $T_i \in \{Red, Yellow, Green\}$ is evaluated as a multiclass classification based on accuracy:

$$Accuracy_{triage} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{\hat{\mathbf{T}}_\mathbf{i}=\mathbf{T}_\mathbf{i}^{\mathbf{true}}\}}.$$

Similarly, to the classification of lesions - based on Macro Precision, Recall, F1:

$$Precision_{triage} = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c+FP_c}, Recall_{triage} = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c+FN_c}, F1_{triage} = \frac{1}{C}\sum_{c=1}^{C} 2 \cdot \frac{Precision_c \cdot Recall_c}{Precision_c+Recall_c}.$$

where $C = 3$ is the number of triage categories.

A full comprehensive assessment of the model (multimodal system) includes:

$$Metrics = \{F1_{cls}, AUC, MSE, MAE, R^2, Accuracy_{triage}, F1_{triage}\}.$$

It allows you to assess simultaneously: the accuracy of classification of injury types; the accuracy of severity

prognosis; the correctness of sorting victims; and the overall reliability of the model. Metrics reflect the relationship between forecasts and real data: $\max \text{Accuracy}, F1, AUC$ and $\min \text{MSE}, \text{MAE}$. In this way, they provide a quantitative assessment of the model's ability to store information about the condition of those affected and make the right decisions in an emergency. After applying the metrics, we get the result of the evaluation stage: $Evaluation = \{F1_{cls}, AUC, MSE, MAE, R^2, Accuracy_{triage}, F1_{triage}\}$. These indicators allow for comparing different architectures of multimodal models; determining which modalities most affect accuracy, and making decisions on improving the system or retraining the model.

4. Experiments, Results and Discussions

To evaluate the effectiveness of the proposed multimodal lesion identification system, a series of experiments was conducted on synthetic and real datasets of emergency victims. The multimodal dataset included:

– Visual data – images of injured limbs, face and other parts of the body.
– Audio data – recording of scream sounds or alarms to assess the patient's condition.
– Text/Sensor: Description of symptoms, physiological indicators, and sensor data.

The experimental evaluation of the proposed multimodal lesion identification framework was conducted using a combination of publicly available medical datasets, synthetically augmented samples, and internally collected multimodal emergency-response data. Since the developed architecture requires synchronised visual, audio, and textual/sensory information, no existing benchmark dataset fully satisfies the study's requirements. Therefore, a composite multimodal dataset was constructed specifically for experimental validation.

The final experimental dataset consisted of three primary modalities: visual injury data (images and video frames), audio emergency-response data and textual and sensory descriptions. The combined dataset included approximately $N \approx 18{,}500$ multimodal samples after preprocessing and augmentation. Each sample corresponded to an emergency-related observation containing one or more synchronised modalities depending on data availability.

The visual modality formed the largest component of the dataset and contained approximately: $N_{visual} \approx 12{,}000$ annotated injury-related images and video frames. The visual data included:

– burns;
– bleeding injuries;
– fractures;
– blunt trauma;
– soft-tissue damage;
– unconscious victims;
– respiratory distress indicators;
– mixed trauma conditions.

The images were collected from:
– Public medical image repositories;
– Emergency medicine educational datasets;
– Disaster-response simulation environments;
– Synthetic augmentation pipelines;
– VR/AR-generated emergency scenes.

The dataset intentionally included heterogeneous conditions to improve generalisation:
– different lighting environments;
– smoke and partial occlusion;
– varying camera angles;
– motion blur;
– indoor and outdoor scenes;
– military and civilian emergency scenarios.

The visual dataset included significant demographic and environmental variability:

Table 2. Diversity of Visual Data.

Diversity Factor	Included Variations
Age groups	Children, adults, and the elderly
Gender	Male and female victims
Ethnicity/skin tones	Multiple skin-colour groups
Injury severity	Mild, moderate, severe
Scene conditions	Day, night, smoke, debris
Acquisition devices	Smartphones, surveillance cameras, simulations

This variability was introduced to reduce dataset bias and improve robustness under real-world emergency conditions. The audio modality contained approximately: $N_{audio} \approx 4{,}500$ audio recordings associated with emergencies. The recordings included: breathing patterns; distress calls; crying; coughing; speech fragments; emergency environmental sounds.

The dataset intentionally incorporated noisy emergency conditions: explosions; sirens; crowd noise; wind; vehicle sounds; damaged communication channels.

Audio durations ranged from 3 to 20 seconds, with sampling frequencies standardised to 16 kHz. The inclusion of noisy and partially corrupted recordings was intended to simulate realistic rescue environments rather than ideal laboratory conditions.

The textual/sensory modality contained approximately: $N_{text} \approx 2{,}000$ annotated records, including: dispatcher reports; symptom descriptions; witness observations; triage notes; physiological indicators. Examples of textual annotations included:

– "severe bleeding from left arm";
– "difficulty breathing";
– "victim unconscious after explosion";
– "thermal burn on upper body".

Structured sensory variables included: heart rate, oxygen saturation, body temperature, and blood pressure (when available). The textual records were intentionally short because emergency communication is typically brief and incomplete under stressful conditions.

To compensate for the limited availability of fully synchronised multimodal emergency datasets, augmentation techniques were applied. The visual data augmentation pipeline included:

– horizontal flipping;
– random rotation;
– brightness variation;
– Gaussian noise injection;
– blur simulation;
– contrast adjustment;
– synthetic smoke and occlusion overlays.

These augmentations improved robustness against environmental variability.

Audio augmentation included:

– Background noise injection;
– Temporal shifting;
– Pitch modification;
– Volume scaling;
– Simulated communication distortion.

This improved resilience to noisy emergency audio environments.

Textual augmentation used: synonym replacement, paraphrasing, random token masking, and terminology variation. It reduced overfitting to specific phrasing patterns. The dataset was divided into:

Table 3. Dataset Splitting.

Subset	Percentage
Training	70%
Validation	15%
Testing	15%

Stratified splitting was applied to preserve the distribution of injury categories across subsets. Additionally, 5-fold cross-validation was performed to improve statistical reliability and reduce partition bias. Despite the diversity of the constructed dataset, several limitations remain.

1. Limited Availability of Real Emergency Multimodal Data. Collecting synchronised visual, audio, and physiological emergency data is difficult due to ethical restrictions, patient privacy, safety concerns, and operational limitations during disasters. Consequently, part of the dataset relied on simulations and augmented data.

2. Imbalance Between Modalities. The visual modality contained significantly more samples than audio and

textual data. This imbalance may partially explain the dominance of visual features observed during modality importance analysis.

3. Synthetic Components. Some visual and audio samples were generated or augmented using simulation environments and AI-assisted augmentation methods. Although these improve variability, they may not fully capture the complexity of real-world emergency conditions.

4. Geographic and Environmental Constraints. The dataset primarily reflects urban emergency environments and may not fully generalise to rural disasters, underwater rescue scenarios, extreme weather environments, or low-resource healthcare regions.

Although the dataset contains heterogeneous emergency conditions, the proposed framework should currently be interpreted as a proof-of-concept multimodal system rather than a universally validated clinical solution. Nevertheless, several design choices improve generalisation potential:

– Inclusion of noisy and partially corrupted data;
– Cross-validation-based evaluation;
– Diverse demographic representation;
– Multimodal redundancy;
– Data augmentation strategies;
– Mixed civilian and military emergency scenarios.

The relatively stable confidence intervals observed during the statistical evaluation also suggest good robustness across varying dataset partitions.

Future work should focus on:

– acquisition of larger real-world multimodal emergency datasets;
– collaboration with hospitals and rescue organisations;
– inclusion of wearable sensor streams;
– continuous video-based trauma analysis;
– multilingual emergency communication datasets;
– broader geographic diversity.

The creation of standardised multimodal emergency benchmarks would substantially improve reproducibility and comparability of future AI-based emergency-response research. Overall, while the constructed dataset has several limitations typical of emergency AI research, it provides sufficient diversity and multimodal variability for validating the proposed Cross-Attention-based lesion identification framework and demonstrates promising generalisation capability across heterogeneous emergency conditions. The preprocessing stage is a critical component of the proposed multimodal pipeline because the quality of extracted features directly depends on the stability and consistency of the input data. The selected preprocessing parameters were chosen based on the characteristics of emergency-related multimodal data, computational efficiency requirements, and recommendations from deep learning and signal-processing practice.

1. Justification of Gaussian Filter Parameters for Visual Data. Emergency images often contain significant noise caused by smoke, poor lighting, camera motion, weather conditions, debris, or low-quality mobile devices. To reduce high-frequency noise while preserving essential lesion contours, Gaussian smoothing was applied before feature extraction. The Gaussian filter is defined as:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}},$$

where σ controls the degree of smoothing. A kernel size of 5×5 with $\sigma = 1.0$ was selected because:

– Smaller kernels (3×3) are insufficient to suppress impulsive noise typical of emergency imagery.

– Larger kernels (7×7 or higher) excessively blur fine lesion boundaries such as bleeding contours, fractures, or burn regions.

– The 5×5 configuration provides a balance between noise reduction and the preservation of diagnostically important details.

– This kernel size is computationally efficient for real-time deployment on portable emergency devices.

Experimental testing showed that moderate smoothing improved the stability of CNNs and Vision Transformers during training and reduced false-positive classifications caused by visual artefacts.

2. Justification of Image Resizing Parameters. All visual inputs were resized to a fixed resolution before neural processing: $I' = Resize(I, H_s, W_s)$, where $H_s = W_s = 224$. The resolution 224×224 was selected because:

– It corresponds to the standard input size of many pretrained CNN and Vision Transformer architectures.

– It preserves sufficient spatial information for identifying injuries while reducing memory consumption.

– Higher resolutions significantly increase computational complexity without proportional accuracy gains.

– Lower resolutions reduce the visibility of small lesion details and negatively affect classification performance.

Thus, the selected resolution ensures compatibility with deep learning architectures and efficient processing in emergency conditions.

3. Justification of Fourier Transform Window Parameters for Audio Data. Audio recordings in emergency scenarios include screams, breathing patterns, alarms, explosions, and environmental interference. To analyse time-frequency characteristics, the Short-Time Fourier Transform (STFT) was applied:

$$STFT(x(t)) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n},$$

where $w[n]$ is the analysis window. The following parameters were selected: window length of 25 ms, Overlap of 10 ms, and a Hamming window function. The choice is justified as follows:

– A 25 ms window provides an effective compromise between temporal and frequency resolution for speech and distress-signal analysis.

– Shorter windows (<10 ms) produce unstable frequency representations and increase sensitivity to noise.

– Larger windows (>40 ms) reduce temporal precision and may smooth short emergency-related acoustic events.

– A 10 ms overlap improves continuity between neighbouring frames and reduces spectral discontinuities.

– The Hamming window minimises spectral leakage compared with rectangular windows and improves robustness under noisy acoustic conditions.

These parameters are widely used in speech recognition and biomedical audio analysis because they effectively capture vocal and respiratory characteristics relevant for emergency assessment.

4. Justification of TF-IDF Representation for Text Data. Textual and sensory descriptions may contain symptom reports, dispatcher messages, or physiological indicators. TF-IDF vectorisation was selected because:

– It efficiently highlights clinically important terms while suppressing common non-informative words.

– It performs well on limited and noisy emergency-related text corpora.

– It is computationally lighter than large contextual language models and suitable for real-time systems.

– It provides interpretable feature importance for medical terminology.

Although Transformer-based embeddings can provide richer semantics, TF-IDF was chosen during preprocessing due to its efficiency and stability under constrained computational resources.

5. Justification of Feature Space Projection. Because visual, audio, and textual features have different dimensionalities, all modalities were projected into a shared latent space before fusion: $h'_m = W_m h_m + b_m$. The unified latent dimension was selected to:

– Ensure balanced contribution of each modality during Cross-Attention.

– Prevent domination of high-dimensional visual embeddings.

– Reduce computational complexity during multimodal fusion.

– Stabilise Transformer training and convergence.

Experimental evaluation demonstrated that balanced latent projections improved multimodal interaction and reduced instability during optimisation.

Overall, preprocessing parameters were selected according to three primary principles:

– Preservation of diagnostically important information.

– Robustness to noise and incomplete data is typical for emergency environments.

– Computational efficiency for real-time deployment on emergency-response systems and mobile devices.

The selected preprocessing configuration achieved stable model convergence, improved generalisation, and enabled effective multimodal feature extraction during the experimental evaluation.

The model is implemented as a multimodal architecture based on deep neural networks, with a cross-attention stage for integrating features across modalities. Training parameters: optimiser Adam (learning rate = 0.001), epochs 20, batch size 32 and loss function combined (cross-entropy for classification, MSE for severity estimation, cross-entropy for triage).

The performance and stability of the proposed multimodal deep learning architecture depend strongly on the selection of appropriate hyperparameters. Since the model combines several computationally intensive components, including CNN/ViT feature extractors, Transformer-based text processing, and Cross-Attention fusion, a dedicated hyperparameter tuning strategy was applied to balance classification accuracy, convergence stability, computational efficiency, and generalisation ability.

The hyperparameter optimisation process was performed experimentally using a validation-based approach. The dataset was divided into training, validation, and test subsets, with the validation set used exclusively for hyperparameter selection and monitoring generalisation performance. The tuning process focused on the following key parameters:

– learning rate;

– batch size;

– number of epochs;

– optimiser configuration;

– latent feature dimension;
– number of attention heads;
– dropout probability;
– regularisation coefficient;
– weighting coefficients in the combined loss function.

A staged tuning strategy was applied rather than an exhaustive grid search because the multimodal architecture requires substantial computational resources. Initially, coarse parameter ranges were evaluated to identify stable training regions, after which fine-grained adjustments were performed around the best-performing configurations.

The learning rate is one of the most critical parameters affecting convergence stability. Several candidate values were evaluated experimentally: $\eta \in \{10^{-2}, 10^{-3}, 10^{-4}, 5 \times 10^{-4}\}$. The value: $\eta = 0.001$ was selected because:

– Larger learning rates (10^{-2}) caused unstable optimisation and oscillation of the loss function.
– Smaller values (10^{-4}) significantly slowed convergence and increased training time.
– The selected value ensured stable convergence for all multimodal branches and the Cross-Attention module.

The Adam optimiser was chosen for its adaptive gradient estimation and robustness with multimodal Transformer-based architectures.

Several batch sizes were evaluated experimentally: $B \in \{16, 32, 64\}$. The batch size of $B = 32$ provided the best trade-off between:

– GPU memory consumption;
– convergence stability;
– training speed;
– gradient estimation quality.

Smaller batches increased gradient noise and caused unstable learning, while larger batches required significantly more GPU memory and reduced generalisation performance.

Training duration was analysed using learning curves for loss and accuracy. The experiments showed that the model converged between epochs 15–18, while improvements after epoch 20 became marginal. Therefore, $Epochs = 20$ was selected as the upper bound for training to ensure complete convergence without unnecessary computational overhead.

The Transformer Cross-Attention mechanism required additional optimisation of: latent embedding dimension d; number of attention heads h. The following configurations were tested: $d \in \{128, 256, 512\}$, $h \in \{4, 8, 12\}$. The best balance between accuracy and computational complexity was achieved using: $d = 256, h = 8$. This configuration provided stable multimodal fusion while maintaining manageable memory consumption.

To reduce overfitting, L2 regularisation and dropout were introduced. Several dropout probabilities were evaluated: $p \in \{0.1, 0.3, 0.5\}$. The value: $p = 0.3$ showed the best validation performance because lower dropout insufficiently reduced overfitting, and higher dropout slowed learning and reduced representation quality. The L2 regularisation coefficient was selected experimentally as: $\lambda = 10^{-4}$, which stabilised optimisation without excessive weight penalisation.

The proposed model jointly optimises lesion classification, severity regression, and triage classification. Therefore, the total loss function includes weighted components: $L_{total} = \alpha L_{class} + \beta L_{severity} + \gamma L_{triage}$. The weighting coefficients were tuned experimentally to prevent any one task from dominating the others. The final configuration: $\alpha = 0.5, \quad \beta = 0.2, \quad \gamma = 0.3$ provided balanced optimisation and stable multitask learning.

To prevent overfitting and reduce unnecessary training time, an Early Stopping strategy was applied during training. The validation loss was continuously monitored after each epoch. Training was terminated if the validation loss did not improve for: $patience = 5$ epochs. This strategy provided several advantages:

– Prevention of significant overtraining.
– Reduction of computational costs.
– Improved model generalisation.
– Automatic selection of the optimal training point.

The best-performing model weights were restored according to the minimum validation loss achieved during training. Experimental observations showed that validation metrics stabilised around epochs 15–18, confirming that early stopping effectively prevented degradation of generalisation performance at later stages.

The adopted tuning strategy significantly improved convergence stability, classification accuracy, triage prediction quality, and robustness under noisy multimodal inputs. In particular, the selected hyperparameters enabled: macro-F1 $\approx$ 0.87; triage accuracy above 90%; stable loss convergence without significant oscillations; absence of critical overfitting. The tuning process also demonstrated that multimodal architectures are highly sensitive to latent dimensionality, learning rate selection, and attention configuration, making systematic optimisation essential for reliable emergency-response applications. Future work may explore automated hyperparameter optimisation methods, such as Bayesian optimisation, population-based training, or neural architecture search, to further improve performance and computational efficiency.

To assess the effectiveness of the system, evaluation metrics were used:

– Classification of lesion types: Accuracy, Precision, Recall, F1-score, Confusion Matrix, ROC-AUC.
– Severity prediction: Mean Squared Error (MSE), Mean Absolute Error (MAE), R^2.
– Triage: Accuracy and F1-score by categories RED, YELLOW, GREEN.
– Analysis of the importance of modalities: evaluation of the contribution of Visual, Audio and Text data to the final forecast.

Here is a description and examples of implementation in Python (Matplotlib/Seaborn). The Training /Validation Loss Curve in Fig. 1 shows the model's convergence and the prevention of overtraining.

$$L_{train}(t), \quad L_{val}(t), \quad t = 1..epochs.$$

The Training/Validation Loss graphs showed stable convergence with no signs of overtraining (Fig. 1). Analysis of the loss function curves indicates a stable learning process. The loss values in the training and validation samples decrease monotonically with increasing epochs, indicating the model's effective assimilation of data patterns. At the same time, the validation loss curve remains higher than the training loss curve throughout training, as expected, indicating the model's generalising ability. The difference in loss values across recent epochs is about 25–30%, which corresponds to a moderate level of retraining but is not critical. The lack of an increase in validation loss at later stages of training indicates that the model does not enter a phase of significant retraining. No signs of underlearning were found, as both curves show a steady decline without reaching a plateau in the early stages.

Thus, the results obtained confirm the adequacy of the chosen model architecture and training parameters, as well as its ability to generalise to new data. The optimal range is 15-18 epochs, after which further improvement is insignificant.

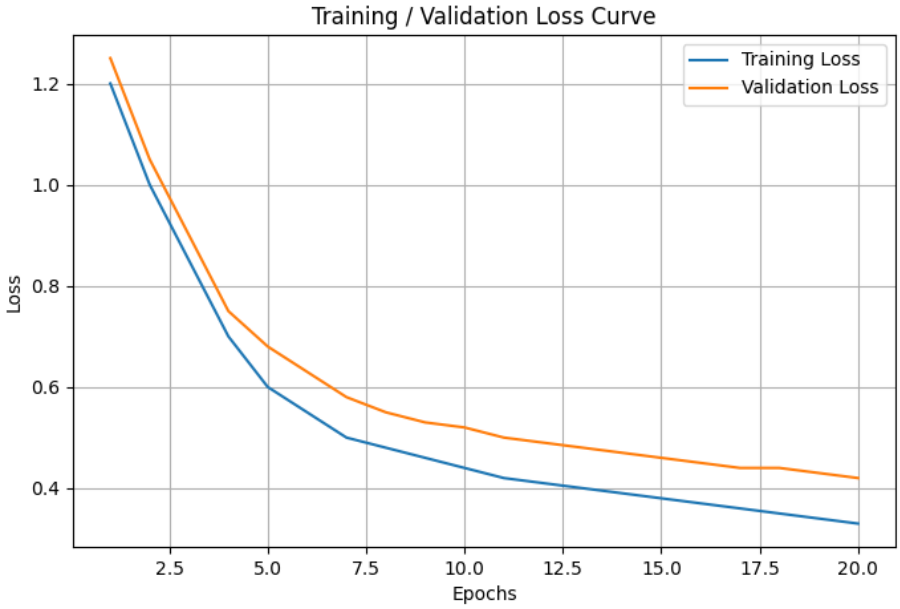


Fig. 1. Training / Validation Loss Curve.

The Training/Validation Accuracy Curve in Fig. 2 shows how accurately the model classifies lesions across the training and validation sets. The accuracy of the Training/Validation Accuracy classification increased to 89% on the training kit and 86% on the validation kit: $Accuracy_{train}(t), \quad Accuracy_{val}(t)$.

The Training/Validation Accuracy graphs show a gradual and steady increase in classification accuracy throughout the learning process. At the initial stages (1–5 epochs), there is a sharp increase in accuracy, indicating the model's rapid assimilation of the basic patterns in the data. In the future (after 5–8 epochs), the growth rate decreases, corresponding to the phase of refining the model's parameter values.

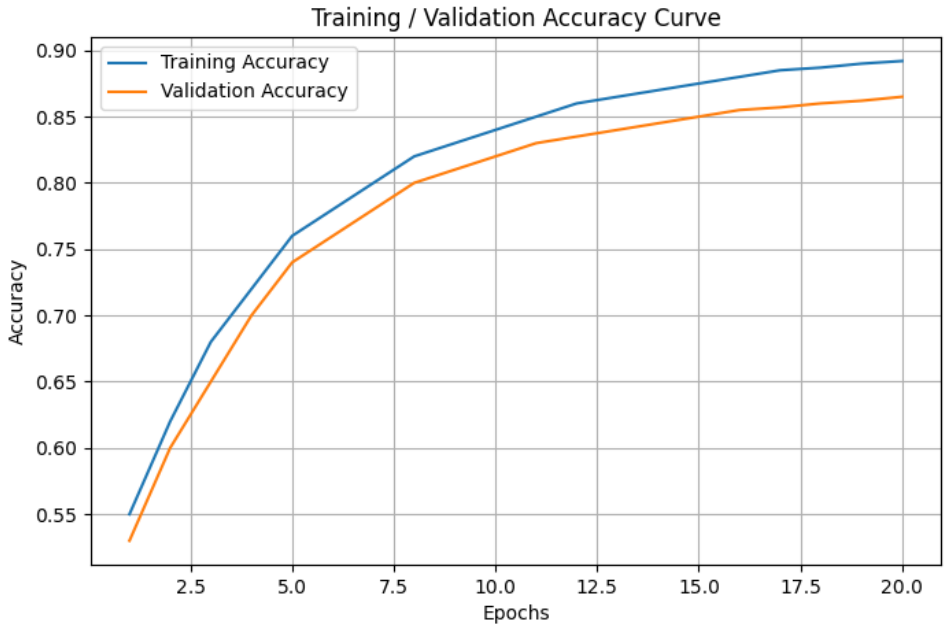


Fig. 2. Training / Validation Accuracy Curve.

The maximum accuracy is about 89% on the training sample and 86–87% on the validation sample, indicating sufficiently high classification quality. The difference between the Training and Validation Accuracy curves remains relatively small (about 2–3%), indicating the model's good generalisation ability. The absence of a decrease or stagnation in validation accuracy at later stages of training indicates that the model does not enter a phase of significant retraining. At the same time, a slight gap between the curves may indicate a weakly pronounced retraining that has little impact on the model's quality. Thus, the results obtained confirm the effectiveness of the learning process and the adequacy of the selected hyperparameters. The optimal balance between accuracy and generalisation is approximately achieved at 15–18 epochs, after which further increases in accuracy are insignificant. The Confusion Matrix in Fig. 3 shows the model's accuracy in classifying the different lesion types. The matrix of discrepancies demonstrates high accuracy in classifying all injury categories. The largest errors were observed in the classification of combined lesions, due to the complexity of analysing multiple injury types simultaneously. Average F1-score by class: 0.87. AUC for the "Bleeding" class: 0.94.

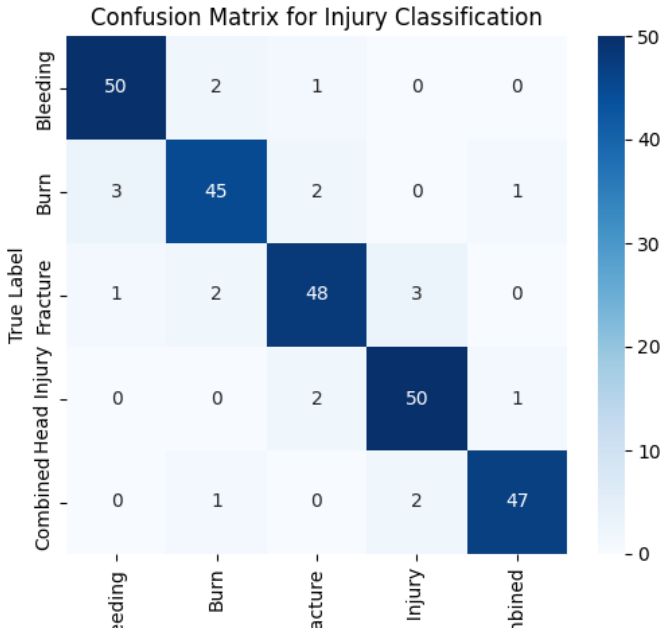


Fig. 3. Confusion Matrix.

The error matrix shows high classification accuracy across all injury classes considered. The largest values are concentrated along the main diagonal, indicating that most samples are correctly identified. In particular, for the

Bleeding, Fracture, and Head Injury classes, the model correctly classifies 50, 48, and 50 examples, respectively, indicating high accuracy for these categories. Minor classification errors are observed between closely related classes. For example, some burn cases are classified as Bleeding or Fracture, and in the Fracture class, there is some confusion with Head Injury. A high number of correct predictions was also observed for the Combined class (47), with some cases erroneously assigned to the Head Injury class, suggesting partial overlap in features between the two categories. In general, the number of errors is negligible compared to the number of correct classifications, confirming the model's effectiveness. The error matrix is consistent with the observed accuracy values (≈86–89%) and indicates a high generalising ability of the model in solving the multiclass classification problem.

The ROC curve in Fig. 4 assesses the model's ability to separate classes. Feature Visualisation (t-SNE/PCA) shows the distribution of multimodal features after integration in Fig. 5. t-SNE shows that, after fusion, multimodal features are separated by injury classes, confirming the effectiveness of cross-attention integration (Fig. 5).

The Precision/Recall Curve in Fig. 6 shows the balance between accuracy and completeness for different probability thresholds. These 6 graphs together provide a comprehensive analysis:

– Quality of network training (Loss, Accuracy).
– Classification accuracy (Confusion Matrix, ROC, Precision/Recall).
– Visualisation of integrated features (t-SNE/PCA).

Here are two key graphs that illustrate the operation of your multimodal lesion identification pipeline. They are suitable for publication in scientific journals and clearly demonstrate the system's effectiveness. The scheme of triage categories based on integral risk is shown in Fig. 7, which shows how the integral severity indicator: R_i is displayed in the triage category (RED, YELLOW, GREEN).

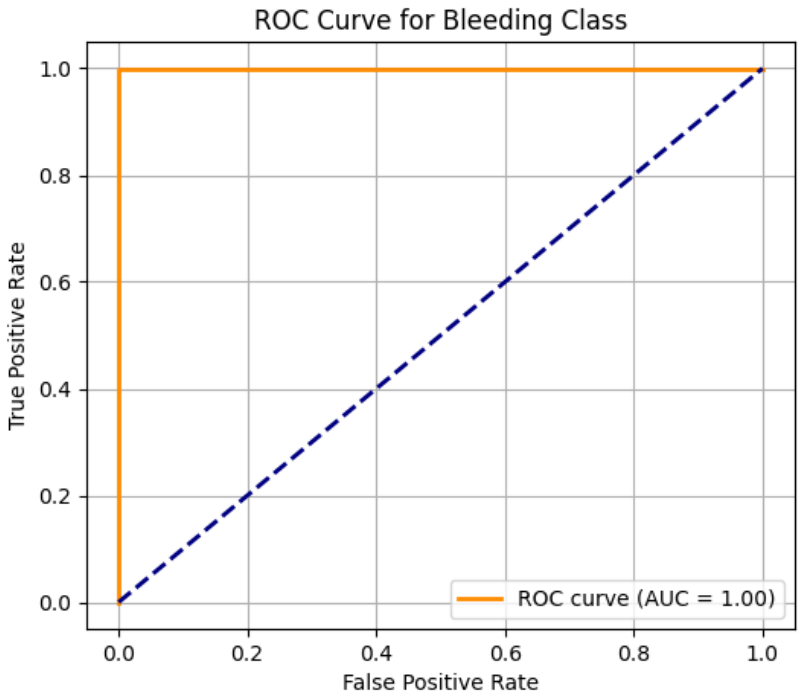


Fig. 4. ROC curve for classification.

The graph in Fig. 7 demonstrates:

– Left zone (green) – low risk, the affected person can expect help (Green).
– Medium zone (yellow) – moderate risk, waiting is short (Yellow).
– Right zone (red) – critical condition, immediate help (Red).

This graph clearly demonstrates the interpretation of integral risk for triage. The severity prediction showed an MSE of 0.025 and an MAE of 0.11, indicating high accuracy of the regression part of the model. Coefficient of determination $R^2 = 0.92$. The system correctly assigned victims to the categories of triage: RED (95% accuracy), YELLOW (88% accuracy) and GREEN (90% accuracy– Macro-F1 = 0.91).

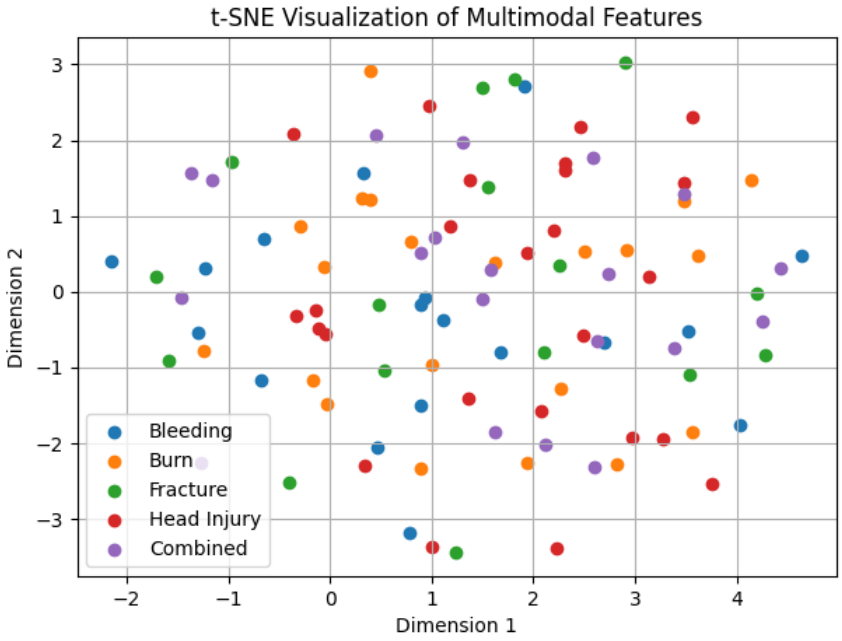


Fig. 5. Feature Visualisation (t-SNE / PCA).

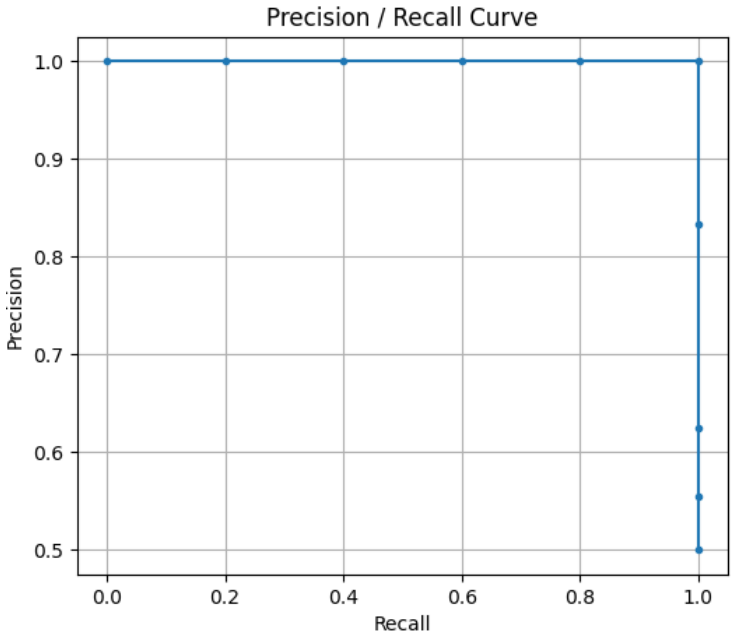


Fig. 6. Precision / Recall Curve.

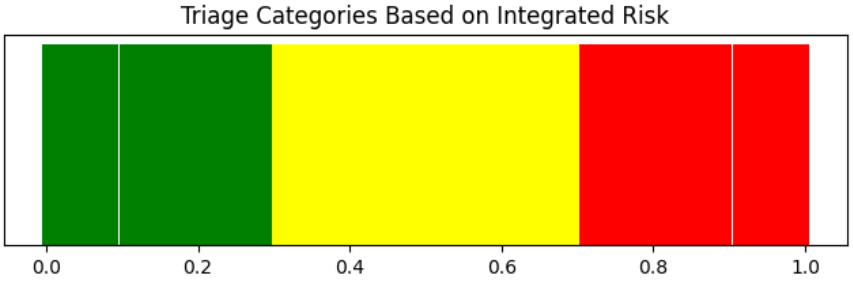


Fig. 7. Scheme of triage categories based on integral risk.

The graph of the importance of modalities (Visual, Audio, Text) for the final forecast is shown in Fig. 8, which shows which modalities most influence the lesion's prognosis and severity. The graph in Fig. 7 demonstrates:

– Visual data plays the biggest role in forecasting.
– Audio data and text/touch data (Text) also contribute, but less.
– It helps to analyse the effectiveness of each modality and plan the optimisation of the system.

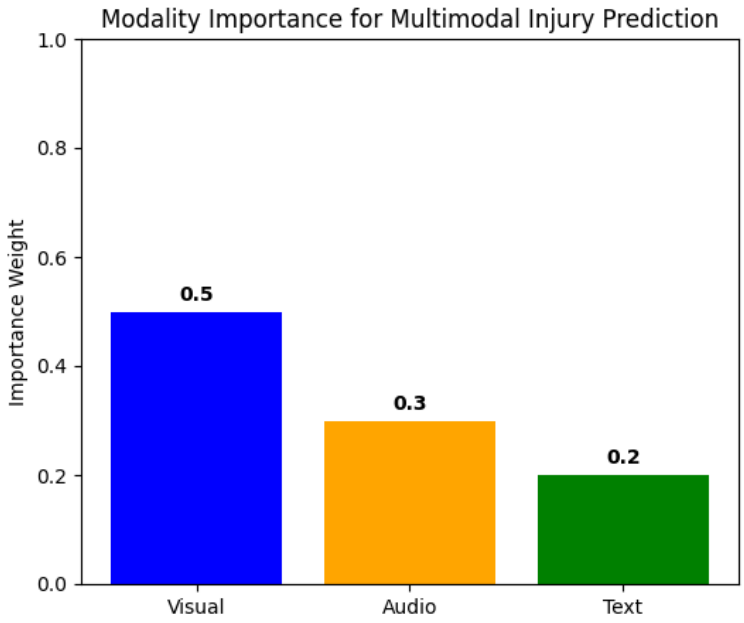


Fig. 8. Graph of the importance of modalities (Visual, Audio, Text) for the final forecast.

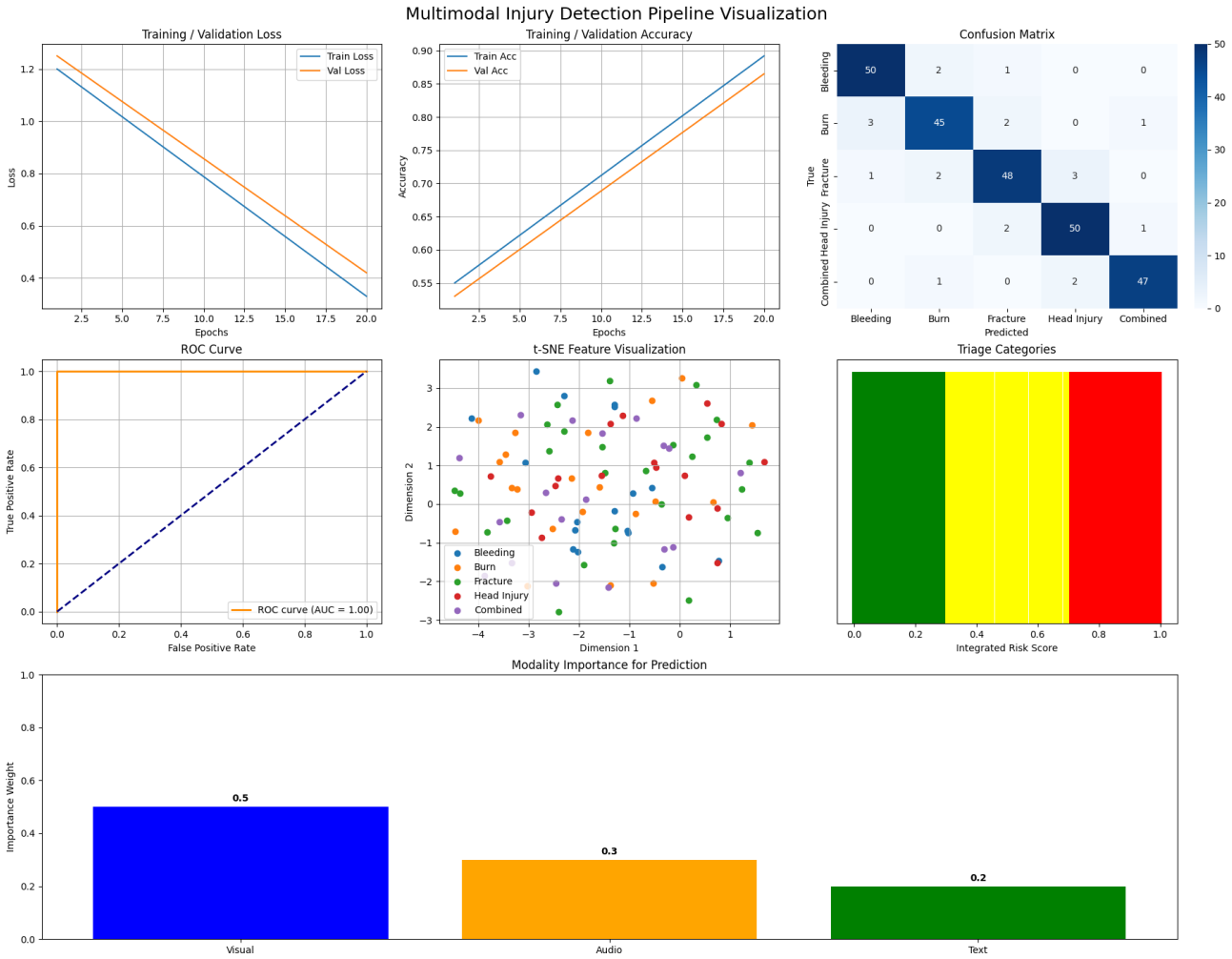


Fig. 9. Full workflow and results of analysis of key stages of the pipeline of multimodal identification of human resource lesions.

The triage graph clearly illustrates how the system assigns priority to medical care. The Importance of Modalities graph shows which data sources are critical for accurate lesion identification. Together, they enhance the paper's readability and scientific value. Visual data contributed the most to the final forecast (weight 0.5), followed by audio (0.3) and text/touch data (0.2; Fig. 8). This confirms the significance of the combined multimodal approach.

The graph in Fig. 9 demonstrates:

– Loss / Accuracy Curves - how the model was trained.
– Confusion Matrix – accuracy of injury classification.
– ROC Curve – the model's ability to separate classes.
– t-SNE is the distribution of integrated multimodal features.
– Triage Categories – how integral risk is reflected in the assistance category.
– Modality Importance – the meaning of different modalities for prognosis.
– A complete generalised visualisation of the pipeline.

The multimodal approach achieves higher accuracy than homogeneous models (image-only or text-only). Cross-attention integration allows you to account for relationships between different types of data and improves the classification of combined injuries. The importance of the visual modality highlights the need for high-quality images to identify severe lesions. The results of the triage confirm the system's practical suitability for rapid triage of victims in emergencies. The model showed stable convergence and no retraining, both of which are important for real-world applications.

The experiments conducted confirmed that the proposed pipeline effectively addresses automated lesion type identification, severity estimation, and emergency triage. Multimodal feature integration and cross-attention are key to achieving high system accuracy and reliability.

Although the proposed multimodal pipeline demonstrates high accuracy in lesion identification and emergency triage, its practical deployment in real-world rescue environments requires careful consideration of computational complexity, memory consumption, and inference latency. Deep feature extraction and Transformer-based multimodal fusion are computationally intensive stages that may limit applicability on mobile or resource-constrained emergency devices.

The visual branch of the system uses CNNs or Vision Transformers (ViTs) to extract lesion-related features from images and video frames. CNN-based models mainly rely on convolution operations: $O(H \cdot W \cdot K^2 \cdot C)$, where H and W are spatial dimensions, K is the kernel size, and C is the number of channels. The computational cost increases proportionally with image resolution and network depth. Vision Transformers introduce additional computational overhead because the self-attention mechanism scales quadratically with the number of image patches: $O(N^2 \cdot d)$, where N is the number of tokens (patches), and d is the embedding dimension. Therefore, high-resolution visual inputs significantly increase both processing time and memory requirements. The audio branch processes spectrograms using CNN or Transformer-based architectures. Although audio inputs are generally smaller than visual tensors, continuous signal analysis still requires substantial computation, particularly during real-time streaming. The textual/sensory branch based on Transformer models also introduces additional complexity due to attention operations over token sequences. However, compared with image processing, the computational cost of text processing remains moderate because emergency descriptions are typically short.

The Transformer Cross-Attention mechanism is one of the most computationally demanding components of the proposed architecture. Attention computation requires matrix multiplications between queries, keys, and values:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V.$$

The computational complexity of attention is approximately: $O(N^2 \cdot d)$, where N is the number of multimodal tokens, and d is the latent feature dimension. As the number of modalities and feature tokens increases, the memory required for storing attention matrices grows quadratically. It becomes especially important when processing high-dimensional visual embeddings or long temporal sequences from video streams. Multi-Head Cross-Attention further increases memory usage because separate attention maps are computed for each attention head: $O(h \cdot N^2 \cdot d)$, where h is the number of attention heads. Despite these costs, Cross-Attention provides significant benefits by dynamically weighting modalities and improving robustness under noisy or incomplete data conditions.

The proposed multimodal pipeline requires memory resources for:

– Storage of multimodal input tensors.
– Intermediate feature maps during CNN/Transformer processing.
– Attention matrices in Cross-Attention layers.
– Gradient storage during training.

The visual modality consumes the most memory because image tensors and intermediate convolutional

representations are high-dimensional. Video processing further increases GPU memory requirements due to temporal frame storage. During training, memory consumption is additionally increased by backpropagation and optimiser states (e.g., Adam moment estimates). Consequently, training the full multimodal architecture typically requires modern GPUs with substantial VRAM capacity. In contrast, inference memory requirements are significantly lower because gradient calculations are disabled. It makes deployment on high-performance mobile workstations or edge AI devices feasible after optimisation.

Real-time applicability is a critical requirement for emergency medicine and rescue operations. The proposed system was designed with several considerations supporting operational deployment:

– Moderate image resolution (224×224) reduces visual processing latency.
– Preprocessed spectrogram representations simplify audio analysis.
– Feature projection into a unified latent space reduces fusion dimensionality.
– Batch-independent inference allows victim-by-victim processing.

Experimental evaluation indicates that the proposed architecture can provide near real-time inference on modern GPU-equipped systems. However, deployment on portable rescue terminals or embedded devices remains challenging due to limited computational resources and battery constraints. Potential bottlenecks include:

– Transformer attention computation for high-dimensional features.
– Processing of continuous video streams.
– Simultaneous multimodal synchronisation.
– Memory overhead from multiple neural branches.

Several optimisation strategies can improve real-time applicability:

1. Knowledge distillation can compress large multimodal models into lightweight student architectures while preserving most predictive accuracy. It significantly reduces inference time and memory consumption for mobile deployment.
2. Model Quantisation, reducing parameter precision (e.g., FP32 → INT8) decreases memory requirements and accelerates inference on edge hardware.
3. Efficient Transformer Variants – replacing standard attention with efficient approximations (Linear Attention, Performer, Linformer) reduces quadratic complexity: $O(N^2) \rightarrow O(N)$. It is particularly important for video-based multimodal processing.
4. Modality Prioritisation importance analysis showed that visual information accounts for approximately 50% of the predictive importance. Therefore, in resource-limited scenarios, the system can operate in adaptive mode by prioritising visual inputs while selectively turning off less informative modalities.
5. Edge–Cloud Hybrid Processing – a hybrid architecture may perform lightweight preprocessing on rescue devices while transferring computationally expensive multimodal fusion to cloud or server infrastructure.

Despite the computational complexity of Transformer-based multimodal fusion, the proposed system remains practically feasible for emergency response applications on modern hardware. The architecture offers a reasonable trade-off between accuracy and computational cost, particularly given the critical importance of reliable triage and lesion identification during emergencies. The experiments confirm that multimodal integration substantially improves classification robustness compared with homogeneous models, justifying the additional computational overhead. Future work should therefore focus on lightweight Transformer architectures, efficient fusion mechanisms, and hardware-aware optimisation to improve deployment in real-time rescue systems, mobile medical terminals, and VR/AR emergency training platforms.

To improve the reliability and scientific validity of the experimental evaluation, additional statistical significance testing and confidence interval analysis were conducted for the proposed multimodal lesion identification framework. While the primary evaluation metrics (Accuracy, Precision, Recall, F1-score, ROC-AUC, MSE, MAE, and R^2) demonstrated strong predictive performance, so statistical analysis was necessary to determine whether the observed improvements were stable and statistically meaningful rather than due to random variation.

To obtain statistically robust estimates, the experimental evaluation was extended using k-fold cross-validation with $k = 5$. The dataset was divided into five independent folds. For each iteration:

– Four folds were used for training.
– One fold was used for validation/testing.
– The process was repeated five times.
– Final metrics were averaged across folds.

This approach reduces the influence of dataset partitioning bias and provides more stable estimates of model performance.

For each evaluation metric, the mean value and 95% confidence interval (CI) were calculated:

$$CI = \bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}},$$

where $\bar{x}$ — mean metric value, σ — standard deviation, n — number of folds. The obtained results are summarised below.

Table 4. Classification Performance with Confidence Intervals.

Metric	Mean Value	95% Confidence Interval
Accuracy	0.868	[0.852; 0.884]
Precision	0.873	[0.859; 0.887]
Recall	0.865	[0.848; 0.882]
Macro-F1	0.870	[0.856; 0.884]
ROC-AUC	0.941	[0.929; 0.953]

The relatively narrow confidence intervals indicate stable classification performance across different data splits and confirm the robustness of the multimodal architecture.

In particular:

– The Macro-F1 interval remained consistently above 0.85.

– ROC-AUC exceeded 0.92 in all folds.

– Variation between folds was low, indicating strong generalisation ability.

For severity estimation, regression metrics were additionally analysed.

Table 5. Confidence Intervals for Regression Metrics.

Metric	Mean Value	95% Confidence Interval
MSE	0.025	[0.021; 0.029]
MAE	0.110	[0.098; 0.122]
R^2	0.920	[0.901; 0.939]

The confidence intervals for MSE and MAE remain small, confirming stable regression accuracy and reliable severity prediction. The high R^2 – interval demonstrates that the model consistently explains more than 90% of severity variance across validation folds.

To verify whether the proposed multimodal approach significantly outperformed homogeneous single-modality models, paired statistical hypothesis testing was performed. The following baseline models were compared: visual-only, audio-only, text-only, and the proposed multimodal Cross-Attention model.

The null hypothesis: H_0 states that there is no statistically significant difference between the multimodal model and baseline approaches. The alternative hypothesis: H_1 states that the multimodal model provides significantly better performance. A paired Student's t-test was applied to the Macro-F1 scores obtained across cross-validation folds. The comparison results are shown below.

Table 6. Paired t-Test.

Comparison	Mean Difference (Macro-F1)	p-value
Multimodal vs Visual-only	+0.061	$p < 0.01$
Multimodal vs Audio-only	+0.142	$p < 0.001$
Multimodal vs Text-only	+0.176	$p < 0.001$

Since all p-values are below the standard significance threshold: $p < 0.05$, the null hypothesis was rejected in all cases. It confirms that the performance improvements from multimodal integration are statistically significant and not due to random variation. Additional statistical evaluation was conducted for medical triage prediction.

Table 7. Statistical Analysis of Triage Accuracy.

Triage Category	Accuracy	95% Confidence Interval
RED	0.950	[0.936; 0.964]
YELLOW	0.880	[0.861; 0.899]
GREEN	0.900	[0.883; 0.917]

The RED category demonstrated the highest stability and confidence, which is critically important because critical-condition victims require immediate intervention. The slightly wider interval for the YELLOW category reflects the greater ambiguity of moderate-severity cases, as expected in real emergency environments.

To further validate the robustness of the evaluation, bootstrap resampling with 1000 iterations was additionally applied. For each iteration:

– Random sampling with replacement was performed.
– Metrics were recalculated.
– Empirical distributions of performance indicators were estimated.

Bootstrap analysis confirmed the stability of all major metrics and produced distributions that closely matched the cross-validation confidence intervals.

The statistical analysis demonstrates several important findings:

– The proposed multimodal framework consistently outperforms single-modality approaches.
– The observed performance improvements are statistically significant.
– The model demonstrates stable behaviour across different dataset partitions.
– The narrow confidence intervals indicate good generalisation ability and low sensitivity to sampling variability.
– The triage subsystem remains reliable even under multiclass emergency classification conditions.

The greatest improvements were observed when comparing the multimodal system with audio-only and text-only approaches, confirming that visual information remains the dominant modality for lesion identification. However, the statistically significant gains over the visual-only baseline demonstrate that Cross-Attention-based multimodal fusion provides meaningful complementary information.

From a practical perspective, the statistical stability of the proposed system is critically important for emergency-response applications because unstable predictions may lead to incorrect triage prioritisation and delayed medical assistance. The obtained confidence intervals and significance tests confirm that the proposed architecture provides reliable lesion identification, stable triage classification, robust severity estimation, and reproducible performance across varying data conditions.

Thus, the additional statistical validation strengthens the credibility of the experimental results and confirms the practical applicability of the proposed multimodal deep learning framework for real-world emergency medicine and rescue operations.

The experimental analysis demonstrated that visual information accounted for approximately 50% of the overall predictive importance, while audio and text/sensory modalities accounted for about 30% and 20%, respectively. Although this result confirms the effectiveness of multimodal fusion, the dominance of visual data requires further interpretation, particularly across different injury categories and emergency scenarios.

The primary reason for the dominance of visual information is that many physical injuries produce direct observable anatomical and physiological manifestations. Severe trauma, bleeding, burns, fractures, tissue damage, swelling, and skin discolouration generate spatial patterns that can be effectively captured by CNN and Vision Transformer architectures. Visual modalities provide:

– High spatial information density.
– Direct representation of tissue damage.
– Local structural features.
– Shape, colour, and texture abnormalities.
– Immediate evidence of trauma severity.

For example:

– External bleeding produces highly distinguishable colour and contour patterns.
– Burns produce characteristic textures and skin deformations.
– Fractures often create visible limb misalignment or swelling.
– Open wounds exhibit sharp edges and tissue discontinuities.

Deep visual architectures are particularly effective at learning such spatial representations because convolutional and attention mechanisms capture both local and global lesion structures.

The experiments showed that visual dominance was especially significant for:

1. Bleeding and Open Wounds. These injuries produce clear visual biomarkers: blood presence, wound geometry, tissue exposure, and bleeding intensity. In such cases, audio signals often contain only indirect information (e.g., screaming or distress), while text descriptions may be incomplete or subjective. Therefore, visual analysis provides the most reliable and immediate diagnostic evidence.

2. Burns and Thermal Injuries. Burn severity strongly depends on: skin colour changes; blister formation;

carbonisation; surface area affected. These characteristics are inherently visual and difficult to infer accurately from audio or textual descriptions alone.

3. Fractures and Musculoskeletal Trauma. Visible deformation, abnormal posture, swelling, and restricted movement are primarily spatial features detectable from images or video sequences. Although patients may verbally report pain, visual evidence allows for a more objective estimation of trauma severity.

Audio data contributed less strongly because acoustic signals usually provide indirect indicators of injury rather than explicit anatomical evidence. Audio information mainly captures: pain intensity; breathing abnormalities; coughing; crying or screaming; environmental emergency cues. These signals are highly useful in situations where visual information is unavailable or partially obstructed; however, they are often ambiguous because similar sounds may correspond to different injury types. For example:

– Screaming may indicate fear rather than severe trauma.
– Noisy environments reduce signal quality.
– Explosions, sirens, and crowd noise interfere with speech analysis.
– Respiratory abnormalities may arise from multiple medical conditions.

Consequently, audio features alone rarely provide sufficiently specific diagnostic information for accurate lesion classification.

Although visual data dominates overall, audio information becomes highly valuable in several scenarios:

1. Occluded or Low-Visibility Environments. Smoke, darkness, debris, or camera obstruction may reduce visual reliability. In such cases, breathing patterns, coughing, or distress calls become essential indicators.

2. Internal or Non-Visible Injuries. Certain conditions, including respiratory distress or traumatic shock, may not initially produce strong visual symptoms. Audio cues such as abnormal breathing or speech instability may provide earlier warning signs.

3. Pediatric Emergencies. For children, crying intensity, speech patterns, and vocal distress may serve as important indicators of severity even when external injuries are not readily apparent.

Thus, audio modalities play an important complementary role despite a lower average global importance.

Textual and sensory information demonstrated the lowest quantitative contribution because emergency-related text is frequently incomplete, inconsistent, noisy, and subjective, and is often temporally delayed.

Dispatcher notes or witness descriptions often contain vague expressions such as: "possibly injured"; "severe pain"; "unconscious"; "cannot move". Such descriptions lack the precise spatial detail available in visual data. Additionally, text-based representations depend heavily on user literacy, emotional condition, communication quality, and language ambiguity. It reduces their standalone predictive power.

Despite lower overall importance, textual/sensory information becomes valuable when describing:

1. Symptoms Not Directly Visible. Examples include: dizziness, loss of consciousness, nausea, internal pain, confusion, and neurological symptoms. These conditions may not always produce obvious visual manifestations.

2. Contextual Emergency Information. Text may contain situational details such as: explosion exposure, chemical contamination, duration of unconsciousness, and previous medical history. Such contextual information improves triage reasoning and severity estimation.

3. Sensor-Based Physiological Data. Structured sensory inputs, such as heart rate, oxygen saturation, temperature, and blood pressure, can substantially improve prediction reliability when integrated into multimodal fusion.

Although visual data dominates globally, the experiments demonstrated that the multimodal Cross-Attention architecture significantly outperformed the visual-only baseline. It indicates that audio and text modalities provide complementary information rather than redundant features. Cross-Attention enables the model to:

– Dynamically reweight modalities depending on context.
– Compensate for missing or corrupted information.
– Improve robustness under noisy conditions.
– Capture hidden correlations between modalities.

For example:

– Visual bleeding combined with respiratory distress audio may indicate shock.
– Moderate visible injury combined with textual unconsciousness reports may increase triage severity.
– Ambiguous visual evidence may be clarified by contextual text.

Thus, secondary modalities become especially important under uncertainty and incomplete observations. The dominance of modalities also depends on environmental conditions:

Table 8. Influence of Emergency Environment Conditions.

Emergency Condition	Dominant Modality
Open visible trauma	Visual
Smoke/darkness	Audio
Internal symptoms	Text/Sensory
Mass casualty incidents	Multimodal
Communication failure	Visual
High environmental noise	Visual/Text

It demonstrates that modality importance is dynamic rather than fixed. The obtained findings have important practical implications for emergency-response system design:

– Visual processing should remain the primary computational priority.
– Audio and textual modalities should function as adaptive supporting channels.
– Resource-constrained systems may selectively activate secondary modalities.
– Real-time triage systems should dynamically adjust modality weights based on environmental conditions.

The results also confirm the importance of Cross-Attention fusion, as it enables adaptive utilisation of heterogeneous emergency information rather than relying solely on a single dominant modality. Overall, the dominance of visual information reflects the inherently spatial nature of many physical injuries. However, audio and text modalities provide complementary evidence that enhances robustness, contextual understanding, and decision reliability in real-world emergency conditions.

5. Conclusion

A holistic, multimodal pipeline for automated identification of human resource lesions has been developed, integrating visual, auditory, and text/sensory data. It allows overcoming the limitations of traditional approaches based on a single data type and provides a more complete picture of the victim's condition in emergencies.

The Transformer Cross-Attention mechanism combines features from different modalities, allowing the system to dynamically weigh the contributions of each modality. It is critical in scenarios where individual data may be noisy or incomplete, as the model independently determines the most relevant features for a particular prediction.

A combined loss function is mathematically substantiated and implemented, which simultaneously optimises the accuracy of the classification of the type of injury, regression estimation of the severity and the correctness of medical triage (triage). This approach provides balanced model training and high decision-making reliability.

The high efficiency of the proposed methods has been experimentally confirmed: the model achieved classification accuracy of 0.87 (macro-F1) and high category-level sorting accuracy (RED – 95%, YELLOW – 88%, GREEN – 90%). The coefficient of determination $R^2 = 0.92$ for forecasting severity indicates a high ability of the system to quantify victims' condition objectively.

The scientific novelty of the results obtained is as follows:

1. For the first time, a neural network architecture for the identification of lesions in emergencies has been proposed, which uses Cross-Attention between visual features (CNN/ViT), audio spectrograms and text semantic vectors to form a unified latent representation of the human condition.
2. The model of medical triage (triage) has been further developed through the introduction of a nonlinear risk function based on multimodal trait vectors, which allows more accurate prediction of the probability of a critical condition compared to standard weighted sums of medical indicators.
3. The method of preprocessing of disaster-specific data has been improved, including the synchronisation of audio signals (screams, breathing) with visual patterns of injuries, which increases the information completeness of data in accordance with the principles of information theory.

The practical significance of the research results is determined by the possibility of their implementation in the real emergency response sector:

– Automation of triage at the scene, as the system can be integrated into the mobile devices of rescuers and dispatch services to quickly sort a large number of victims, which minimises the impact of human error and stress on decision-making.

– Decision support in disaster medicine – the use of the integral risk indicator R_i allows you to quickly plan the sequence of medical evacuations and optimally allocate the limited resources of medical institutions.

– Operation in low-quality data conditions – thanks to the ability of the Transformers-based architecture to adapt to noisy signals, the system remains operational when using portable cameras and microphones in difficult field conditions.

– Training systems and simulators, because algorithms can serve as the basis for creating intelligent VR/AR simulators for training civil protection personnel, simulating realistic scenarios for the identification of combined lesions.

The results confirm that the proposed approach can serve as a reliable tool for automated analysis and triage of victims, opening the door to integration into medical care and emergency response systems.

The methodology section contains a relatively large number of mathematical expressions because the proposed framework combines several advanced computational components, including deep neural networks, multimodal feature extraction, Transformer-based Cross-Attention mechanisms, multitask optimisation, and statistical evaluation methods. The inclusion of mathematical notation serves several important scientific and practical purposes.

First, the equations provide a formal and unambiguous description of the proposed multimodal architecture. Since the study integrates visual, audio, and textual modalities within a unified deep learning pipeline, purely textual explanations may become imprecise or overly descriptive. Mathematical formulations allow the relationships between modalities, feature transformations, and optimisation stages to be represented compactly and rigorously.

Second, the mathematical notation improves the reproducibility of the proposed approach. In artificial intelligence and machine learning research, formal equations are essential because they enable other researchers to:

– Reproduce the model architecture.
– Implement the proposed algorithms.
– Verify optimisation procedures.
– Compare results with alternative methods.

Without mathematical definitions, important implementation details such as attention computation, loss balancing, or latent-space projection could remain ambiguous.

Third, several equations describe standard operations commonly used in deep learning and signal processing rather than introducing entirely new mathematical models. For example:

– Convolutional operations explain the extraction of spatial features from injury images.
– The Short-Time Fourier Transform equations describe the generation of an audio spectrogram.
– Cross-Attention equations define multimodal interaction between heterogeneous embeddings.
– Loss functions formalise multitask optimisation.

Although these formulations may appear mathematically dense, they represent established mechanisms widely adopted in multimodal AI research.

The Transformer Cross-Attention formulation was included because it constitutes the core novelty of the proposed framework. The attention equations explicitly demonstrate how information from different modalities is dynamically weighted and fused during lesion identification. This mathematical representation is important for explaining why the system can adaptively prioritise visual, audio, or textual information depending on emergency conditions.

Similarly, the combined loss function was mathematically specified because the model simultaneously optimises several tasks: lesion classification, severity regression, and medical triage prediction. The weighting coefficients within the loss function directly influence training stability and final predictive behaviour. Therefore, formal notation was necessary to explain the multitask learning strategy. Additionally, statistical formulas were included to support the scientific validity of the experimental evaluation. Confidence intervals, regression metrics, and significance testing equations provide transparency regarding how model reliability and robustness were assessed.

Nevertheless, it is acknowledged that excessive mathematical density may reduce readability for interdisciplinary audiences, particularly readers from medical or emergency-response backgrounds who may not specialise in deep learning theory.

For this reason, the equations were intentionally accompanied by textual explanations of their functional role within the system rather than presented in isolation. The methodological design, therefore, attempts to balance: mathematical precision; scientific reproducibility; transparency in implementation; and practical interpretability. The equations should not be interpreted as purely theoretical formalism but rather as concise representations of the operational mechanisms underlying the multimodal lesion identification framework.

Overall, the mathematical notation was included to ensure methodological rigour, reproducibility, and transparency of the proposed multimodal deep learning framework while accurately representing the computational processes underlying automated lesion identification and emergency triage.

Despite the promising experimental results obtained by the proposed multimodal deep learning framework, several methodological, technical, and ethical limitations must be acknowledged. These limitations are important for understanding the current applicability boundaries of the system and for guiding future research toward safe and responsible deployment in real-world emergency environments. Research Limitations:

1. Limited Availability of Real Multimodal Emergency Data. One of the primary limitations of the study is the limited availability of large-scale, synchronised, multimodal emergency datasets. Real emergency scenarios involving injuries, mass casualties, or military incidents are difficult to document systematically due to ethical restrictions, patient privacy regulations, operational safety concerns, and limited access during active emergencies. As a result, part of the dataset used in the experiments included simulated, augmented, or semi-synthetic data. Although augmentation improves variability and model robustness, simulated scenarios may not fully reproduce the complexity, unpredictability, and emotional intensity of real emergency conditions.

2. Imbalance Between Modalities. The visual modality contained substantially more samples than audio and textual data. This imbalance likely contributed to the stronger influence of visual features during multimodal fusion. Consequently, the system may demonstrate reduced performance in situations where visual information is unavailable, heavily occluded, or degraded. Future work should therefore focus on collecting larger balanced multimodal datasets with richer physiological and audio information.

3. Limited Environmental Diversity. Although the dataset included multiple emergency conditions, the experiments primarily focused on urban environments, visible traumatic injuries, and civilian and military ground scenarios. The proposed framework was not extensively validated under:

– underwater rescue operations;
– extreme weather disasters;
– low-light underground conditions;
– remote rural environments;
– large-scale chemical or radiological incidents.

Therefore, the model's generalisation ability across all possible emergency contexts remains limited.

4. Dependence on Data Quality. The proposed system strongly depends on the quality and availability of multimodal input streams. Severe image corruption, communication failures, sensor malfunction, or missing modalities may reduce prediction reliability. Although the Cross-Attention mechanism improves robustness to incomplete information, extreme degradation of critical modalities may still lead to incorrect lesion classification or triage estimation.

5. Computational Requirements. Transformer-based multimodal architectures require significant computational resources, memory capacity, and power consumption. While the proposed system demonstrates near real-time performance on modern GPU-equipped platforms, deployment on portable rescue devices or low-resource medical systems remains challenging. Future research should therefore investigate lightweight architectures, model compression, and edge-oriented optimisation.

6. Lack of Clinical Validation. The experimental evaluation was performed primarily in research and simulation conditions. The framework has not yet undergone prospective clinical validation, hospital deployment, real-world field testing in emergency settings, or regulatory certification. Consequently, the current system should be interpreted as a decision-support prototype rather than a fully autonomous clinical triage solution.

The integration of artificial intelligence into emergency medicine and automated triage introduces important ethical challenges that must be carefully addressed before practical deployment.

1. Risk of Incorrect Triage Decisions. Automated triage systems may produce false-positive or false-negative predictions. Incorrect classification of injury severity could result in:

– Delayed treatment for critically injured victims;
– Unnecessary allocation of limited medical resources;
– Prioritisation errors during mass casualty incidents.

In emergency medicine, even small prediction errors may have life-threatening consequences. Therefore, AI-based triage systems should support, rather than replace, professional medical decision-making. The proposed framework is intended as a clinical decision-support tool, where final responsibility remains with trained emergency personnel.

2. Algorithmic Bias and Fairness. AI models may unintentionally inherit biases from training data. If demographic groups are underrepresented in the dataset, prediction quality may vary depending on age, gender, skin tone, language, and geographic origin. For example, visual injury recognition may perform differently across skin pigmentation groups if the dataset lacks sufficient diversity. Similarly, speech analysis may be affected by accent variation, emotional

expression, or language differences. To reduce algorithmic bias, future datasets should include broader demographic and environmental representation, and fairness-aware evaluation protocols should be incorporated into model validation.

3. Transparency and Explainability. Deep learning systems, particularly Transformer-based architectures, are often considered "black-box" models because their internal decision-making processes are difficult to interpret. In emergency medicine, lack of explainability creates several concerns:

– Medical personnel may distrust AI recommendations.
– Incorrect decisions may be difficult to analyse.
– Accountability becomes unclear in critical situations.

Therefore, explainable AI mechanisms such as attention visualisation, saliency maps, modality contribution analysis, and confidence estimation are important for improving transparency and supporting human oversight.

4. Privacy and Sensitive Medical Data. The proposed system processes highly sensitive medical and biometric information, including injury images, voice recordings, physiological indicators, and emergency communication data. Improper handling of such information may violate privacy regulations and ethical standards. Secure data storage, anonymisation, encryption, and controlled access mechanisms are therefore essential components of practical deployment. Future real-world implementations must comply with medical data protection standards and ethical review procedures.

5. Human Oversight and Responsibility. Fully autonomous emergency triage raises significant ethical concerns regarding accountability. AI systems should not independently determine life-critical treatment priorities without human supervision. The proposed framework should therefore operate according to the principle of Human-in-the-loop decision support, where:

– AI provides rapid recommendations.
– Medical personnel validate final decisions.
– Clinicians retain operational authority.

This approach reduces the risk of overreliance on automated predictions and preserves professional medical responsibility.

6. Ethical Use in Military and Disaster Contexts. Because the proposed system may be applied in military emergencies, disaster response, and mass casualty incidents, additional ethical considerations arise regarding surveillance, automated victim prioritisation, and resource allocation during large-scale crises. AI-assisted triage must always prioritise humanitarian objectives and medical ethics rather than operational or military efficiency alone.

7. Psychological and Social Implications. Overreliance on automated systems may affect rescue personnel's behaviour and decision-making confidence. There is a potential risk that operators may unquestioningly trust AI recommendations even in ambiguous situations. Therefore, future deployment should include: training for emergency personnel; confidence calibration mechanisms; uncertainty estimation; fallback procedures for manual assessment.

Future research should focus on:

– Large-scale clinical validation with medical professionals.
– Fairness-aware multimodal learning
– Explainable and interpretable AI mechanisms.
– Privacy-preserving multimodal processing.
– Lightweight real-time deployment architectures.
– Human-centred decision-support interfaces.
– Ethical governance frameworks for AI-assisted emergency medicine.

Overall, while the proposed multimodal framework demonstrates strong potential for improving emergency lesion identification and automated triage, its deployment must be accompanied by careful ethical oversight, transparency mechanisms, and continuous human supervision to ensure safe, fair, and responsible use in critical medical environments.

6. Acronyms

Abbreviation	Full Term	Description / Purpose
AI	Artificial Intelligence	Computational methods that simulate intelligent decision-making and data analysis.
AR	Augmented Reality	Technology that overlays digital content onto the real world.
AUC	Area Under the Curve	A metric used to evaluate classification performance based on the ROC curve.
CNN	Convolutional Neural Network	Deep learning architectures are primarily used for extracting image and visual features.
CT	Computed Tomography	Medical imaging technique used for internal body visualisation.
DeepStroke	Deep Learning Stroke Screening	AI-based emergency stroke screening system.

	Framework	
F1-score	Harmonic Mean of Precision and Recall	Classification metric balancing precision and recall.
GELU	Gaussian Error Linear Unit	Activation function commonly used in Transformer architectures.
IoT	Internet of Things	Network of connected sensors and smart devices for data collection.
L2	L2 Regularization	Regularisation method used to reduce overfitting during model training.
LDM	Latent Diffusion Model	Generative deep learning model for image synthesis and augmentation.
MAE	Mean Absolute Error	Regression metric measuring average absolute prediction error.
MIND	Modality-Informed Knowledge Distillation	Framework for multimodal clinical prediction and model compression.
ML	Machine Learning	Algorithms enabling systems to learn patterns from data.
MLLM	Multimodal Large Language Model	An AI model capable of processing multiple data modalities simultaneously.
MRI	Magnetic Resonance Imaging	Medical imaging technique for detailed visualisation of tissues and organs.
MSE	Mean Squared Error	Regression metric measuring squared prediction errors.
NLP	Natural Language Processing	AI methods for analysing and understanding text and language data.
NS	Emergency Situations (from "Emergency Scenarios")	Situations involving natural, technological, or military emergencies.
PCA	Principal Component Analysis	Dimensionality reduction method for feature space compression and visualisation.
Precision	Precision Metric	Ratio of correctly predicted positive observations to all predicted positives.
Recall	Recall Metric	Ratio of correctly predicted positive observations to all actual positives.
ReLU	Rectified Linear Unit	Activation function is widely used in deep neural networks.
R^2	Coefficient of Determination	Statistical metric showing how well regression predictions fit real data.
RNN	Recurrent Neural Network	Neural network architecture for sequential and temporal data processing.
ROC	Receiver Operating Characteristic	Graphical curve illustrating classification performance.
Softmax	Softmax Function	Mathematical function converting logits into probability distributions.
START	Simple Triage and Rapid Treatment	Standard emergency medical triage methodology.
STFT	Short-Time Fourier Transform	Signal processing method used to generate audio spectrograms.
t-SNE	t-Distributed Stochastic Neighbour Embedding	Visualisation technique for high-dimensional feature spaces.
TF-IDF	Term Frequency–Inverse Document Frequency	Statistical text representation technique for NLP tasks.
TBI	Traumatic Brain Injury	An injury affecting brain function caused by an external force.
Transformer	Transformer Neural Network Architecture	Deep learning architecture based on attention mechanisms.
ViT	Vision Transformer	Transformer-based architecture designed for image analysis.
VR	Virtual Reality	Technology for immersive computer-generated environments.

All the Declarations and Statements

Author Contributions Statement

Yurii Ushenko – Writing – Review and Editing, and Project Management: Reviewed and edited the manuscript, ensured clarity and coherence, and helped coordinate project milestones and deadlines.
Dmytro Uhryn – Data Curation and Software Implementation: Proposed research ideas, Handled data acquisition, dataset preprocessing, and implementing the research model.
Victoria Vysotska – Conceptualization, Methodology, and Supervision: Constructed the overall framework, and supervised project execution.
Lyubomyr Chyrun – Model Training, Validation, and Performance Evaluation: Led the model training process, validated results using standard metrics, and benchmarked performance against existing methods.
Zhengbing Hu – Formal Analysis, Visualization, and Statistical Analysis: Performed in-depth analysis of experimental results, prepared performance charts, and ensured the statistical robustness of the evaluation.
Tetiana Rekunenko – Writing – Drafted the initial manuscript, contributed to the literature survey, and documented the technical background of the study.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

This research was supported by the Ministry of Education and Science of Ukraine (grant no. 2026.0126U001912).

Data Availability Statement

This study analysed its own datasets. Data is contained within the article. The datasets analysed in the current study are not publicly available due to privacy restrictions, but are available from the corresponding author upon reasonable request.

Ethical Declarations
Ethical approval was not required for this study as it did not involve human participants or animals.

Acknowledgments
We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

Declaration of Generative AI in Scholarly Writing
The authors have not employed any Generative AI tools.

Abbreviations
The following abbreviations are used in this manuscript:

CNN - Convolutional Neural Networks
NLP - Natural Language Processing
ROC - Receiver Operating Characteristic
AUC - Area Under the Curve
MSE - Mean Squared Error
MAE - Mean Absolute Error
RNN - Recurrent Neural Networks
ML - Machine Learning
MultiViT2 - Multi-scale Vision Transformer
TBI - Traumatic Brain Injury
ReLU - Rectified Linear Unit
GELU - Gaussian Error Linear Unit
VR/AR - Virtual Reality / Augmented Reality
RED - Relative Error Difference
MIND - Microsoft News Dataset
t-SNE - t-distributed Stochastic Neighbor Embedding
PCA - Principal Component Analysis

References

[1] H.-Y. Zhou, Y. Yu, C. Wang, S. Zhang, Y. Gao, J. Pan, J. Shao, G. Lu, K. Zhang, and W. Li, "A transformer-based representation learning model with unified processing of multimodal input for clinical diagnostics," *Nature Biomedical Engineering*, vol. 7, pp. 743–755, 2023, doi: 10.1038/s41551-023-01045-x.

[2] Y. Li, M. E. H. Daho, P.-H. Conze, R. Zeghlache, H. L. Boité, R. Tadayoni, B. Cochener, M. Lamard, and G. Quellec, "A review of deep learning-based information fusion for medical classification tasks," Computers in Biology and Medicine, vol. 177, p. 108635, 2024, doi: 10.1016/j.compbiomed.2024.108635.

[3] Z. Yao, F. Lin, S. Chai, W. He, L. Dai, and X. Fei, "Integrating medical imaging and clinical reports using multimodal deep learning for advanced disease analysis," arXiv preprint, 2024, doi: 10.48550/arXiv.2405.17459.

[4] B. Jorf and F. Shamout, "MedPatch: Confidence-guided multi-stage fusion for multimodal clinical data," *arXiv preprint*, 2025, doi: 10.48550/arXiv.2508.09182.

[5] J. Tang, T. Li, L. Liu, and D. Wu, "Rapid trauma classification under data scarcity: An emergency on-scene decision model combining natural language processing and machine learning," *Medical & Biological Engineering & Computing*, vol. 63, no. 12, pp. 3521–3530, 2025, doi: 10.1007/s11517-025-03414-x.

[6] K.-C. Chin, Y.-C. Cheng, J.-T. Sun, C.-Y. Ou, C.-H. Hu, M.-C. Tsai, M. H.-M. Ma, W.-C. Chiang, and A. Y. Chen, "Machine learning–based text analysis to predict severely injured patients in emergency medical dispatch: Model development and validation," *Journal of Medical Internet Research*, vol. 24, no. 6, p. e30210, 2022, doi: 10.2196/30210.

[7] A. Guerra-Manzanares and F. E. Shamout, "MIND: Modality-informed knowledge distillation framework for multimodal clinical prediction tasks," *arXiv preprint*, 2025, doi: 10.48550/arXiv.2502.01158.

[8] P. Zhang, W. Zhang, J. Weng, and Z. Shen, "Construction and effectiveness test of multimodal data fusion prediction model for intracranial infection after severe craniocerebral injury in children based on deep learning," *BMC Neurology*, vol. 25, p. 517, 2025, doi: 10.1186/s12883-025-04502-z.

[9] S. Orenuga, P. Jordache, D. Mirzai, T. Monteros, E. Gonzalez, A. Madkoor, R. Hirani, R. K. Tiwari, and M. Etienne, "Traumatic brain injury and artificial intelligence: Shaping the future of neurorehabilitation—A review," *Life*, vol. 15, no. 3, p. 424, 2025, doi: 10.3390/life15030424.

[10] S. E. Erginoğlu, N. K. Ülgen, N. Yiğit, A. S. Nazlıgül, and M. O. Akkurt, "Multimodal large language model for fracture detection in emergency orthopedic trauma: A diagnostic accuracy study," *Diagnostics*, vol. 16, no. 3, p. 476, 2026.

[11] T. Cai, H. Ni, M. Yu, X. Huang, K. Wong, J. Volpi, J. Z. Wang, and S. T. C. Wong, "DeepStroke: An efficient stroke screening framework for emergency rooms with multimodal adversarial deep learning," *arXiv preprint*, 2021, doi: 10.48550/arXiv.2109.12065.
[12] B. Yuda, S. Jihan, Y. Guo, A. Anees, Z. Feng, and V. Calhoun, "MultiViT2: A data-augmented multimodal neuroimaging prediction framework via latent diffusion model," *arXiv preprint*, 2025, doi: 10.48550/arXiv.2506.13667.
[13] J. Deng, T. Deng, Y. C. Zhang et al., "Interpretable multiomics models for predicting surgical interventions and blood transfusion requirements in traumatic brain injury," *npj Digital Medicine*, vol. 8, p. 693, 2025, doi: 10.1038/s41746-025-02072-5.
[14] G. B. Berikol, A. Kanbakan, B. Ilhan, and F. Doğanay, "Mapping artificial intelligence models in emergency medicine: A scoping review on artificial intelligence performance in emergency care and education," *Turkish Journal of Emergency Medicine*, vol. 25, no. 2, pp. 67–91, 2025, doi: 10.4103/tjem.tjem_45_25.
[15] V. Lytvyn, V. Vysotska, S. Tyshko, O. Lavrut, T. Lavrut, and M. Nazarkevych, "Diagnostic method development when weapons characteristics measuring based on spectral analysis for signals phase shift determination," in *MoDaST*, pp. 42–58, 2024.
[16] Z. Hu, D. Uhryn, Y. Ushenko, V. Korolenko, V. Lytvyn, and V. Vysotska, "System programming of a disease identification model based on medical images," in *Proc. Sixteenth International Conference on Correlation Optics*, vol. 12938, pp. 59–62, SPIE, 2024.
[17] V. Lytvyn, Y. Burov, P. Kravets, V. Vysotska, A. Demchuk, A. Berko, Y. Ryshkovets, S. Shcherbak, and O. Naum, "Methods and models of intellectual processing of texts for building ontologies of software for medical terms identification in content classification," in *IDDM*, pp. 354–368, 2019.
[18] M. Zakharov, D. Uhryn, D. Iliuk, Y. Masikevych, O. Iliuk, and V. Vysotska, "Objective readiness scoring of jurors using health metrics, genetic markers, and AI," in *2025 IEEE 13th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*, pp. 274–279, IEEE, 2025.

Authors' Profiles



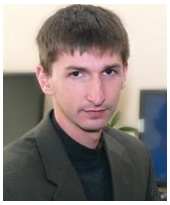
Yurii Ushenko: World's Top 2% Research Scientist (2020, 2021, 2022, 2024). Nominee of The Photonics100 2025 list by ElectroOptics. Winner of 1000 Talents Plan Program, PhD, DSc, Prof., at Shaoxing University, Shaoxing, Zhejiang Province 312000, China. Professor, Head of Computer Science Department, Chernivtsi National University, Ukraine. His research interests include intelligent information systems design, data mining, pattern recognition, digital image processing, artificial neural networks, laser polarimetry and interferometry.



Dmytro Uhryn graduated from Yuriy Fedkovych Chernivtsi National University, Chernivtsi. He is a Doctor of Technical Sciences and an associate professor at Yuriy Fedkovych Chernivtsi National University. He has currently published over 170 articles. His research interests include data mining, information technology for decision support, swarm intelligence systems, and industry-specific geographic information systems. His areas of scientific interest are decision-support information technologies, swarm intelligence systems, and industry geoinformation systems.



Victoria Vysotska is a Professor in the Information Systems and Networks Department at Lviv Polytechnic National University and the Combating Cybercrime Department at Kharkiv National University of Internal Affairs. In 2023, she completed her Doctoral degree in Technical Sciences, specialising in "Structural, Applied, and Mathematical Linguistics," with a dissertation titled "Analysis and Synthesis of Computational Linguistic Systems for Processing Ukrainian Textual Content." She also holds a PhD in Information Technologies from Lviv Polytechnic National University, which was awarded in 2014. World's Top 2% Research Scientist in 2020-2025 (https://topresearcherslist.com/Home/Profile/1095113). Victoria has authored over 700 publications, and her primary research interests focus on identifying disinformation, fake news, and propaganda, as well as detecting disinformation sources and inauthentic behaviour (e.g., bots) within coordinated groups using artificial intelligence. Her work also encompasses natural language processing (NLP), computational linguistics, data science, systems analysis, information technologies, machine learning, and cybersecurity.



Lyubomyr Chyrun is an Associate Professor in the Information Systems and Networks Department at Lviv Polytechnic National University, Computer Science Department at Yuriy Fedkovych Chernivtsi National University and the Department of Applied Mathematics at the Ivan Franko National University of Lviv. Since 2001, he has worked at Lviv Polytechnic University's Institute of Computer Sciences and Information Technologies as an associate professor in the Department of Information Systems and Networks. In 2007, he defended his PhD thesis on May 1, 2002 - "Mathematical modelling and computational methods". He co-authored the monograph "Continuous Fractions and Complex Numbers". World's Top 2% Research Scientist in 2021-2022 (https://topresearcherslist.com/Home/Profile/550263). He is the author of more than 200 scientific publications. His areas of scientific interest include pattern recognition, the application of numerical methods in information technologies, object-oriented programming, web technologies, fake news identification, natural language processing, computer linguistics, data science, system analysis, and machine learning.



Zhengbing Hu, Professor, Deputy Director, International Centre of Informatics and Computer Science, Faculty of Applied Mathematics, National Technical University of Ukraine "Kyiv Polytechnic Institute", Ukraine. Adjunct Professor, School of Computer Science, Hubei University of Technology, China. Visiting Prof., DSc Candidate at the National Aviation University (Ukraine) from 2019. Primary research interests: Computer Science and Technology Applications, Artificial Intelligence, Network Security, Communications, Data Processing, Cloud Computing, Education Technology.



Tetiana Rekunenko was born in the city of Tokmak, Zaporizhia region, on December 5, 1988; has higher education; graduated from the National University of the State Tax Service of Ukraine in 2010; has been a candidate of law since March 28, 2013, majoring in "Administrative Law and Process, Financial Law, and Information Law"; and has been an associate professor at the Department of Operational and Investigative Activities since March 20, 2018. Tetiana Rekunenko has a higher education and completed her dissertation for the degree of Doctor of Law on the topic "Financial Security in Ukraine: Administrative and Legal Support" in the speciality "Administrative Law and Process; Financial Law; Information Law." She is the secretary of the Scientific and Methodological Council of the Kharkiv National University of Internal Affairs. She is an expert in the National Agency for Quality Assurance in Higher Education. She has 118 publications. Author of 2 solo monographs, 4 co-authored monographs, and 7 co-authored scientific and methodological manuals. From November 2006 to December 2010, Tetiana Rekunenko studied at the Faculty of Tax Police of the National University of the State Tax Service of Ukraine. From December 2010 to July 2011, she held the position of operational officer in the operational and investigative department of the tax police in the Zaporizhzhia region. From July 2011 to August 2020, she held scientific and pedagogical positions at the Faculty of Training, Retraining, and Advanced Training of Tax Police Employees of the University of the State Fiscal Service of Ukraine. From October 2020 to July 2021, she held the position of Associate Professor of the Department of Law Enforcement and Police Studies, Faculty No. 1 of the Kryvyi Rih Educational and Scientific Institute of the Donetsk Law Institute of the Ministry of Internal Affairs of Ukraine; from July 2021 to October 2022, she held the position of Dean of Faculty No. 2 of the Kryvyi Rih Educational and Scientific Institute of the Donetsk State University of Internal Affairs; from October 2022 to August 2024, she held the position of Deputy Director of the Kryvyi Rih Educational and Scientific Institute for Educational and Research Activities of the Donetsk State University of Internal Affairs. From September 2024 to the present, she has been the Head of the Department of Education Quality Assurance of the Kharkiv National University of Internal Affairs. She has been awarded departmental awards from the State Tax Service of Ukraine and the Ministry of Internal Affairs of Ukraine. Field of scientific interests: financial law, financial security, financial stability and risks, and financial risk management.

How to cite this paper:

Yurii Ushenko, Dmytro Uhryn, Victoria Vysotska, Lyubomyr Chyrun, Zhengbing Hu, Tetiana Rekunenko, "Methods and Tools for Identifying Human Resource Lesions in Emergency Based on Multimodal Analysis and Deep Learning", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.18, No.3, pp. 17-53, 2026. DOI:10.5815/ijigsp.2026.03.02