LUMA: A Benchmark Dataset for Learning from Uncertain and Multimodal Data

Grigor Bezirganyan
grigor.bezirganyan@univ-amu.fr
Aix Marseille Univ, CNRS, LIS
Marseille, France

Sana Sellami
sana.sellami@univ-amu.fr
Aix Marseille Univ, CNRS, LIS
Marseille, France

Laure Berti-Équille
laure.berti@ird.fr
IRD, ESPACE-DEV
Montpellier, France

Sébastien Fournier
sebastien.fournier@univ-amu.fr
Aix Marseille Univ, CNRS, LIS
Marseille, France

Abstract

Multimodal Deep Learning enhances decision-making by integrating diverse information sources, such as texts, images, audio, and videos. To develop trustworthy multimodal approaches, it is essential to understand how uncertainty impacts these models. We propose LUMA, a unique multimodal dataset, featuring audio, image, and textual data from 50 classes, specifically designed for learning from uncertain data. It extends the well-known CIFAR 10/100 dataset with audio samples extracted from three audio corpora, and text data generated using the Gemma-7B Large Language Model (LLM). The LUMA dataset enables the controlled injection of varying types and degrees of uncertainty to achieve and tailor specific experiments and benchmarking initiatives. LUMA is also available as a Python package including the functions for generating multiple variants of the dataset with controlling the diversity of the data, the amount of noise for each modality, and adding out-of-distribution samples. A baseline pre-trained model is also provided alongside three uncertainty quantification methods: Monte-Carlo Dropout, Deep Ensemble, and Reliable Conflictive Multi-View Learning. This comprehensive dataset and its tools are intended to promote and support the development, evaluation, and benchmarking of trustworthy and robust multimodal deep learning approaches. We anticipate that the LUMA dataset will help the research community to design more trustworthy and robust machine learning approaches for safety critical applications. The code and instructions for downloading and processing the dataset can be found at: https://github.com/bezirganyan/LUMA.

CCS Concepts

• **Computing methodologies** → **Computer vision**; **Natural language processing**; **Probabilistic reasoning**; **Supervised learning**; **Uncertainty quantification**; • **Information systems** → **Uncertainty**.

Keywords

multimodal deep learning, uncertainty quantification, dataset

ACM Reference Format:
Grigor Bezirganyan, Sana Sellami, Laure Berti-Équille, and Sébastien Fournier. 2025. LUMA: A Benchmark Dataset for Learning from Uncertain and Multimodal Data. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3726302.3730302

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730302

1 Introduction

In recent years, the use of Machine Learning and Deep Learning has surged across various fields, driving advancements in data analysis and decision-making. In domains such as healthcare, autonomous driving, and finance, information is distributed across multiple modalities including audio, video, text, and images. To better understand the data and improve decision-making capabilities, it is crucial for deep learning models to integrate diverse, multimodal sources of information. Multimodal Deep Learning (MDL) addresses this need and improves the capabilities of uni-modal networks [4, 19, 21, 34].

Another important consideration for deploying deep learning models in safety critical fields is trustworthiness. Traditional deep learning models are often overconfident in their predictions [1], which can lead to catastrophic results in areas such as healthcare or autonomous driving. Although various techniques for uncertainty quantification have been proposed to measure the level of uncertainty in data and model, this remains an open and challenging area. More research and robust benchmarks are needed to advance the field of uncertainty quantification in deep learning [17, 26].

In probabilistic modeling, uncertainty is usually divided into aleatoric (data) and epistemic (model) uncertainties [14]. Aleatoric uncertainty refers to the uncertainty in data due to inherent noise. It is impossible to reduce the amount of aleatoric uncertainty with additional data (hence, it is also often called irreducible uncertainty). Epistemic uncertainty is the uncertainty in model parameters, due to lack of data, hence, it can be reduced with additional data samples. Epistemic uncertainty is also usually high for Out-of-distribution (OOD) data, and is commonly used for OOD detection.

Multimodal uncertainty quantification (MUQ) is a relatively new research area that adapts uncertainty quantification approaches to

multimodal deep learning problems, aiming to enhance the trustworthiness of these models [12]. Due to the unsupervised nature of uncertainty quantification, where the exact extent of uncertainty in the data and the model is unknown, analyzing and benchmarking proposed MUQ methods is challenging. Current multimodal datasets used for benchmarking state-of-the-art models in multimodal uncertainty quantification [8, 11, 12, 22, 35] lack the ability to inject a controlled amount and various types of uncertainties for each modality. This limitation hinders the comprehensive benchmarking of MUQ techniques, which is essential for developing trustworthy and robust multimodal deep learning approaches.

To address this challenge, we introduce LUMA (Learning from Uncertain and Multimodal data), a multimodal dataset specifically designed for benchmarking multimodal learning algorithms on uncertain data. The dataset includes 101,000 images, 135,096 speech audio recordings, and 62,875 text passages, amounting to approximately 3 GB of data. Each modality is independently sourced, reflecting real-world conditions where data is often collected under different conditions and times. For example, in medical contexts, diagnostic data from different modalities such as radiography, MRI, and ECG/EEG are gathered asynchronously, leading to modality-specific uncertainties. The modalities are carefully aligned, ensuring that each text passage is related to the object in the corresponding image, and each audio recording is the pronunciation of the object label in the image. The provided Python toolkit allows the injection of aleatoric and epistemic uncertainties in a controlled and parameterized way into each modality specifically.

To summarize, our contributions are as follows:

(1) We propose LUMA[1], a multimodal dataset specifically designed for learning from uncertain data. It includes audio, image, and textual modalities across 50 distinct classes. We compiled the images from the CIFAR 10/100 dataset [18], extracted, validated, and associated audio samples corresponding to the CIFAR image class labels from three diverse audio corpora [2, 15, 27], and generated the text modality based on the class labels using the Gemma-7B Instruct [25] Large Language Model (LLM). We also performed additional bias analysis of the dataset. Each generated version of the dataset consists of 600 data records per class (500 for training, and 100 for testing) belonging to 42 classes, and 3,859 OOD data points, belonging to the remaining 8 classes.
(2) We offer a Python package[2] that generates dataset samples with varying levels of noise and uncertainty. The uncertainty generator can effectively increase aleatoric uncertainty in the data and epistemic uncertainty in the model.
(3) Finally, we provide baseline models including three different uncertainty quantification methods (Monte-Carlo Dropout [7], Deep Ensemble [20], Reliable Conflictive Multi-View Learning [35]), to serve as a starting point for benchmarking.

2 Limitations of current Datasets for MDL benchmarking

In practice, we often don't know the extent of inherent uncertainties in the data or how accurately they represent the real-world data space. This often makes it hard to evaluate how well uncertainty quantification algorithms work. Moreover, deep learning algorithms may behave differently under different amount of uncertainties (i.e., the robustness to noise may vary). Thus, it may be beneficial to inject additional amount of noise in the data, and observe the change in uncertainty metrics and the performance of the models. Since approaches to quantify different types of uncertainty vary, it is beneficial to have options for injecting various types of uncertainties.

Several datasets are used in multimodal uncertainty quantification settings. A notable line of work [8, 11, 12] has employed datasets such as HandWritten[3], CUB[4], Scene15[5], and Caltech101 [6]. These datasets typically extract different features from unimodal sources to create a multi-view setup. While they have been instrumental, they primarily repurpose unimodal data for multimodal tasks, underscoring the need for more comprehensive and inherently multimodal datasets to better evaluate uncertainty in deep learning models.

Furthermore, the current approaches that introduce uncertainty in the data [8, 11, 12] add Gaussian noise to the views or the extracted features. While Gaussian noise does increase uncertainty, it does not accurately reflect the noise that can be found in real-world datasets and this process lacks fine-grained control over the type of uncertainty being injected.

Additionally, how different modalities' uncertainties interact significantly impacts the overall multimodal uncertainty. When both modalities encode redundant information, the total uncertainty might not decrease. Conversely, conflicting information can lead to increased uncertainty, while complementary information can reduce it. A deeper understanding of these phenomena is crucial. Fine-grained control over individual modalities' uncertainties opens the way for more theoretical research based on empirical observations.

To better understand and analyze uncertain multimodal data, as well as to debug and benchmark uncertainty quantification techniques in the multimodal learning context, we propose a dataset accompanied by an uncertainty generator package. This package includes various techniques for injecting uncertainty, such as controlling data diversity, adding different types of real-world noise, randomly switching labels to their closest class, and injecting out-of-distribution (OOD) data.

3 LUMA Dataset

In this section, we introduce LUMA, a dataset composed of an extensible list of modalities including image, audio, and text modalities, collected from various sources.

3.1 Image modality

For the image modality, our priority was to choose a relatively simple yet well-known dataset, where we could have the option to manually increase the degree of uncertainty. For that purpose, we chose CIFAR-100 and CIFAR-10 [18] datasets since they are

[1] https://huggingface.co/datasets/bezirganyan/LUMA
[2] https://github.com/bezirganyan/LUMA
[3] https://archive.ics.uci.edu/ml/datasets/Multiple+Features
[4] http://www.vision.caltech.edu/visipedia/CUB-200.html
[5] https://serre-lab.clps.brown.edu/resource/hmdb-a-largehuman-motion-database
[6] https://data.caltech.edu/records/mzrjq-6wc02

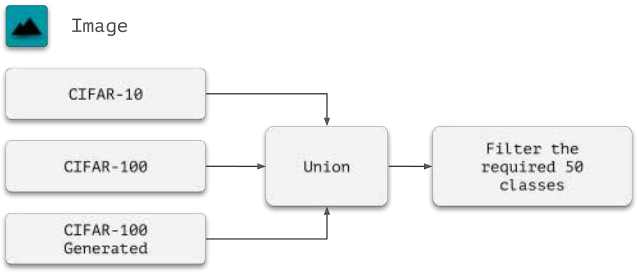

Figure 1: Image collection pipeline

well-known datasets of small 32x32 images, with lots of baseline models. 42 classes were chosen so that after aligning with the other modalities, we would have at least 600 samples in each class per modality. The threshold of 600 classes was selected based on the number of images per class in the CIFAR-100 dataset. We took another 8 classes, which had less than 600 samples after aligning with other modalities, as OOD samples. In total, we took 25,200 images as train/test data, and 3,859 images as OOD data (see the image collection pipeline in Figure 1).

Aside from the main dataset, as described in Section 3.4, another priority was to understand the behaviors of models under different levels of data diversity. To achieve this, we decided to sample 600 data points with different level of diversity from the bigger set CIFAR-10/100. However, in CIFAR-100 dataset, there are no more than 600 samples per class. We alleviated this issue with including images generated with EDM Diffusion-based generative model[7] [13]. We chose EDM-generated images, since the generated samples were already available, and Zheng et al. [36] showed that augmenting CIFAR-10 data with EDM-generated samples significantly improves the classification accuracy.

3.2 Audio modality

For audio modality, the diversity of accent in the pronunciation was an important factor to consider, and we collected samples, where different people would pronounce the corresponding class label of CIFAR 10/100 images. For this task, we used three audio/text parallel corpora and extracted the desired audio segments. More specifically, we used The Spoken Wikipedia [15], LibriSpeech [27], and Mozilla Common Voice [2] corpora. The audio collection pipeline is shown in Figure 2.

The Spoken Wikipedia is a collection of hundreds of hours of phoneme-level aligned audio, where volunteer readers are reading various Wikipedia articles. We used these alignments to extract all the instances of audio segments that pronounced one of the CIFAR-10/100 classes.

The LibriSpeech dataset is a corpus of 1,000 hours of English speech, derived from audiobooks from the LibriVox[8] project, which is a collection of public domain audiobooks. Unfortunately, LibriSpeech doesn't provide word-level alignment, hence, we used force-aligned alignments[9] generated with the Montreal Forced Aligner [24]. Similarly to The Spoken Wikipedia, we looked up the

[7]Retrieved generated samples from https://github.com/wzekai99/DM-Improves-AT
[8]https://librivox.org/
[9]https://github.com/CorentinJ/librispeech-alignments

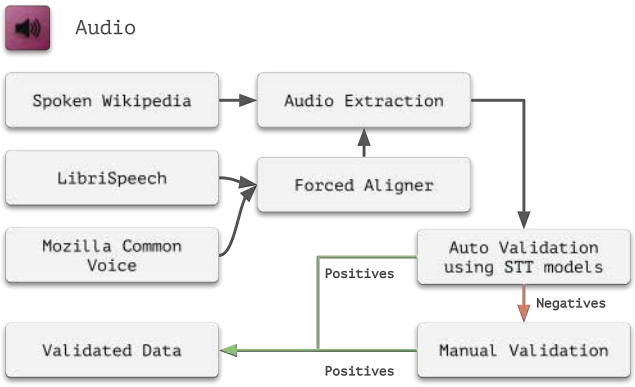

Figure 2: Audio collection, extraction and validation pipeline

CIFAR-10/100 labels in forced aligned textual data, and extracted the corresponding audio segments.

The Mozilla Common Voice corpora, is a crowdsourced open-source collection of voices by volunteer contributors from around the world. Like LibriSpeech, Mozilla Common Voice also doesn't provide word-level alignments, hence, we again used forced aligned alignments[10], and extracted the relevant audio samples.

73 additional recordings of pronunciations belonging to 4 classes ("roses", "telephone", "whale", "wolf") were voluntarily contributed by our colleagues, which were anonymized, trimmed, and added to the dataset.

From these corpora, we used the following rule to extract the samples. First, we extended our class label set with a superset that also contains the plural forms of the words (i.e., for the audio track "horse", the audio track "horses" was added to the set), then we iterated over all aligned transcripts, and for any word included in the formed set, we extracted the corresponding audio sample. We considered the plural forms, since we believe that an extra "s" or "es" does not change the pronunciation of the words much. We did not consider the plural forms if it requires audible changes to the word root (i.e., mouse - mice). Since most of the audio data is collected from forced alignments, it is possible to have misaligned audio segments, which could introduce additional noise to the dataset. Moreover, since part of the audio samples are from voluntary contributions, there can be very noisy samples, which are hard to interpret, or audio samples with a strong accent, which again can be hard to interpret. To remove such extreme cases in audio samples, we performed an automatic validation of the samples. Then, we filtered out the false negatives with manual validation for the negative predictions.

The automatic validation was achieved with the OpenAI's Whisper Large V3 model [29] for audio transcription, and transcribed the extracted audio samples. If the transcription corresponded to the class label (or its plural form), then we considered the sample as valid. Otherwise, the sample was sent for manual validation. Because of the huge output space of the Whisper Large V3 model, the probability of false positives is quite low, so we did not perform a manual validation for positive predictions. To summarize,

[10]https://github.com/JRMeyer/common-voice-forced-alignments

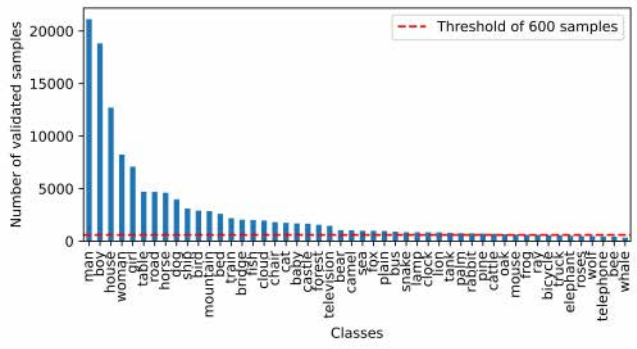

Figure 3: Number of validated audio samples per class. We will include the classes with higher than 600 samples as in-distribution data, and others as out-of-distribution data.

we validated 130,069 out of 178,123 data samples with automatic validation, and we performed a manual validation for the remaining samples.

For manual validation, we decided to check only the classes, which did not have more than 800 samples (to be able to sample 600 samples with different degrees of diversity, as described in Section 3.4). Hence, we filtered 8,372 samples, and scheduled them for manual labeling. We opted for Label Studio [31] to build the labeling interface. The interface provided the audio sample, with the prompt "Is the audio saying the word below? (An extra 's' or 'es' in the pronunciation is okay.)" and answer options of "Yes" or "No". We asked our colleagues (M.Sc. and PhD students, and professors) with advanced to fluent English knowledge to annotate the samples.

In total, we collected 2 annotations per sample, from 17 annotators. We got 71.61% annotation agreement and accepted 5,027 samples, where both annotators confirmed the validity of the sample. Samples with conflicting annotations were rejected. Hence, we took the 42 classes that had more than 600 validated samples (automatically and manually) as training / test data, and the remaining 8 classes as OOD data. In total, the auto-validated and manually validated audio samples combined, LUMA has 135,096 audio samples. The final distribution of audio data across classes can be seen in Figure 3.

3.3 Text Modality

For text modality, the main constraint was that the text segments had to talk about the label of the images. For that, we decided to employ a generative model, and generate text segments about the class label. We utilized Google's Gemma-7B Instruct model [25] to generate more than 1,200 texts samples per class, using 13 different prompts. The 13 prompts are as follows:

```
"You are talking with your friend about some topic.
    Use the word <word> in a sentence with your
    friend. Use casual language. Tone: Casual /
    Conversational, length: short",
"You are the prime minister of the United Kingdom.
    During a press conference you are asked a
    question about <word>. Give a sentence from that
    press conference mentioning the word <word>. Tone
    : Formal, length: medium",
"You are explaining a five year old child what the
    word <word> means. Use very simple and
    explanatory language, so the kid will understand
    the meaning of the word <word>. Tone: Casual,
    complexity: simple",
"Imagine you are writing a science fiction book.
    Write a conversation from that book mentioning
    the <word>.",
"You are the editor in a mainstream journal. Write a
    sentence from a news article about a <word> in
    your journal that mentions the word <word>.",
"You are a teenager writing a post in Facebook about
    <word>. Write the post about the experience you
    had with the <word>.",
"You are playing a word describing game with your
    friend. The word is <word>, and you shall
    describe it without mentioning the word itself,
    so your friend will guess it. Explain it to him
    clearly in a simple language.",
"Think of something else that shares similar
    characteristics or functions with the <word>.
    Draw comparisons or use analogies between that
    other word and the <word>.",
"Place the word <word> within historical context. How
     would you describe it in relation to its origins
    , evolution, or significant historical events? Be
     creative in your description.",
"Consider how the word '<word>' is depicted or
    referenced in popular culture, literature, or
    media. Describe it by referencing these cultural
    elements.",
"Pretend you are a character who sees a <word> for
    the first time in your life. Describe it from the
     character's perspective, considering their
    background, personality, and knowledge.",
"Write a 4 line small poem about the word <word>. Be
    creative, and use casual tone for the poem.",
"You are a musician composing a song inspired by <
    word>. Write the lyrics to the song, capturing
    the mood, emotions, and imagery associated with <
    word>. Use rhythm and melody to convey the
    essence of <word> in your music."
```

The <word> was replaced with the class labels. Gemma-7B Instruct was chosen, since according to their technical report [25], it outperforms other open LLMs with similar size, in 11 out of 18 tasks. Moreover, in our experiments, it provided responses that we found more useful for our intended applications than those from Mistral-7B [10].

To validate that the generated texts accurately represent the labels, we masked all label occurrences in the text and fed the masked text back into the Gemma-7B Instruct model, asking it to

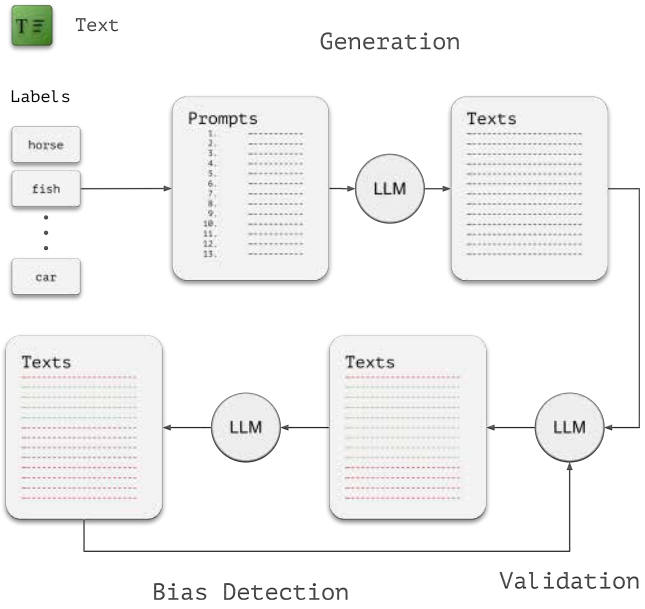

Figure 4: Text generation and validation pipeline

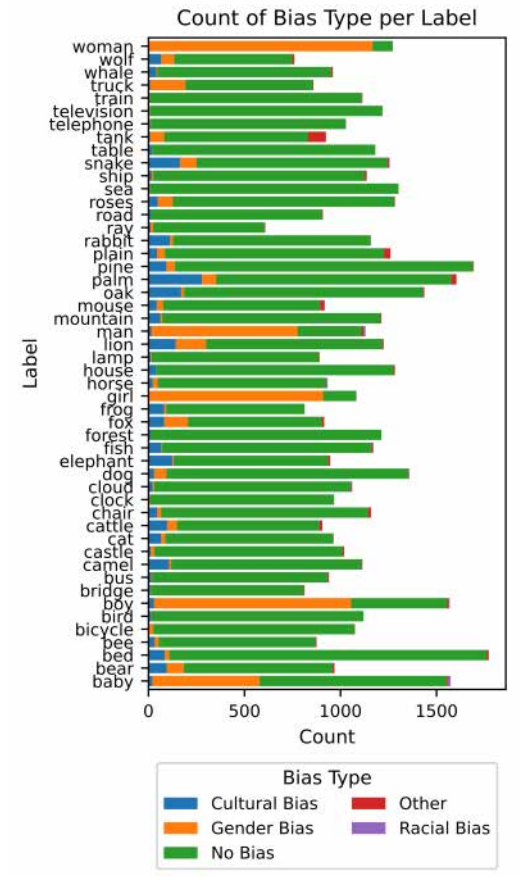

Figure 5: The amount of texts with different biases according to Gemma-7B Instruct model.

classify the text into one of the labels. Based on the prediction of the model, if the prediction matched the ground truth label, we accepted the sample as validated. In total, we accepted 55,953 text samples.

After manually analyzing some of the generated texts, we noticed that there were samples with offensive biases and stereotypes, such as:

```
A woman is a grown-up person who has a soft,
    nurturing personality. She usually takes care of
    her family and friends, and sometimes works
    outside the home. Women are strong and smart,
    they can do many things that men can do.

A tool in a toolbox is an efficient and valuable
    asset that aids in various tasks. Similarly, a
    man is also a valuable asset to any group or
    society. Just like a tool in a toolbox, a man's
    capabilities are tailored to fulfill different
    roles and functions, making him an essential
    component of any endeavor.
```

Particularly, we noticed a lots of gender bias for classes "man", "woman", "boy" and "girl". To find the proportion of the biased data, we asked the Gemma model to find out if the given text contains gender, racial, religious, or cultural biases. We found out, that indeed, the aforementioned 4 classes have a huge amount of gender bias (See Figure 5). Our hypothesis is, that describing a man or woman in an unbiased way is a challenging task for LLM models (as well as for humans), which are trained on unbalanced data [16]. Although the model identified a high level of bias in these classes, indicating a potential for LLM-based bias detection, we are unable to assess how accurate or consistent these detections truly are.

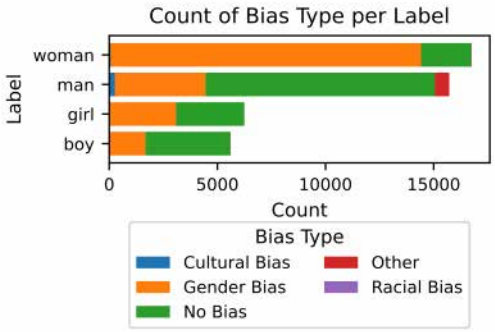

Figure 6: The amount of texts with different biases according to Gemma-7B Instruct model, after reconstructing the prompts for these 4 labels. Although there is still high amount of bias, we can filter them out and still have enough unbiased texts (more than 800 text passages per class).

To reduce the biases for the labels of `man`, `woman`, `boy` and `girl`, we reconstructed the prompts to explicitly provide topics and keywords with occupations, which will minimize the bias. The prompt used is the following:

```
Write two sentences with topic: <topic>, and keywords:
     <keyword1>, <keyword2>.
```

where <topic> was replaced with one of the following words: 'factual', 'fiction', 'history', 'books', 'movies', 'philosophical', and <keyword1> with one of the following words: 'man', 'woman', 'boy', 'girl'. <keyword2> was treated differently depending on the label: For the term 'man', it was replaced with one of the following words:

```
actor, king, scientist, doctor, wizard, duke, lord,
    governor, prime minister, father, sorcerer,
    waiter, chess, director, producer, uncle, singer
```

For the term woman, it was replaced with one of the following words:

```
actress, queen, scientist, doctor, witch, duchess,
    lady, governor, prime minister, mother, sorcerer,
     waitress, chess, director, producer, aunt,
    singer
```

For the term boy, it was replaced with one of the following words:

```
kid, actor, prince, son, nephew, pupil, student,
    singer
```

And for the term girl, it was replaced with one of the following words:

```
kid, actress, princess, daughter, niece, pupil,
    student, singer
```

We then performed another round of bias detection using the Gemma model. While we found that a significant amount of bias still exists, we identified enough unbiased texts (according to the Gemma model) to include in the LUMA dataset. The number of biased and unbiased texts after re-generating the data for these 4 classes can be found in Figure 6.

Since textual data was generated using an LLM, we recognize that the dataset may contain factual inaccuracies, or biases, but our aim is to offer a benchmark to study uncertainty quantification in multimodal classification settings. LUMA shall not be used as a source of knowledge or information.

3.4 Dataset compilation

Based on the collected samples from the 3 modalities (image, text, audio), we first compiled a clean version of the dataset with minimal uncertainty. Our goal was to then provide tools that allow uncertainty to be introduced on demand. We prioritized flexibility, offering multiple options for controlling uncertainty through adjustable parameters such as data diversity, sample-level noise, label corruption, and the injection of out-of-distribution (OOD) samples.

3.4.1 Data Diversity. With a fixed number of data points, increasing the diversity of data enhances the information passed to the model; thereby, it shall reduce the epistemic uncertainty. Conversely, when samples are concentrated at a single point in the latent space, they encode less information, which shall lead to greater epistemic uncertainty in areas where data is scarce (Figure 8). Hence, controlling the diversity of the data allows us to study the behavior of epistemic uncertainties under varying amounts of information.

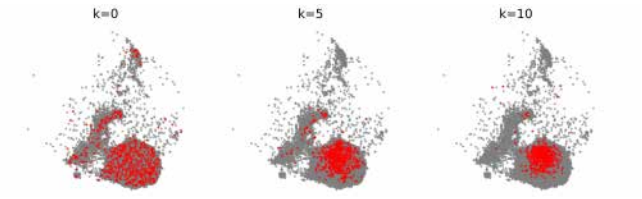

Figure 7: t-SNE [33] visualization of audio data points from class "man" (in gray), and sampled points with different diversity parameter k. The higher is the value of k, the more concentrated (less diverse) the points are.

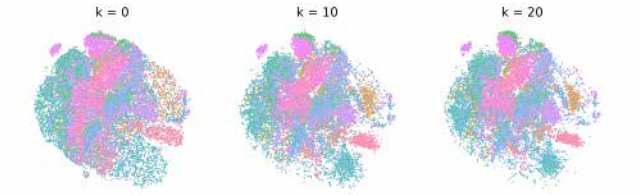

Figure 8: t-SNE [33] visualization of audio data points for all classes, sampled with different diversity parameter k. With higher k we have more concentrated samples, and more separation between classes. The diversity can similarly be controlled for the other modalities.

To control the diversity, we extract deep features from each modality (Wav2Vec [3] for audio, BERT [5] for text and VGG-11 [30] for images), and compute the inverse distance of each sample to the center (mean vector) of its class, raised to the power of k:

$$D_i = \frac{1}{\left\| F_i - \frac{1}{|C|} \sum_{j \in C} F_j \right\|_2^k}, \quad i \in C, \tag{1}$$

where F represents the deep feature vectors extracted from the samples, C is the set of data indices belonging to the class, and $|\cdot|$ measures the cardinality of the set. Then, having the inverse distances, we sample points from the categorical distribution $x_n \sim \text{Cat}(D)$. In Eq. 1, k is the variable controlling the diversity. If $k = 0$ the sampling is uniform. The bigger k, the higher probability of selection will be applied to the samples closer to the center.

Having sufficient samples in image and text modalities, our bottleneck was the number of samples in the audio modality. Since in the 42 in-distribution classes, around 70% have more than 900 audio samples, we considered this enough for diversity control.

3.4.2 Sample Noise. We want to have an option to inject a controlled amount of noise into the data. This may reduce the information in each data sample and increase the classification difficulty. With our hypothesis, this may affect the aleatoric uncertainty degrees. This type of noise can also be very beneficial for estimating the model's robustness to noise. We apply different types of noise to each modality.

For **audio modality**, we added background noise (Human non-speech sounds, urban noises, animal sounds, natural soundscapes,

etc.) from the ESC-50 dataset [28] to each sample, using the `audiomentations`[11] library. The amount of the minimum and maximum signal-to-noise ratio, as well as the proportion of the noisy data, is set as a hyper-parameter.

For **text modality**, we utilize the `nlpaug` [23] library, to add different types of noise. The user has the option to choose a subset of noise types from: 1) Keyboard noise that simulates keyboard distance error; 2) OCR noise that simulates OCR engine noise; 3) Random character noise to insert, substitute, or delete random characters; 4) Antonym noise to swap random words with their antonyms; 5) Random word noise to insert, substitute, or delete random words; 6) Spelling noise to add spelling mistakes according to the spellingling mistake dictionary; and 7) Back-translation noise to translate the text to another language, and then translate back to English. The parameters of these noise types can be specified by the user, and are transferred to the `nlpaug` library for adding the specific type of noise.

For **image modality**, we added different types of noise suggested and implemented by Hendrycks and Dietterich [9]. 15 perturbations are included such as: adding Gaussian noise, shot noise, impulse noise, defocus blur, frosted glass blur, motion blur, zoom blur, snow, frost, fog, changing the brightness, contrast, elasticity, pixelating, and JPEG compressing. Additionally, common transformations such as cropping or skewing are not part of the default set but can be easily applied by the user outside the framework if needed.

3.4.3 Label Noise. Aleatoric uncertainty can also be introduced by injecting label noise into the dataset, i.e., by randomly altering the labels of certain samples. Since this type of uncertainty arises from inherent noise in the data itself, it cannot be mitigated by increasing the dataset size and thus directly contributes to the model's aleatoric uncertainty.

To inject label noise in a more structured manner, we randomly select a subset of samples based on a user-defined probability. For each selected sample, we compute its deep feature representation using modality-specific encoders: Wav2Vec [3] for audio, BERT [5] for text, and VGG-11 [30] for images. We then measure the sample's average distance to the five nearest samples from each class. The sample is reassigned the label of the class with the lowest mean distance, ensuring the new label is still semantically or perceptually close to the original, thereby creating realistic label noise.

3.4.4 OOD Injection. Ideally, the models shall be uncertain on data points from unknown distribution (i.e., distribution they haven't been trained on). In the literature, often the OOD samples are taken from another dataset, which can simplify the problem, because such samples are far from the training data. For this matter, we kept a separate set of samples from the same dataset, but belonging to classes that are not present in the training data, as OOD samples.

4 Baseline models with Uncertainty Quantification

We develop baseline models with three different uncertainty quantification algorithms, to serve as a starting point for other research and benchmarking initiatives. For the sake of simplicity, we choose late fusion approaches, where we have classification networks for each modality, and then fuse their decisions by simply averaging the output logits. These baselines were selected to instantiate unimodal and multimodal architectures, which can be trained on the dataset and are not intended to serve as a comprehensive benchmark, nor did we endeavor to achieve the best possible performance. The architectures of the baseline models are depicted in Figure 9.

For the image modality, we used a simple convolutional neural network. For the audio modality, we extracted 128x128 mel-spectrograms from padded audio samples, and used a convolutional network for classification. For the text modality, we extracted the BERT [6] embeddings for each token, and averaged them out, so that we have one embedding per text passage. Then, we passed the embedding through a simple feed-forward neural network to get the predictions. As depicted in Figure 9, each model includes two output heads: one for the prediction and the other for aleatoric uncertainty, following the methodology outlined by Valdenegro-Toro and Mori [32]. Then, to combine the aforementioned unimodal networks into a multimodal architecture, we adopted the late fusion approach. In the Monte Carlo Dropout and Deep Ensemble methods, we obtained the multimodal prediction by averaging the logits from the final layers of the classifiers. For the Reliable Conflictive Multi-View Learning (RCML) [35], we modified the output of the last layer in each network to produce evidence, as described in [35], and followed their methodology for combining the evidence.

The dropout probability is 0.3, with the deep ensemble comprising 10 networks. Networks were trained for up to 300 epochs, with early stopping after 10 epochs of no validation loss improvement.

4.1 Uncertainty Metrics

For uncertainty quantification, we implemented 3 approaches: Monte Carlo Dropout (MCD) [7], Deep Ensemble (DE) [20], Reliable Conflictive Multi-View Learning (RCML) [35]. In Monte Carlo Dropout and Deep Ensembles, we use the aleatoric entropy and the epistemic entropy as uncertainty measures, and follow Valdenegro-Toro and Mori [32] for disentangling the aleatoric and epistemic uncertainties. $H_{\text{Ale}}(y \mid \mathbf{x}) = \text{entropy}(p_{\text{Ale}}(y \mid \mathbf{x}))$ and $H_{\text{Epi}}(y \mid \mathbf{x}) = \text{entropy}\left(p_{Epi}(y \mid \mathbf{x})\right)$, where p_{Epi} and p_{Ale} are the probabilities obtained according to [32]. For RCML [35], we measure the aleatoric uncertainty with the expected entropy:

$$\mathbb{E}_{p(\boldsymbol{\pi}|\mathbf{x},\hat{\boldsymbol{\theta}})}[H[P(y \mid \boldsymbol{\pi})]] = -\sum_{k=1}^{K} \frac{\alpha_k}{\alpha_0} \left(\psi(\alpha_k + 1) - \psi(\alpha_0 + 1)\right), \quad (2)$$

where α_k is the k-th concentration parameter of the Dirichlet distribution, and α_0 is the sum of all concentration parameters. ψ is the digamma function. As a measure for epistemic uncertainty, we take $\frac{N}{\alpha_0}$, where N is the number of classes. We evaluate the measures of accuracy and uncertainty of the models on the clean dataset, the dataset with reduced diversity, the dataset with increased sample noise, and the dataset with increased label noise[12].

4.2 Results

The results are summarized in Table 1. Since uncertainty is quantified differently in RCML than in MCD and DE, their values cannot

[11] https://github.com/iver56/audiomentations

[12] For noise generation parameters, please refer to our GitHub page

Table 1: Results for UQ with baseline models. The absolute values are reported for clean dataset, and changes in percentages relative to clean dataset are reported for the noisy versions of LUMA dataset.

Method	Clean Dataset		Reduced Diversity		Increased Label Noise		Increased Sample Noise	
	Ale.	Epi.	Ale.	Epi.	Ale.	Epi.	Ale.	Epi.
MCD Image	1.00	1.03	-15.73%	-11.66%	+59.20%	+54.51%	+4.44%	+2.18%
MCD Audio	0.52	0.70	-5.54%	+2.16%	+96.63%	+54.49%	+23.12%	+14.40%
MCD Text	0.37	1.01	-3.91%	-2.62%	+93.59%	+2.41%	+64.96%	-2.03%
MCD Multi.	0.26	0.78	-8.52%	-1.21%	+122.44%	+11.60%	+59.14%	+9.89%
DE Image	1.45	1.40	-37.49%	-8.54%	-7.43%	+0.24%	-18.46%	-3.22%
DE Audio	0.56	0.99	-27.39%	-3.34%	**+156.40%**	+50.43%	**+70.26%**	+34.41%
DE Text	0.42	1.01	+5.02%	-6.15%	+81.26%	-0.51%	+62.24%	-7.11%
DE Multi.	0.31	0.82	-22.80%	-3.40%	+115.15%	+20.62%	+45.97%	+5.54%
RCML Multi.	1.99	0.43	**+8.34%**	**+16.16%**	+64.72%	**+106.16%**	+36.19%	**+58.21%**

Table 2: OOD Detection AUC Values for Different Methods

Method	MCD				DE				RCML Multi.
	Image	Audio	Text	Multi.	Image	Audio	Text	Multi.	
AUC	0.54	0.47	0.53	0.50	0.54	0.49	0.54	0.50	**0.91**

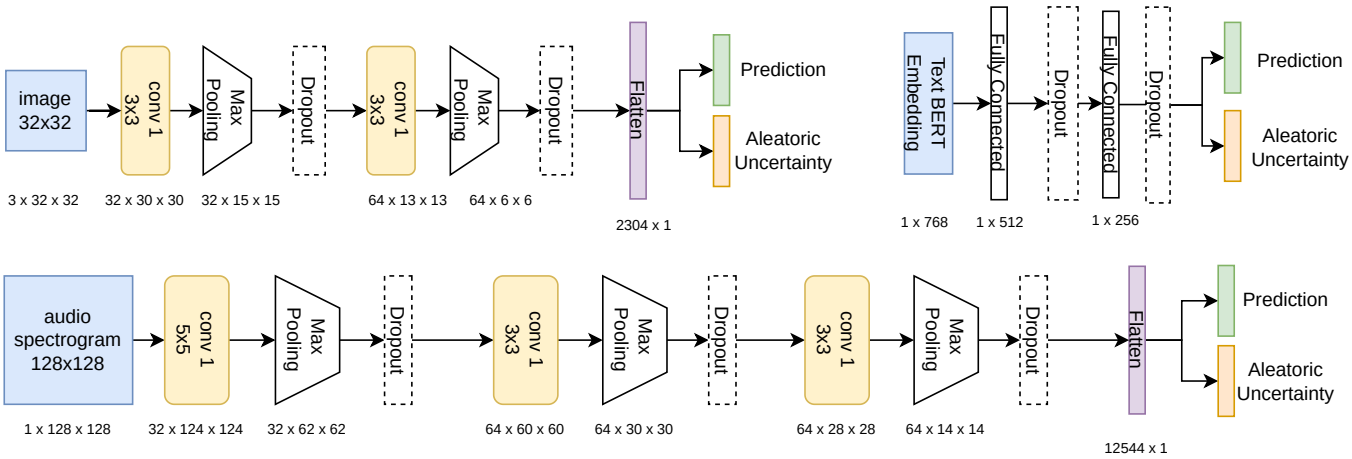

Figure 9: Network architectures used for image, audio and text modalities.

be directly compared, so the table reports the relative changes in both types of uncertainty with respect to the clean dataset.

As we can observe in the table, in most cases, adding label and sample noises effectively increases the epistemic and aleatoric uncertainties. Interestingly, in most MCD and DE models, the uncertainty decreases when they are trained on data with lower diversity. This may indicate that these approaches fail to recognize data points outside their training distribution, which we will further investigate with the OOD detection task.

We evaluate AUC score for OOD detection based on the epistemic uncertainty. The results are summarized in Table 2. We can see that Monte Carlo Dropout and Deep Ensembles fail to provide epistemic uncertainty values suitable for OOD detection in LUMA dataset, with a poor performance of approximately 0.5 AUC value. On the other hand, the RCML achieves an outstanding AUC score of 0.91, indicating that the epistemic uncertainty values quantified with this method can be effectively used for OOD detection.

To further evaluate the qualities of the uncertainties of the different models, we estimate the epistemic and aleatoric uncertainties under different amounts of label noise. Ideally, we expect a good uncertainty quantification algorithm to provide higher uncertainty values for more noisy data. As we can see from Figure 10, only RCML consistently raises the uncertainty estimates under increased label noise, which again shows the higher quality of its uncertainty estimates over the other baselines.

In Table 3, we present the classification accuracy measures for the clean dataset and variations in accuracy under different types of noise for the three baseline models across each modality. We

Table 3: Classification accuracies for the clean dataset and variations in accuracy under different types of noise.

Model	Clean Dataset	Reduced Diversity	Increased Label Noise	Increased Sample Noise
	Accuracy	Difference in accuracy from the clean dataset		
MCD Image	0.335	+0.058	-0.306	+0.019
MCD Audio	0.867	-0.025	-0.784	-0.155
MCD Text	0.965	-0.027	-0.864	-0.144
MCD Multi.	0.991	-0.010	-0.874	-0.063
DE Image	0.387	+0.066	-0.166	+0.019
DE Audio	0.912	-0.003	-0.809	-0.149
DE Text	0.973	-0.023	-0.864	-0.125
DE Multi.	0.996	-0.006	-0.849	-0.042
RCML Multi	0.973	-0.128	-0.833	-0.148

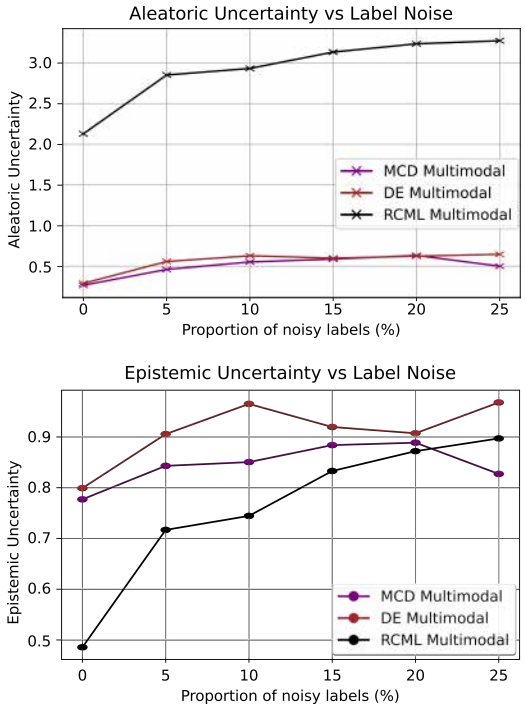

Figure 10: Changes in uncertainty estimations under different proportions of label noise (in percentages). The RCML approach consistently increases both aleatoric and epistemic uncertainties with increased label noise. In contrast, the MCD and DE models sometimes fail to increase the corresponding uncertainty estimations in this experiment.

observe that accuracy always decreases with increasing the label noise, but reducing diversity and increasing sample noise may not always decrease accuracy in the image modality.

In conclusion, the performance of Monte Carlo Dropout and Deep Ensembles indicates a limitation in their suitability for OOD detection in LUMA dataset. This suggests new avenues for further exploratory research to leverage uncertainty estimation for robust detection of out-of-distribution samples. Furthermore, the observed disparities highlight the necessity for a comprehensive benchmarking effort on LUMA dataset, encompassing a broader array of state-of-the-art methodologies.

5 Conclusion

In this paper, we propose LUMA, a multimodal dataset for learning from uncertain and multimodal data. The dataset spans image, audio, and text modalities, and is accompanied by a Python package that allows users to generate customized versions with varying levels and types of noise and uncertainty. To support future research, we also provide a suite of baseline implementations for performance comparison. The dataset can be easily extended with additional modalities and augmented with more data samples. In the future, we plan to include dependent modalities, enabling more comprehensive studies of uncertainty quantification as well as improved applicability in information retrieval contexts. The open-source nature of the data compilation pipeline and code for uncertainty and noise generation facilitates the integration of new contributions from the community to promote multimodal uncertainty studies and benchmarking initiatives.

Acknowledgments

We thank our colleagues and friends for contributing to the audio validation process, and to the HumanSignal team for providing an academic license for Label Studio.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion* 76 (2021), 243–297.

[2] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. 4211–4215.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[4] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. 2022. A Survey on Deep Multimodal Learning for Computer Vision: Advances, Trends, Applications, and Datasets. *The Visual Computer* 38, 8 (Aug. 2022), 2939–2970. doi:10.1007/s00371-021-02166-7

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Association for Computational Linguistics, 4171–4186. doi:10.18653/V1/N19-1423

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, 4171–4186. doi:10.18653/V1/N19-1423

[7] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. PMLR, 1050–1059.

[8] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2023. Trusted Multi-View Classification With Dynamic Evidential Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 2 (2023), 2551–2566. doi:10.1109/TPAMI.2022.3171983

[9] Dan Hendrycks and Thomas Dietterich. 2018. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*.

[10] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[11] Myong Chol Jung, He Zhao, Joanna Dipnall, and Lan Du. 2023. Beyond Unimodal: Generalising Neural Processes for Multimodal Uncertainty Estimation. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 42191–42216. https://proceedings.neurips.cc/paper_files/paper/2023/file/839e23e5b1c52cfd1268f4023a3af0d6-Paper-Conference.pdf

[12] Myong Chol Jung, He Zhao, Joanna Dipnall, Belinda Gabbe, and Lan Du. 2022. Uncertainty estimation for multi-view data: the power of seeing the whole picture. *Advances in Neural Information Processing Systems* 35 (2022), 6517–6530.

[13] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. In *Proc. NeurIPS*.

[14] Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or Epistemic? Does It Matter? *Structural Safety* 31, 2 (March 2009), 105–112. doi:10.1016/j.strusafe.2008.06.020

[15] Arne Köhn, Florian Stegen, and Timo Baumann. 2016. Mining the Spoken Wikipedia for Speech Data and Beyond. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (Portorož, Slovenia, 23-28), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Paris, France.

[16] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*. 12–24.

[17] Ranganath Krishnan and Omesh Tickoo. 2020. Improving model calibration with accuracy versus uncertainty optimization. *Advances in Neural Information Processing Systems* 33 (2020), 18237–18248.

[18] Alex Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images. (2009).

[19] Felix Krones, Umar Marikkar, Guy Parsons, Adam Szmul, and Adam Mahdi. 2024. Review of multimodal machine learning approaches in healthcare. *ArXiv* abs/2402.02460 (2024). https://api.semanticscholar.org/CorpusID:267412288

[20] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017).

[21] Sang Il Lee and Seong Joon Yoo. 2020. Multimodal deep learning for finance: integrating and forecasting international stock markets. *The Journal of Supercomputing* 76 (2020), 8294–8312.

[22] Wei Liu, Xiaodong Yue, Yufei Chen, and Thierry Denoeux. 2022. Trusted Multi-View Deep Learning with Opinion Aggregation. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 7 (June 2022), 7585–7593. doi:10.1609/aaai.v36i7.20724

[23] Edward Ma. 2019. NLP Augmentation. https://github.com/makcedward/nlpaug.

[24] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*. 498–502. doi:10.21437/Interspeech.2017-1386

[25] Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, and et al. 2024. Gemma. (2024). doi:10.34740/KAGGLE/M/3301

[26] Zachary Nado, Neil Band, Mark Collier, Josip Djolonga, Michael W Dusenberry, Sebastian Farquhar, Qixuan Feng, Angelos Filos, Marton Havasi, Rodolphe Jenatton, et al. 2021. Uncertainty baselines: Benchmarks for uncertainty & robustness in deep learning. *arXiv preprint arXiv:2106.04015* (2021).

[27] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR Corpus Based on Public Domain Audio Books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, South Brisbane, Queensland, Australia, 5206–5210. doi:10.1109/ICASSP.2015.7178964

[28] Karol J Piczak. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*. 1015–1018.

[29] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 28492–28518. https://proceedings.mlr.press/v202/radford23a.html

[30] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. http://arxiv.org/abs/1409.1556

[31] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. https://github.com/heartexlabs/label-studio Open source software available from https://github.com/heartexlabs/label-studio.

[32] Matias Valdenegro-Toro and Daniel Saromo Mori. 2022. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 1508–1516.

[33] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

[34] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M López. 2020. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems* 23, 1 (2020), 537–547.

[35] Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. 2024. Reliable Conflictive Multi-View Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 14 (March 2024), 16129–16137. doi:10.1609/aaai.v38i14.29546

[36] Chenyu Zheng, Guoqiang Wu, and Chongxuan Li. 2024. Toward understanding generative data augmentation. *Advances in Neural Information Processing Systems* 36 (2024).