Multi-modal learning for skin lesion (bruise) timeline analysis

Mehrdad Ghyabi
George Mason University
mghyabi@gmu.edu

Abstract

Bruise age estimation is a critical task in forensic science and medical diagnostics, where accurate timing of injuries can have significant legal and clinical implications. Traditional methods rely heavily on visual inspection, which is subjective and often unreliable due to individual variations in bruise healing. This project presents a novel multi-modal deep learning approach that integrates visual and textual data to improve bruise age estimation. A dataset of over 32,000 images, captured under controlled lighting conditions and accompanied by participant-specific physiological attributes, was used for model development. Separate models were first trained on image data using convolutional neural networks (CNNs) and on structured textual data using both tabular encoding and BERT-based text embeddings. The final architecture fuses image and text-derived features to produce individualized age predictions. Experiments were conducted under multiple settings to assess the impact of data representation and fusion strategies. Root Mean Squared Error (RMSE) was used as the primary metric to evaluate accuracy. Results demonstrate that the multi-modal fusion approach outperforms single-modality models, highlighting the value of integrating heterogeneous data sources for personalized bruise age estimation.

1 Introduction

This project presents a novel approach to the problem of bruise age estimation by combining two traditionally separate streams of information: visual image data and textual data derived from structured clinical and demographic records. While previous research has largely focused on user aided unimodal methods—either applying computer vision techniques to bruise images or using clinical data in isolation—this project proposes a genuinely multi-modal learning framework that integrates both data types to enable personalized, context-aware bruise age prediction.

The innovation lies not only in the fusion of modalities but also in how each modality is processed and utilized. On the visual side, convolutional neural networks (CNNs) are employed to learn features describing temporal changes in bruise color, texture, and morphology to capture evolution of a bruise over time. Simultaneously, the project introduces a novel application of tabular data transformation: structured participant attributes (e.g., age, skin tone, body fat, etc.) are encoded and embedded into a unified feature space, allowing the model to account for inter-individual variability in healing rates. This effectively turns patient-specific metadata—often underutilized or discarded in computer vision pipelines—into an active component of the learning process.

Although multi-modal learning has been explored in other domains (e.g., image-captioning or medical imaging paired with clinical notes), its application to bruise age estimation is original. The use of structured tabular information as a complementary modality in deep learning, combined with image-based analysis, is a creative extension of existing methodologies. Rather than simply reusing standard architectures, the project takes an integrative stance, implementing feature-level fusion techniques

Workshop on Instruction Tuning and Instruction Following at NeurIPS 2023.

to jointly learn from heterogeneous data sources. This is assumed to allow the model to uncover interactions between how a bruise appears visually and the underlying physiological conditions influencing that appearance—something that is probably harder for either modality to achieve in isolation.

This framework is defined to push the boundaries of current forensic technology by offering a more objective, data-rich alternative to bruise aging. It is supposed to allow more accurate, individualized assessments that could enhance the reliability of forensic evidence and medical diagnostics.

2 Related Work

Estimating the age of bruises has long been a significant challenge in both forensic and clinical contexts. Traditional methods rely heavily on visual assessments by professionals, which can be highly subjective and inconsistent. Several studies have attempted to standardize this process by introducing image processing techniques and colorimetric analysis, but these methods often fall short due to inter-individual variability in bruise development and healing [1]. More recent advances in machine learning and computer vision have opened new avenues for objective and automated bruise age estimation, particularly using convolutional neural networks (CNNs) to extract meaningful features from bruise images.

In the computer vision domain, CNN-based approaches have proven highly effective for medical image analysis. For instance, deep CNNs have been successfully used for skin lesion classification, fracture detection in X-rays, and wound healing analysis. However, in the specific context of bruises, existing studies such as [2] have primarily focused on using CNNs to classify bruise presence or severity, without accounting for the temporal progression of healing or individual-specific variables. Thus, these models may not generalize well to diverse populations or accurately estimate bruise age over time.

Multi-modal learning has gained traction as a promising strategy to address the limitations of unimodal approaches in various domains, including healthcare, natural language processing, and affective computing. Multi-modal models are designed to combine complementary information from different data types—such as images and text—thereby capturing richer patterns than single-modality models. For instance, Hayat et al. [3] demonstrated that combining radiology images with patient metadata led to significantly improved diagnostic accuracy in detecting thoracic diseases.

A particularly relevant advance for handling textual data in a multi-modal setup is the development of BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. [4]. BERT models excel at producing context-aware embeddings from raw, unstructured text and have been widely adopted in clinical applications to process electronic health records, radiology reports, and other medical documentation. Their bidirectional nature allows them to understand text more holistically compared to other models like word2vec or GloVe. For example, Hao et al. [5] used BERT embeddings to encode clinical narratives and achieved superior performance in predicting patient outcomes compared to traditional methods.

Feature-level fusion, the strategy, has been shown to be particularly effective for integrating multi-modal data. Studies like the one by and Zhang et al. [6] have outlined various fusion techniques—early, intermediate (feature-level), and late fusion—and have found that feature-level fusion provides a balance between expressivity and flexibility, especially when dealing with heterogeneous data sources.

3 Problem Statement

Accurately estimating the age of bruises is a critical challenge in both forensic and clinical settings due to the non-uniform, individualized nature of the healing process. Traditional bruise age assessments rely primarily on subjective visual inspection, which is not only prone to human error but also fails to account for a host of biological and contextual variables—such as patient demographics, skin tone, body composition, and medication use—that significantly influence bruise development and recovery. While machine learning methods have been applied to image-based bruise analysis, these approaches are limited by their unimodal focus and often overlook the broader physiological context that governs healing variability across individuals.

The core challenge, therefore, is to develop an automated, objective system that can estimate bruise age, with an acceptable level of error, by learning from both the visual evolution of bruises and the diverse set of patient-specific attributes that affect recovery. This requires a solution that can not only model complex temporal patterns in bruise appearance but also effectively integrate heterogeneous data modalities—including unstructured textual health records and structured demographic or physiological variables.

This project aims to address this problem by designing a multi-modal deep learning framework that fuses visual and textual data to learn personalized, data-driven mappings from input features to bruise age. Key challenges include the standardization of irregular textual data for model integration, the modeling of time-dependent visual changes using CNNs, and the effective fusion of visual and non-visual data at the feature level. The solution space is intentionally open-ended to allow for exploration of various data encoding, fusion, and learning strategies that can generalize across diverse populations and imaging conditions. The overarching goal is to build a system that improves the accuracy, consistency, and interpretability of bruise age estimation in real-world forensic and medical environments.

4 Methodology

To develop a model for bruise age estimation, several experiments were conducted to evaluate different data modalities individually before integrating them into a unified multi-modal learning framework. Each experiment involved training deep learning models on specific data representations—images, tabulated features, and text—resulting in a final architecture that combines both image-based and text-derived features for improved prediction accuracy.

4.1 CNN-Based Image Regressor

The first experiment focused solely on the visual modality. A convolutional neural network (CNN) regressor was trained to estimate the age of a bruise from a single RGB image. The input images were resized to 320×320 pixels to ensure a reasonable training computational cost. The architecture consisted of two main parts:

- Feature Extraction Block: A stack of five convolutional layers was used to extract hierarchical spatial features from the input image. These layers transformed the 320×320×3 input into a compact feature representation of 512 dimensions.
- Regression Block: A series of five fully connected (FC) layers was used to map the 512-dimensional feature vector into a single scalar value representing the estimated bruise age in hours.

4.2 Deep Neural Network on Tabular Data

In the second experiment, a deep neural network (DNN) was trained exclusively on structured tabular features associated with each image. From an initial set of 108 recorded attributes, 33 features were selected based on relevance and availability. After applying one-hot encoding to nominal variables, the feature vector length expanded to 66 dimensions. The network architecture consisted of four fully connected layers that performed non-linear transformations of the input, eventually regressing it into a single numeric output—again representing estimated bruise age. This experiment assessed the informativeness of structured physiological and demographic data when used independently of visual inputs.

4.3 BERT-Based Text Feature Regressor

The third experiment aimed to simulate the use of textual data derived from clinical records by transforming structured tabular data into natural language-like input. Each row in the tabular dataset was converted into a string by concatenating textual representations of the feature values. These strings were then passed through a BERT encoder, a smaller and more efficient version of the BERT language model, which encoded the input text into a 256-dimensional embedding vector representing the semantic context of the data. A fully connected neural network with five hidden layers was trained to map these 256-dimensional text embeddings into bruise age estimates. Importantly, this

experiment was conducted under two settings to test the sensitivity of BERT-based embeddings to the order of input features:

- Original Order Setting: Text strings were constructed using feature columns in their original order.
- Shuffled Order Setting: The order of features in each string was randomized prior to encoding.

Separate models were trained for each setting to evaluate the impact of feature ordering on text-based performance.

4.4 Multi-Modal Feature Fusion Model

The final and most comprehensive experiment combined visual and textual representations to capture both surface-level and contextual cues related to bruise healing. This multi-modal model integrated latent features extracted from both image and text modalities:

- • Image Features: An autoencoder was trained on 320×320 images to learn a compressed representation of bruise appearances. The encoder block, composed of five convolutional layers, reduced each image into a 256-dimensional latent vector. The decoder, also with five layers, was used solely during training to reconstruct the original image from its encoded representation.
- • Text Features: As in the previous experiment, BERT-mini was used to encode tabular text into a 256-dimensional embedding. Both original and shuffled input orders were tested, and separate models were trained for each.

The 256-dimensional vectors from the image encoder and BERT encoder were concatenated into a 512-dimensional fused vector. This combined feature representation was passed through a six-layer fully connected regression network to estimate the age of the bruise. By fusing the two modalities, this final model aimed to leverage the complementary strengths of visual cues and physiological context, enabling more accurate and individualized bruise age estimation.

5 Experiments

5.1 Data

The dataset used in this project was collected during a prior research study focused on bruise development and healing across a diverse population [7]. It consists of over 32,000 high-resolution images documenting the progression of bruises over time in 156 participants representing six different skin tones. Each participant underwent controlled bruising using two standardized methods: (1) paintball gunshots to the upper arms and (2) steel ball drops to the lower arms. These controlled methods ensured consistency in bruise formation across the study population, which is crucial for training robust machine learning models.

Each participant was monitored over a four-week period, visiting the lab 21 times for imaging sessions. During each session, multiple images of each bruise were captured under various alternate light source (ALS) and filtering conditions. This resulted in a dataset that captures temporal and spectral variations in bruise appearance, including subtle changes in color, texture, and shape that occur during healing.

In addition to image data, the dataset includes rich textual and tabular metadata for each imaging session. This metadata contains both fixed attributes—such as age, sex, natural hair color, and skin tone—and transient physiological factors—such as body fat percentage, localized skin fat content, bruise size, and bruise color. These contextual factors are known to influence bruise healing. This information was recorded in a structured tabular format.

To prepare this dataset for multi-modal modeling, several preprocessing steps were applied. For tabular data, a subset of 33 features was selected based on relevance, and categorical variables were one-hot encoded, resulting in a 66-dimensional feature vector. For textual modeling, tabular rows were transformed into natural-language phrases and aggregated into input strings suitable for

transformer-based models like BERT. This transformation allowed the model to learn contextual relationships between features, even in the absence of a fixed schema.

The resulting dataset supports various learning tasks: training CNN-based image regressors, deep networks on tabular features, text embedding models using BERT, and fusion models that integrate both modalities.

5.2 Experimental Setup

To ensure reliable evaluation and reproducibility of model performance, all experiments in this study were conducted under controlled and consistent data partitioning strategies and training protocols.

5.2.1 Data Splitting

For most experiments, the dataset was randomly divided into training and testing subsets using an 85-15 split. Specifically:

- Training Set: 85 percent of the dataset was used for model training.
- Test Set: The remaining 15 percent was reserved for evaluating model performance on unseen data.

In experiment involving 10-fold cross-validation, the dataset was divided such that:

- 10 percent of the data was used as a validation set.
- 10 percent was used as the test set.
- The remaining 80 percent served as the training set.

Each fold ensured that samples were randomly selected and stratified as appropriate to maintain a balanced distribution of bruise ages and patient characteristics across all partitions.

5.2.2 Loss Function

Across all models and experiments, the Mean Squared Error (MSE) was consistently used as the loss function. This choice was made due to its effectiveness in penalizing large deviations between predicted and actual bruise ages. The MSE loss is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the true bruise age and $\hat{y}_i$ is the predicted age for the i^{th} sample. This loss function was minimized during training using backpropagation and an optimizer (Adam with 0.001 for learning rate), and served as the primary performance metric throughout training and evaluation.

5.2.3 Baseline Model

The baseline for performance comparison across all experiments was the tabular data regressor, which used selected structured features recorded at each imaging session (such as demographic and physiological data). This model consisted of a deep feedforward neural network trained solely on preprocessed tabular inputs to predict bruise age. It did not incorporate any image or textual data.

All subsequent models—including the CNN-based visual regressor, BERT-based textual model, and the multi-modal fusion network—were evaluated against this baseline to quantify the benefit of incorporating richer modalities of data into the learning process.

5.3 Experimental Results

This section presents the performance outcomes of the various models trained in this project, including single-modality baselines and the final multi-modal fusion model. The evaluation was primarily conducted using Mean Squared Error (MSE) as the performance metric.

5.3.1 CNN-Based Regressor

The initial experiment involved training a convolutional neural network (CNN) regressor using only visual features derived from bruise images. All input images were resized to 320×320 pixels, and the model's predictive performance was evaluated on the test set. As illustrated in Figure 1, the model exhibited no signs of convergence, as neither the training loss nor the test loss demonstrated a consistent downward trend throughout the training process. To investigate whether the lack of convergence was due to insufficient model complexity, a ResNet-18 architecture pretrained on ImageNet was employed. The feature extraction layers were retained with frozen weights, and a fully connected regression head was appended for transfer learning. However, this approach also failed to achieve convergence. These results suggest that relying solely on visual information through CNN-based regression is inadequate for accurate bruise age estimation in this context.

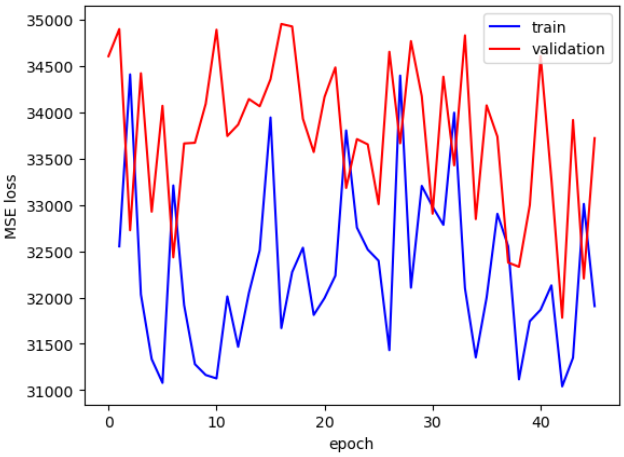

Figure 1: CNN regressor failure to converge

5.3.2 Tabular Deep Neural Network Regressor

The second model was trained using selected structured features extracted from the dataset (e.g., gender, skin tone, body fat percentage). After one-hot encoding categorical variables and normalizing continuous ones, a deep feedforward network was trained. Figure 1 shows loss reduction during the training process of this model. This model is being used as baseline and it achieved MAE of 6.5 hours on test set.

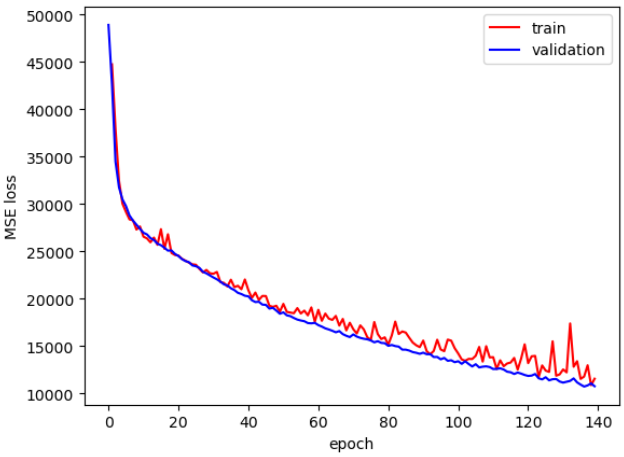

Figure 2: NN model loss reduction

5.3.3 BERT-Based Textual Regression

Two models were trained using BERT embeddings of text strings generated from the tabular metadata:

- Model A: Preserved original column order as shown in figure 3.
- Model B: Used randomly shuffled column order as shown in figure 4.

Both models used BERT-mini to generate a 256-dimensional text embedding, followed by a feedforward network. Models A and B achieved MAEs of 116 hours and 119 hours respectively.

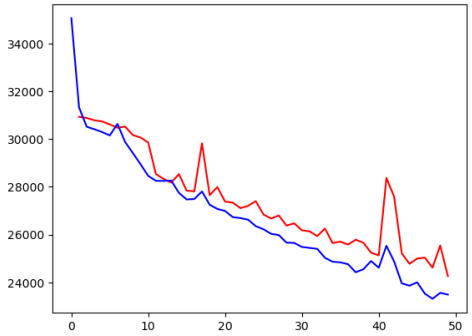

Figure 3: BERT-NN model loss reduction with ordered elements

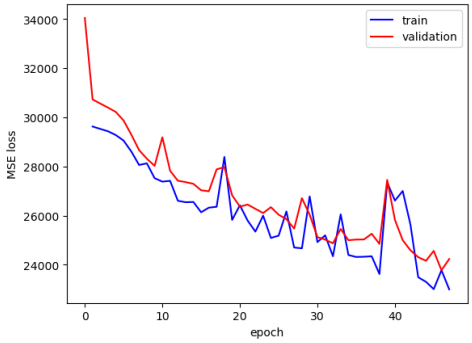

Figure 4: BERT-NN model loss reduction with random elements

5.3.4 Multi-Modal Fusion Model

The final model concatenated 256-dimensional image embeddings and 256-dimensional BERT text embeddings into a 512-length feature vector. This vector was passed through a 6-layer fully connected network to predict bruise age. This variation also had two trained models:

- Model A: Preserved original column order as shown in figure 5.
- Model B: Used randomly shuffled column order as shown in figure 6.

Models A and B achieved MAEs of 50 hours and 55 hours respectively.

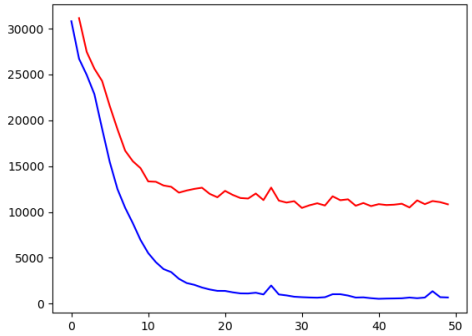

Figure 5: Multi-modal model loss reduction with ordered elements

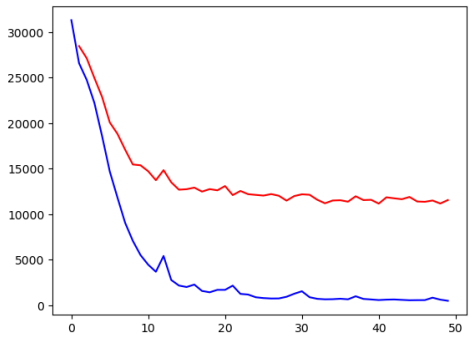

Figure 6: Multi-modal model loss reduction with random elements

5.3.5 Comparison

Performance of different models are sumarized in table 1. It shows that adding text features extracted by BERT can be enhanced with image features captured by convolutional layers.

Table 1: Comparision of different model performances

		NN trained on tabular data	NN trained on BERT features	Multi-modal
MAE	Original text input	1.5 h	116	50
	Random text order		119	55

6 Discussions

This project introduces a novel approach to bruise age estimation by leveraging a multi-modal deep learning framework that integrates both visual and textual data, addressing key limitations of existing methodologies. The originality of the work lies in its methodological innovations and its ability to model bruise healing as a function of both appearance and individual physiological context. The main contributions of the project are outlined as follows:

I Multi-Modal Feature Fusion:

Unlike conventional medical diagnosis models that rely on visual inputs, this project proposes a new architecture that combines features extracted from bruise images with those derived from patient-specific metadata. This fusion enables the model to learn correlations not only within a single modality but also across visual and contextual domains.

II Latent Textual Representation via BERT Embeddings:

The project transforms structured tabular data into natural language text to leverage pretrained BERT embeddings, enabling the model to capture semantic relationships between features in a

context-aware manner. Two variations were tested—preserving the original feature order and using randomized order—to examine BERT's ability to retain relational meaning.

III Image Feature Encoding through Convolutional Autoencoders:
To encode the visual characteristics of bruise healing, a custom autoencoder was trained to compress high-resolution images into a compact 256-dimensional latent space. These latent image vectors capture essential features such as color progression and shape deformation, providing a representation for regression tasks.

IV End-to-End Fusion Model for Bruise Age Estimation:
A deep learning pipeline was developed that concatenates the BERT-generated textual embeddings with the image-derived features, forming a unified 512-dimensional feature vector. This fused representation is then passed through a multi-layer perceptron regressor to estimate bruise age in hours.

V Systematic Experimental Framework:
A series of standalone models were implemented—including CNN-only, tabular DNN, text-based regression, and multi-modal fusion—to rigorously assess the value of each data modality. These controlled experiments not only validate the final architecture but also offer insights into the relative contributions of each feature set.

7 Submission of papers to NeurIPS 2023

This project presented a multi-modal deep learning approach for bruise age estimation by integrating visual and textual data. Traditional bruise assessment methods suffer from subjectivity and disregard individual physiological differences, which limits their reliability in forensic and medical contexts. To overcome these challenges, experiments were done with multiple architectures, including standalone CNN regressors, deep neural networks trained on tabular features, BERT-based text embeddings, and a fused model combining both image and text-derived features.

Through these experiments, we demonstrated that combining features from different modalities enhances prediction accuracy by capturing both the visual progression and patient-specific context of bruise healing.

This project utilized a standardized, laboratory-generated dataset to ensure consistency in bruise formation and controlled imaging conditions. A critical direction for future work involves adapting the proposed model to real-world scenarios by incorporating images of naturally occurring bruises across various body regions and caused by diverse mechanisms of injury. This step is essential for enhancing the model's generalizability and practical applicability in clinical and forensic settings.

Deliverables can be found at the following Github repository. This includes all the codes, benchmark trained models, and a small batch of dataset both in image and table format for the viewer to be able to run the codes.

`https://github.com/mghyabi/multi-modal-bruise-age-estimation`

References

References follow the acknowledgments in the camera-ready paper. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Hughes, V. K., P. S. Ellis & N. E. I. Langlois. (2006) Alternative light source (polilight®) illumination with digital image analysis does not assist in determining the age of bruises. , *Forensic Science International*, pp. 158, 104–107.

[2] Anwar, Syed Muhammad, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, & Muhammad Khurram Khan (2018) Medical image analysis using convolutional neural networks: a review. *Journal of medical systems* 42 (2018): 1-13.

[3] Hayat, Nasir, Krzysztof J. Geras, & Farah E. Shamout (2022) MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. *Machine Learning for Healthcare Conference* pp. 479-503.

[4] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* pp. 4171-4186.

[5] Hao, Boran, Henghui Zhu, & Ioannis C. Paschalidis (2020) Enhancing clinical BERT embedding using a biomedical knowledge base.. *28th international conference on computational linguistics (coling 2020)*.

[6] Zhang, Dongdong, Changchang Yin, Jucheng Zeng, Xiaohui Yuan, & Ping Zhang (2020) Combining structured and unstructured data for predictive models: a deep learning approach. *BMC medical informatics and decision making 20* 1-11.

[7] Scafide KN, Sheridan DJ, Downing N, & Hayat MJ. (2020) Detection of inflicted bruises by alternate light: Results of a randomized controlled trial. *Journal of Forensic Sciences* 65(4):1191-1198.