Engineering Applications of Artificial Intelligence 139 (2025) 109643

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Multimodal transformer for early alarm prediction

Nika Strem [a,*], Devendra Singh Dhami [d], Benedikt Schmidt [c], Kristian Kersting [a,b]

[a] TU Darmstadt, Hochschulstr. 1, 64289, Darmstadt, S1/03 - 073, Germany
[b] Hessian Center for Artificial Intelligence (hessian.ai), Germany
[c] ABB AG Corporate Research Center Germany, Kallstadter Straße 1, Mannheim, 68309, Germany, Germany
[d] Eindhoven University of Technology, 5612 AZ Eindhoven, Netherlands

ARTICLE INFO

Keywords:
Multimodal transformer
Multimodal fusion
Industrial processes
Alarm management
Alarm prediction

ABSTRACT

Alarms are an essential part of distributed control systems designed to help plant operators keep the processes stable and safe. In reality, however, alarms are often noisy and thus can be easily overlooked. Early alarm prediction can give the operator more time to assess the situation and introduce corrective actions to avoid downtime and negative impact on human safety and environment. Existing studies on alarm prediction typically rely on signals directly coupled with these alarms. However, using more sources of information could benefit early prediction by letting the model learn characteristic patterns in the interactions of signals and events. Meanwhile, multimodal deep learning has recently seen impressive developments. Combination (or fusion) of modalities has been shown to be a key success factor, yet choosing the best fusion method for a given task introduces a new degree of complexity, in addition to existing architectural choices and hyperparameter tuning. This is one of the reasons why real-world problems are still typically tackled with unimodal approaches. To bridge this gap, we introduce a multimodal Transformer model for early alarm prediction based on a combination of recent events and signal data. The model learns the optimal representation of data from multiple fusion strategies automatically. The model is validated on real-world industrial data. We show that our model is capable of predicting alarms with the given horizon and that the proposed multimodal fusion method yields state-of-the-art predictive performance while eliminating the need to choose among conventional fusion techniques, thus reducing tuning costs and training time.

1. Introduction

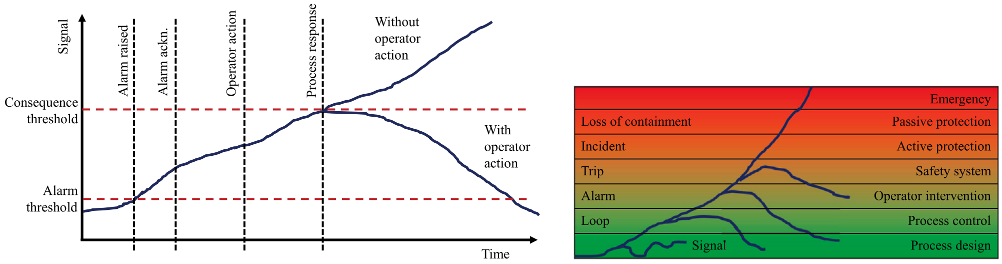
Current trends in deep learning are largely associated with multimodal learning, which implies merging heterogeneous modalities to leverage implicit correlations between multiple sources of information as well as additional features contained in individual modalities to improve the representation learning capacity of the model. Such approaches have shown impressive outcomes across multiple applications (Jabeen et al., 2023). In particular, for state-of-the-art Transformer models (Vaswani et al., 2017) it has been shown that multimodal attention is a significant factor contributing to a model's performance, more than other aspects such as depth or dimensionality (Hendricks et al., 2021).

Yet the method of combining modalities makes a difference and, depending on the dataset and the task, either early or late fusion can yield better results (Snoek et al., 2005; Perez-Rua et al., 2019; Boulahia et al., 2021). Studies like Ma et al. (2022) stress that the optimal fusion is dataset dependent even for the same Transformer model and that no fusion method works best in all cases. Thus, in addition to an already large search space of hyperparameters for the Transformer architecture, including embedding dimensionality, number of heads, MLP ratio, and network depth, as well as type of positional embeddings (Chen et al., 2021b; Chitty-Venkata et al., 2022), another architectural choice becomes necessary.

The multimodal fusion research for the most part has been using large Transformer models trained on huge, curated datasets, with the vast majority of implementations focused on combining common modalities such as image, text and audio (Rahman et al., 2020; Akbari et al., 2021; Chen et al., 2021a). At the same time, despite the success of multimodal machine learning methods in general and Transformers in particular in academic research, to the best of our knowledge, there have been no attempts yet to explore its potential in the industrial domain, where models need to be reasonably compact and where data is noisy and highly unbalanced. This is a glaring omission, because, on the one hand, accurate predictions made by data-driven models can be

[*] Corresponding author.
E-mail addresses: nika.strem@cs.tu-darmstadt.de (N. Strem), d.s.dhami@tue.nl (D.S. Dhami), benedikt.schmidt@de.abb.com (B. Schmidt), kristian.kersting@cs.tu-darmstadt.de (K. Kersting).

https://doi.org/10.1016/j.engappai.2024.109643
Received 15 October 2023; Received in revised form 16 September 2024; Accepted 6 November 2024
Available online 30 November 2024



(a) Alarm response timeline (adapted from IEC (2014))

(b) Layers of protection (adapted from Stauffer and Clarke (2016))

Fig. 1. Alarms notify operators of an equipment failure or process deviation requiring an intervention to restore normal operation. Alarm thresholds are defined taking into account such factors as operator response time, process schedule and response time, as well as severity of potential consequences. Without operator's response, an anomaly can result in equipment damage and downtime, or even major hazards, or a plant shutdown.

of great value for many industrial tasks, and, on the other, the amount of data logged by modern plants is by now sufficient to enable the training of state-of-the-art deep learning models, thereby paving the way for industrial-scale multimodal learning.

In the industry, increasingly comprehensive data tracking can be overwhelming for plant operators, who need to monitor numerous process indicators to ensure normal operation and process safety. Yet, while confusing to a human, the abundant data presents a unique opportunity for data-driven solutions. Whereas incipient changes in a system leading to abnormal behavior cannot be noticed by the operator, a machine learning model trained on historic data could capture characteristic patterns and interactions among various signals and events. Such a model can be used to predict deviations before they become obvious to give the operator more time to resolve the issue and, in addition, to help identify the cause of the problem.

In practice, since the number of measurements to track and assess is overwhelming for a human, process monitoring systems incorporate alarms to indicate an impending critical situation, equipment malfunction, process deviation, or abnormal condition requiring a timely response from the operator (IEC, 2014). Alarms are defined based on thresholds that may not be exceeded by individual sensor measurements, such as temperature, flow, level or pressure. When an alarm is raised, the operator must take a corrective action, such as opening or closing a valve, or changing a setpoint value to bring the process back to normal conditions (Fig. 1(a)). If the operator fails to respond timely to prevent further deterioration of the situation, this can lead to failures, damage of equipment, downtime and hazards to employees (Fig. 1(b)). In reality, however, due to complex interactions in production processes, which are too hard to capture with hand-crafted rules, alarms are often too noisy and thus can be easily overlooked by operators. Trained on historic data and relevant alarms, a machine learning model could capture complex interactions in a process and enable early alarm prediction to give the operator more time to react and introduce corrective actions to avoid downtime and negative impact on human safety and the environment.

The two major data sources that could be used include *signals* (measurements of physical quantities like temperatures, flows, pressures) and *events* (changes automatically registered in the system or introduced by an operator). Existing approaches to early alarm prediction usually concentrate on a particular alarm type and rely on signals corresponding to the alarm in question, which allows to predict only a small subset of problems (Li et al., 2013; Langone et al., 2014; Proto et al., 2019; Koltsidopoulos Papatzimos et al., 2019; Chatterjee and Dethlefs, 2020; Villalobos et al., 2021). Other studies predict alarms based on past alarms, which does not add the value of an early warning (Zhu et al., 2016; Cai et al., 2019; Wang and Liang, 2020). By contrast, we propose to train a multimodal model on both signals and events to predict alarms across the entire plant in conditions which appear normal to the operator, that is, in the absence of other alarms. Further, to overcome the aforementioned complexity of choosing the most appropriate multimodal fusion technique, we propose a model trained end-to-end, which implicitly learns the optimal representation of the data from multiple fusion strategies.

To this end, we introduce **MUltifuSion Transformer** (MUST), a multimodal Transformer-based model which learns the best fusion automatically. The model is validated on the task of early alarm prediction based on the combination of recent events and signal data. Given a window of several minutes of event logs and signal data, the model predicts whether an alarm is going to be triggered after the next few minutes. In addition, while analyzing data coming in from an entire plant, the model learns to identify the problematic area within the plant where the alarm is predicted to happen, and it also predicts the alarm location. In MUST, we combine three fusion techniques: the most common state-of-the-art 'early' and 'late' fusion, as well as a novel 'deep late' fusion. Whereas depending on the dataset and the task, different fusion techniques can prove more accurate, MUST can automatically learn the optimal representation of the data from multiple fusion strategies during an end-to-end training and potentially outperform individual fusion techniques. This eliminates the necessity for preliminary fine-tuning, thereby reducing both development costs and GPU runtime. MUST is validated on a real-world customer use case.

Overall, we make the following contributions:

1. *We present the first work on using Transformers in a multimodal setting for real industrial data, thereby paving the way for industrial transformation.*
2. *We propose a MultiFusion method which automatically learns the optimal representation of the data from multiple fusion strategies and leverages the best fusion of modalities, thus eliminating the need for extra fine-tuning.*
3. *We show that MultiFusion Transformer is effective in dealing with heavily unbalanced real-world datasets with over 100 classes.*

2. Related work

Machine learning techniques are widely studied in industrial applications like alarm prediction, however, they mostly rely on unimodal data.

2.1. Alarm prediction techniques

Data-driven approaches to alarm prediction commonly use signal data. For instance, Langone et al. (2014) train a nonlinear autoregressive model for temperature prediction. Based on its forecast, a binary classifier predicts future alarms. Similarly, Villalobos et al. (2021) forecast sensor measurements with an LSTM model and apply a ResNet classifier to predict alarms based on the forecast. Proto et al. (2019) predict alarms using tree-based classifiers using summary statistics over process variables as tabular inputs. In a similar vein, Li et al. (2013) train a customized SVM to predict bearing related alarms from statistics over sensor measurements. Koltsidopoulos Papatzimos et al. (2019) predict wind turbine alarms based on wind speed distribution analysis. Chatterjee and Dethlefs (2020) use a Transformer model to predict an alarm class from sensor measurements of wind turbines. In other studies, next alarm is predicted based on previous alarm records, for example, Zhu et al. (2016) convert alarm sequences into n-grams and apply maximum likelihood estimation to predict the next alarm. Cai et al. (2019) convert alarm sequences into word embeddings and predict the next alarm using an LSTM network. Wang and Liang (2020) train binary LSTM-based classifiers per each alarm type and predict the next alarm using alarm clustering and model voting. In relying on a sequence of already triggered alarms however, such studies lie out of the scope of early alarm prediction.

To the best of our knowledge, multimodal learning has not yet been used to solve the task of early alarm prediction.

2.2. Fault detection techniques

In a related task of fault detection, several studies can be found where modalities are combined. Inceoglu et al. (2021) use a multimodal classifier for failure detection. The model uses early fusion to combine RGB and depth frames in convolutional and convLSTM layers, then late-fuses the output via concatenation with that of convolutional layers for audio data, then applies fully connected layers and produces a binary output. In a similar fashion, Li et al. (2021) use a stacked denoising autoencoder to extract audio features and a CNN to process images. The concatenated outputs are passed through linear and softmax layers to predict a fault. Yang et al. (2021) use separate CNNs for images and text records and a fully connected network for structured maintenance data. The concatenated outputs are passed through a regression layer to output degradation level. Limoyo et al. (2023) late-fuse outputs from image and signal CNNs in a GRU network to predict real-valued 2D position commands. The majority of approaches however also use unimodal data, namely, process variables: (Datong et al., 2009; Di Lello et al., 2013; Li et al., 2017; Helbing and Ritter, 2018; Zhao et al., 2018; Giurgiu and Schumann, 2019; Langone et al., 2020; Lucke et al., 2020; Lomov et al., 2021; Reinartz et al., 2021; Fadzail et al., 2022; Wang et al., 2023; Song et al., 2023).

2.3. Multimodal learning outside of industrial context

Most studies in multimodal learning rely on the state-of-the-art Transformer models (Vaswani et al., 2017). In particular, multimodal Transformers pretrained on large datasets (with millions of samples) have become popular and rely on different types of fusion and pretraining strategies. Lu et al. (2019) build a joint model using cross-attention on intermediate representations of visual and linguistic modalities extensively pretrained on multimodal tasks. Similarly, Tan and Bansal (2019) incorporate a cross-modal attention layer on top of two unimodal encoders, pretrained on multimodal tasks: masked language modeling, masked object prediction, cross-modality matching, and image question answering. Sun et al. (2019) late-fuse a language and a video encoder and train the model on the speech recognition task using a cross-modal noise contrastive estimation loss. Tsai et al. (2019) combine bidirectional cross-modal attention blocks over pre-extracted language, visual and audio features. They use six pairs of crossmodal Transformers, plus one Transformer per modality to merge two corresponding crossmodal blocks. The model has no decoder and is trained on a classification task. Pramanik et al. (2020) pass image, text and video to separate 'spatial' and 'temporal' encoders (with the 'spatial' dimension averaged) and combines the two with a decoder. The model is pretrained simultaneously on several unimodal and multimodal tasks. Radford et al. (2021) jointly pretrain image and text encoders oh a huge dataset to align samples using a contrastive loss and enable zero-shot image classification. Wang et al. (2021) use a single Transformer network as an image encoder, a text encoder, and a fusion network in different pretraining tasks with an image-text contrastive loss, an image-text matching loss, and a masked language modeling loss. Hu and Singh (2021) pass concatenated outputs of modality-specific encoders to a shared decoder with task-specific output heads. It is jointly trained on all tasks. Cho et al. (2021) use an early fusion encoder taking concatenated text and visual embeddings, and pretrain on multimodal tasks framed as generative predictions. In a similar vein, Wang et al. (2022) present an encoder–decoder framework pretrained on unimodal and multimodal tasks. Ma et al. (2022) compare early and late fusion with respect to robustness to missing modalities. Ijaz et al. (2022) and Feng et al. (2023) use cross-attention fusion on top of modality-specific encoders. Zhang et al. (2022) use dedicated modality-specific encoder models and apply cross-attention fusion. Similarly, Roy et al. (2023) proposes an early fusion based Transformer adapted for hyperspectral data.

State-of-the-art multimodal transformers typically rely on either early, late or cross-fusion or their slight modification and leverage pre-training on huge datasets to obtain strong multimodal representations. In industrial settings, where datasets are small and noisy and represent highly complex processes, the choice of a fusion method is a more important factor of success. However, since the search for the optimal fusion for each dataset is costly, *we propose a model that learns the best fusion implicitly*.

3. MultiFusion transformer

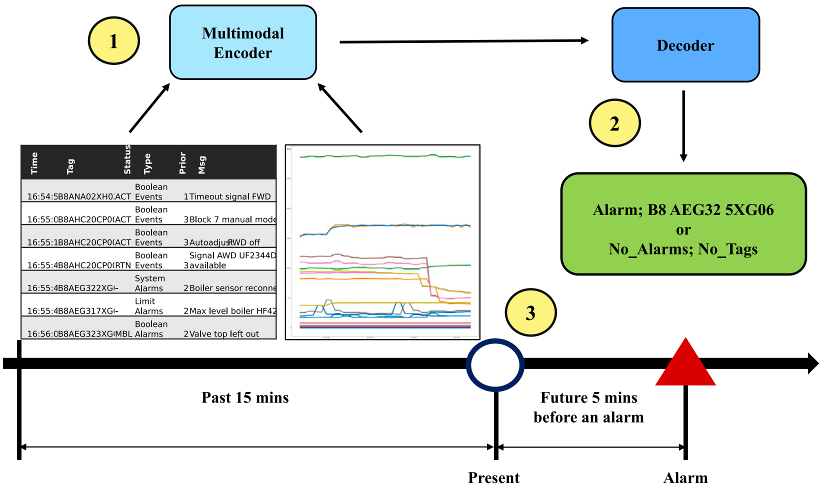
Although numerous fusion methods have been proposed within multimodal machine learning, no single methods outperforms others in all settings. However, finding a fusion that best fits the data and task at hand can be costly. We introduce **MultiFusion Transformer** (MUST), a multimodal attention-based model that automatically learns the optimal representation of the data from multiple fusion strategies to leverage the information from complementary modalities and ensure accurate predictions.

3.1. Early alarm prediction use case

In industrial processes, events and sensor measurements are continuously logged by the control system. Anomalous conditions are indicated to the operator by triggering alarms when predefined thresholds are crossed. Assuming that deviations in the process can start manifesting themselves in events and in the interaction of various signals before a threshold of a specific signal is reached, the goal is to detect such deviations in a data-driven fashion combining different data sources and give the operator an early warning. To achieve this, at every point in time t:

1. A window of N minutes of most recent events and signal data from $t - N$ to t is fed into the model's encoder — Fig. 2(1).
2. The model, trained on historic data, makes a prediction M minutes into the future including a binary output (whether and alarm is going to fire at time $t + M$) and a multiclass output (at which tag an alarm is going to be triggered) — Fig. 2(2).
3. If an alarm is predicted, an early warning is shown to the operator so that he has more time to take corrective actions — Fig. 2(3).

We treat the alarm forecasting task as a sequence-to-sequence learning problem and tackle it using an encoder–decoder Transformer model. The encoder projects the multimodal input sequence of events and continuous multivariate signal data from the past several minutes to a joint hidden representation, which is then fed into the decoder, which in its turn generates a prediction based on this representation.



Time	Tag	Status	Type	Prior	Msg
16:54:5	B8ANA02XH0	ACT	Boolean Events	1	Timeout signal FWD
16:55:0	B8AHC20CP0	ACT	Boolean Events	3	Block 7 manual mode
16:55:1	B8AHC20CP0	ACT	Boolean Events	3	Autoadjust FWD off
16:55:4	B8AHC20CP0	RTN	Boolean Events	3	Signal AWD UF2344D available
16:55:4	B8AEG322XG	-	System Alarms	2	Boiler sensor reconne
16:55:4	B8AEG317XG	-	Limit Alarms	2	Max level boiler HF42
16:56:0	B8AEG323XG	MBL	Boolean Alarms	2	Valve top left out

Fig. 2. Input and output modalities: The encoder creates a representation of the past 15 min of events and signal data, which is used by the decoder to forecast an alarm 5 min into the future.

3.2. Multimodal encoder

Existing approaches to alarm prediction rely either on signals or events separately and usually focus on specific alarm types. Yet to tackle the complex task of predicting alarms across the entire plant and also identifying the specific tag where an alarm would fire, a model must learn a rich representation of the underlying industrial process. In particular, it should capture not only patterns in the individual process variables, but also their interactions with other signals and events. Being contextually related, different modalities provide complementary information, reinforcing salient features and contributing additional ones. With MUST, we propose to combine two complementary modalities to leverage the model's representation learning capacity and thereby maximize the predictive performance of the model.

The two modalities include signal data and events. Signal data is represented as continuous multivariate time series, whereas events are logged as structured entries indicating state changes (each containing a tag, status, event type, priority, operator message, etc.), which are concatenated and treated as text.

Both signal and event data have timestamps, however, the two modalities are not aligned: unlike process variables, which can be resampled at a desired rate, events are unevenly distributed over time. For this reason, samples are created using sliding windows. For each alarm, time windows are selected starting 20 min before the alarm and ending 5 min before it, and balanced with randomly sampled time windows of the same duration not followed by alarms.

To enable multimodal processing, time windows of recent events and sensor readings first undergo a series of transformations in the input layers of MultiFusion Transformer. In the first layer, a window of scaled process variables undergoes a linear transformation, while a sequence of event tokens is passed through an embedding layer such that both modalities are projected into the same hidden dimension D. Input features of each sample can be denoted by $X_{\{S,E\}} \in \mathbb{R}^{T_{\{S,E\}} \times D_{\{S,E\}}}$, where T denotes the number of timestamps within a window for signal data S, and the number of tokens within a concatenated sequence of events E (sequences of event tokens are trimmed to the length of 300). Both the embedding layer for events and the linear layer for signal data have a hidden dimension of 512 to allow the model to learn a rich representation of the input. A learnable positional encoding is then added to the transformed inputs of both modalities separately. Finally, a linear transformation is applied to both modalities to reduce dimensionality to 32 and force the model to narrow the representation down to the most salient patterns.

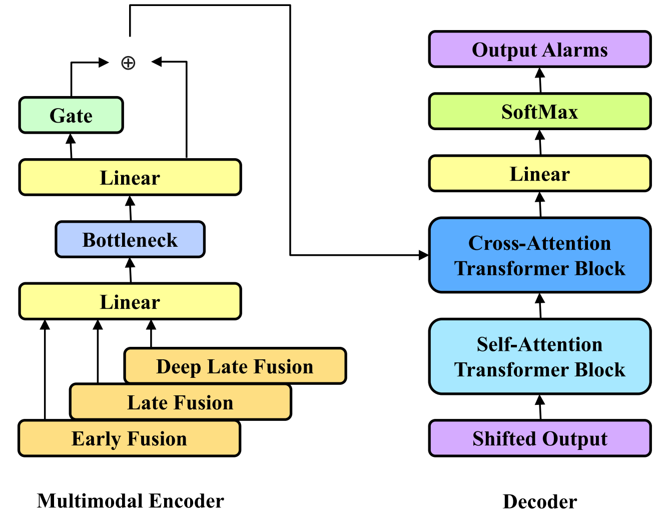
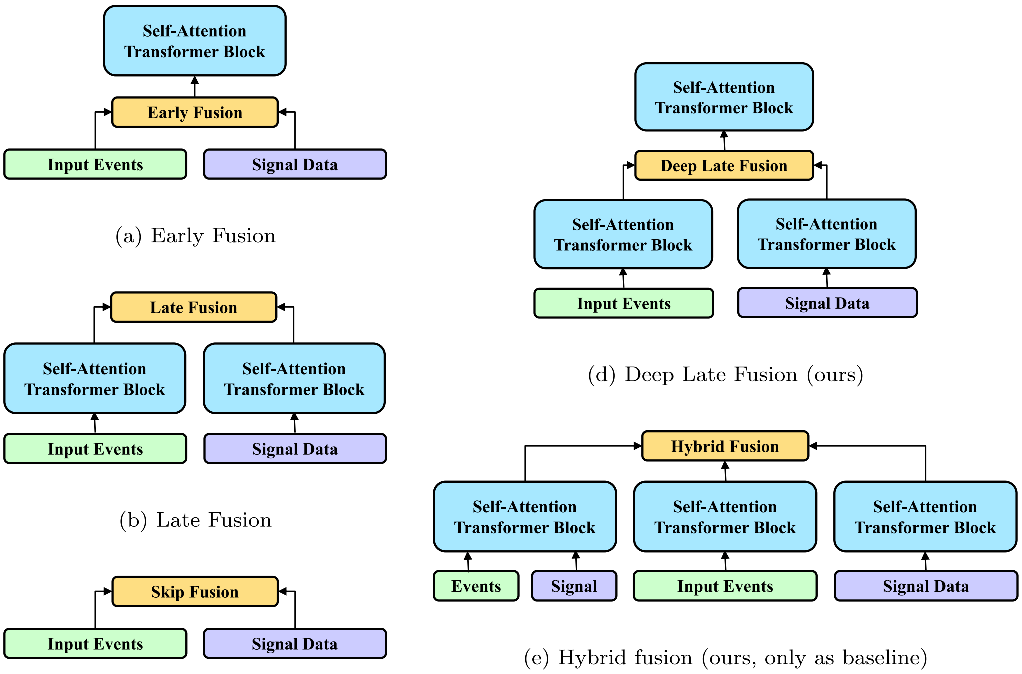


Fig. 3. The MUST Architecture. The encoder incorporates multiple fusion layers, concatenated and passed through a bottleneck layer guided by a sigmoid gate. The resulting representation is passed to the cross-attention block of the decoder.

The basic building blocks of the model are standard Transformer layers, each including a four-headed attention sublayer and a fully connected feed-forward sublayer, both followed by a residual connection. In line with previous works (Klein et al., 2017; Vaswani et al., 2018), MUST uses pre-normalization, whereby a normalization layer precedes a Transformer layer. The arrangement of Transformer blocks is described in detail below.

3.3. MultiFusion

Traditionally, multimodal transformers rely on one of the state-of-the-art fusion techniques, such as early or late fusion (see Section 2.3). However, in practice, no single fusion can surpass others in all situations: depending on the data and the problem to be solved, different techniques may score better (Snoek et al., 2005; Perez-Rua et al., 2019; Boulahia et al., 2021; Ma et al., 2022). Unfortunately, implementing various fusion methods and grid-searching through them is costly and impractical. Therefore, to eliminate the time-consuming extra step of manual model architecture search, we propose to let the model itself learn the optimal representation of the data from multiple fusion techniques.


Self-Attention Transformer Block
Early Fusion
Input Events
Signal Data
Late Fusion
Self-Attention Transformer Block
Self-Attention Transformer Block
Input Events
Signal Data
Skip Fusion
Input Events
Signal Data
Self-Attention Transformer Block
Deep Late Fusion
Self-Attention Transformer Block
Self-Attention Transformer Block
Input Events
Signal Data
Hybrid Fusion
Self-Attention Transformer Block
Self-Attention Transformer Block
Self-Attention Transformer Block
Events
Signal
Input Events
Signal Data

(a) Early Fusion

(b) Late Fusion

(c) Skip Fusion (ours, only as baseline)

(d) Deep Late Fusion (ours)

(e) Hybrid fusion (ours, only as baseline)

Fig. 4. Fusion types: the state-of-the-art early and late fusion and the novel skip, deep late and hybrid fusion. Only early, late and deep late fusion are used as building blocks in MultiFusion Transformer. All the fusion types are used as baselines.

The overall architecture of MUST is outlined in Fig. 3. As the building blocks, the multimodal encoder incorporates three fusion types, namely, the state-of-the-art early and late fusion, as well as a novel deep late fusion:

1. In ***early fusion***, the transformed inputs from both modalities are concatenated and together passed through a single Transformer block (Fig. 4(a)).
2. In ***late fusion***, signal data and events are passed through separate Transformer blocks and concatenated after (Fig. 4(b)).
3. ***Deep late fusion*** combines the early and late fusion: after being passed through separate Transformer blocks, signal data and events are concatenated and passed together through a third Transformer block (Fig. 4(d)).

The outputs of these blocks are concatenated along the sequence dimension, then normalized and transposed. Further, they are passed through a bottleneck consisting of a linear layer reducing the dimensionality by 0.5 to enable a more compact representation, followed by a GELU, and a second linear layer restoring the dimensionality. The output is split in two parts. A sigmoid gate is applied to the first part and the resulting representation is then added to the other part. This results in guiding the model to have a stronger focus on more salient parts of the input representation. Finally, the dimensions are transposed to the original shape, and the encoder output is fed into the cross-attention block of the decoder.

3.4. Alarm decoder

In addition to the binary prediction of whether an alarm would occur within the given horizon, the model also predicts its tag, i. e. the specific component where the alarm would fire. Tags are encoded according to the KKS standard (V.G.B. Kraftwerkstechnik GmbH Essen, 2021) as an alphanumeric string designating the hierarchy of a component's location in the plant: the plant, function, equipment and component (e. g., X0 HNA70 FQ013 XH52).

The prediction is made sequentially, beginning with a binary identifier Alarm/No_alarms, which indicated whether any alarm would fire after the forecasting horizon, followed by the token No_tag in case of No_alarms, else by the tag representation split into tokens corresponding to the plant, function, equipment and component (e. g., Alarm X0 HNA70 FQ013 XH52).

The decoder generates its prediction for each subsequent token based on the hidden representation produced by the multimodal encoder (via the cross-attention Transformer block) and the previously generated tokens (via the self-attention Transformer block — see Fig. 3). In the final decoder layer, the outputs of the Transformer block are passed through a linear layer to bring the output dimension in line with the token vocabulary size, followed by a softmax, approximating the probability distribution over tokens at each step. Finally, they are mapped to the output tokens by picking the most probable one.

4. Experimental evaluation

The performance of MUST is compared on the alarm forecasting task against Transformers with traditional fusion techniques to verify that it can match the best fusion method. Experiments on alarm forecasting were performed using two real datasets and extended through an experiment on the operation forecasting task on a simulated dataset.

4.1. Datasets

We use two proprietary real-world industrial datasets from two waste incineration plants and one simulated dataset which is used to predict upcoming operations in a chemical batch process to validate MUST.

4.1.1. Waste incineration plants datasets

Municipal solid waste incineration is a traditional method of waste disposal and power generation: the heat released from the incineration of the municipal solid waste can be used as the input energy to thermal power plants (Yazdani et al., 2020). Fig. 5 shows an example of such

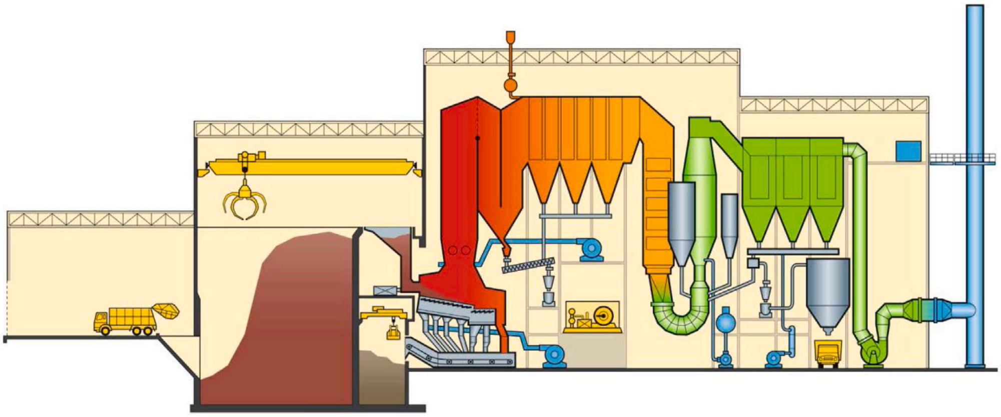


Fig. 5. An example of a waste incineration plant scheme (courtesy of Doosan Lentjes GmbH). Waste is stored and mixed in the reception area. From there, it is moved on the grate to the combustion chamber where it is burnt using flue gas and pre-heated air, and the resulting steam is used to generate electricity, yet it can also be used directly, e. g., for district heating. Finally, gases are cleaned before being emitted from the plant.

a plant. Waste-to-energy technologies reduce the environmental impact of waste management and at the same time decrease the dependency on fossil fuels (Psomopoulos et al., 2009). Due to the heterogeneous nature of waste and, hence, different combustion and transport behaviors, the incineration process is unstable and generates a fair number of alarms (Wissing et al., 2017; Ye et al., 2021). Early alarm prediction can give the operator more time to introduce corrective actions to eliminate possible failures, downtime and negative impact on human safety and the environment.

For the early alarm prediction, real data from two incineration plants (plant A and plant B) is used, recorded over 6 months. In both datasets, the signal data includes 31 continuous process variables selected by an expert and resampled at a frequency of 30 seconds (such as flow rates and temperatures of primary air, natural gas and flue gas, amount of incinerated waste, or feed water temperatures and pressures). The event log consists of automatically generated entries indicating state changes throughout the plant, covering different functional areas. These entries are composed of attributes such as timestamp, tag (a code specifying the component), status, event type, priority, message, etc. Overall, there are over 8M events logged over the given period in one dataset and 9M in the other.

The model is trained on time windows preceding alarms such that there are no other alarms in the same functional area 20 min before, balanced with an equal number of time windows not followed by alarms. Specifically, for every sample, a window is taken starting 20 min before an alarm and ending within a forecasting horizon of 5 min before the alarm.

There are 10,600 samples (half positive, half alarm-free) in the dataset from plant A, which is randomly split into train, dev and test sets of 8608, 996 and 996 samples, respectively (0.80/0.10/0.10). In the plant B dataset, there are 33,496 samples, split into train, dev and test sets of 27,084, 3206, 3206 samples, respectively. Thus, the datasets are balanced w. r. t. binary prediction and highly unbalanced w. r. t. multiclass prediction, with half of the samples labeled as having No_Tag and the rest being distributed as in Figs. 6(a) and 6(b).

To sum up, both datasets pose several challenges:

1. The datasets are small, considering the number of classes and the number of samples per class, as well as task complexity;
2. The distribution of classes is highly unbalanced, with the vast majority of classes represented by very few samples;
3. The input events are very sparse and spread across numerous plant functions, equipment and components;
4. The data comes from real plants, characterized by a high degree of randomness in the process, and contains a lot of noise;
5. Early alarm prediction means that the underlying changes causing the alarm might not have manifested themselves in the data yet;
6. The available process variables are not linked to alarm tags. The task could be easier if all signals corresponding to different alarm tags could be used: in that case, given enough data, the model could learn the thresholds. However, a mapping from alarm tags to appropriate process variables is unavailable, and only 31 signals are used to predict up to 163 alarm types.

4.1.2. Simulated dataset

In addition to real datasets, our approach is validated on a more tractable task of forecasting operations in a chemical batch process using a simulated dataset [dataset] (Tan, 2022). Batch production processes consist of cyclic sequences of operations like heating, cooling, chemical reactions, or stirring. The specific instances of such operations can vary considerably even for the same product (Just et al., 2022), e. g. due to variation in the raw material. A reliable prediction when such operation will end are valuable inputs for production, logistics and personnel planning.

The dataset covers an equivalent of over 2 months of a batch production process and contains 4268 events of 18 classes designating the start of a new operation (such as filling, processing, draining and cleaning) and signal data with 5 continuous process variables sampled at 15 seconds (vessel filling levels, motor rotation speed, cooling water flow rate and steam flow).

The model is trained on time windows preceding each new operation starting 36 min before it begins and ending within a forecasting horizon of 6 min. These windows are balanced with an equal number of time windows not followed by new operations within the forecasting horizon. The task here is to first make a binary prediction of whether a new operation is going to begin after the forecasting horizon and, if yes, which operation it would be. The simulated dataset contains 7440 samples, randomly split into train, dev and test sets of 5616, 912 and 912 samples, respectively.

4.2. Baseline models

We implement separate Transformer models with various fusion types, namely, the state-of-the-art early and late fusion, as well as the competitive novel ‘skip’, ‘deep late’ and hybrid fusion (Strem et al., 2025). In addition to the individual fusion techniques incorporated in the MultiFusion (see Section 3.3), we compare the model against:

1. ***Deep late fusion***, adds a self-attention Transformer block on top of the late fusion (Fig. 4(d)).

2. ***Skip fusion***, implies a simple concatenation of the transformed inputs from both modalities, the result of which is passed to the decoder directly, without self-attention, like a residual connection (Fig. 4(c)).
3. ***Hybrid fusion***, combines the early and late fusion: on the one hand, signal data and events are concatenated and passed through a single Transformer block, on the other, the same inputs are passed through separate Transformer blocks and concatenated after. Then, the outputs of both the early-fused and the late-fused blocks are concatenated and passed to the decoder (Fig. 4(e)).

4.3. Training parameters and scoring

The models are trained using Adam optimizer (Kingma and Ba, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and a learning rate of 1.0^{-3}. The dropout rate is 0.3. The batch size is 128. The model is trained using the cross-entropy loss averaged over all output tokens. All experiments have been run with 5 random seeds on NVIDIA GeForce RTX 2080 Ti graphics card with 12 GB memory with Ubuntu 22.04.2 LTS.

Due to the highly unbalanced distribution of alarm classes, the main evaluation metric used is F1 score, computed separately for binary and multiclass predictions. To calculate F1 score for binary predictions, only the first token of each sample is considered (Alarm/No_alarms). For multiclass predictions, the subsequent tokens are concatenated, and F1 is calculated on these concatenated tokens (e. g., X0HNA70FQ013-XH52). This way, even if a tag prediction is partially correct (e. g., X0HNA 70 CP001 XH52 instead of X0HNA 70 FQ013 XH52), the whole tag is considered incorrect.

Since the multiclass F1 is strict and ignores partial matches, the evaluation also includes the BLEU score (Papineni et al., 2002) (used for Transformer evaluation in traditional NLP tasks) to compare representation learning capacity of MultiFusion Transformer against the baseline models.

4.4. Results and discussion

The proposed MUST model incorporating the novel MultiFusion method is compared against traditional fusion types such as early and late fusion, as well as the novel skip, deep late and hybrid fusion. To keep the models compact and efficient, an important requirement in online industrial settings, and prevent overfitting given the limited size of the datasets for all the models, we use a hidden dimensionality of 32.

The results for real datasets are summarized in Table 1 (plant A) and Table 3 (plant B) where average F1 scores for binary (Alarm/No_alarms) and multiclass predictions (specific tags) for the two plants across different models are presented. For a more fine-grained analysis, we also calculate BLEU score. The results for the simulated dataset are summarized in Table 5. In addition, Fig. 7 shows the confusion matrices for the simulated dataset.

Table 1
Performance of MultiFusion compared to SotA fusion techniques based on experiments on the real dataset, plant A (best in **bold**).

	F1 Bin ↑	F1 MC ↑	BLEU ↑	Params (M)	Time
Skip Fusion	69.2 ± 1.3	37.0 ± 0.5	35.5 ± 2.0	1.2 ± 0.0	0.7 ± 0.2
Early Fusion	66.7 ± 0.9	37.7 ± 0.3	35.9 ± 0.6	1.4 ± 0.0	2.0 ± 0.5
Late Fusion	67.5 ± 1.1	37.6 ± 1.0	35.4 ± 1.3	1.7 ± 0.0	2.6 ± 0.9
Deep Late	66.5 ± 1.3	38.3 ± 0.8	36.6 ± 1.9	2.0 ± 0.0	4.2 ± 0.7
Hybrid Fusion	67.8 ± 1.3	**38.7 ± 1.1**	36.7 ± 2.9	2.0 ± 0.0	5.3 ± 2.2
MultiFusion	**70.3 ± 0.9**	**38.7 ± 0.3**	**39.0 ± 1.2**	2.3 ± 0.0	2.8 ± 1.0

Table 2
Effects of dimensionality on predictive performance averaged across baseline fusion types based on experiments on the real dataset (plant A).

	F1 Bin ↑	F1 MC ↑	BLEU ↑	Params (M)	Time
32	67.5 ± 1.5	37.9 ± 1.0	36.0 ± 1.8	1.7 ± 0.3	3.0 ± 1.9
64	68.8 ± 1.0	36.3 ± 1.1	31.1 ± 4.0	2.5 ± 0.7	1.0 ± 0.7
128	68.9 ± 0.9	35.3 ± 0.9	29.3 ± 3.4	4.5 ± 1.4	0.5 ± 0.4
256	69.2 ± 1.1	35.2 ± 0.5	27.2 ± 3.3	9.1 ± 3.1	0.5 ± 0.3

Table 3
Performance of MultiFusion compared to SotA fusion techniques based on experiments on the real dataset, plant B (best in **bold**).

	F1 Bin ↑	F1 MC ↑	BLEU ↑	Params (M)	Time
Skip Fusion	84.1 ± 0.2	41.2 ± 0.2	37.8 ± 1.0	1.6 ± 0.0	1.8 ± 0.4
Early Fusion	84.1 ± 0.3	41.9 ± 0.5	39.2 ± 1.0	1.9 ± 0.0	4.2 ± 0.9
Late Fusion	84.0 ± 0.3	42.5 ± 0.2	39.4 ± 1.2	2.2 ± 0.0	6.2 ± 1.3
Deep Late	84.0 ± 0.4	42.5 ± 0.5	39.6 ± 1.4	2.5 ± 0.0	7.9 ± 2.1
Hybrid Fusion	84.2 ± 0.3	42.2 ± 0.3	39.2 ± 0.5	2.5 ± 0.0	7.9 ± 1.4
MultiFusion	**84.3 ± 0.3**	**42.7 ± 0.8**	**40.4 ± 1.5**	2.9 ± 0.0	4.1 ± 1.0

Table 4
Effects of dimensionality on predictive performance averaged across baseline fusion types based on experiments on the real dataset (plant B).

	F1 Bin ↑	F1 MC ↑	BLEU ↑	Params (M)	Time
32	84.1 ± 0.3	42.1 ± 0.6	39.0 ± 1.2	2.1 ± 0.3	5.6 ± 2.7
64	84.3 ± 0.4	41.1 ± 0.6	38.0 ± 1.5	3.0 ± 0.7	2.4 ± 1.1
128	84.3 ± 0.3	40.6 ± 0.6	36.9 ± 1.6	5.0 ± 1.4	1.7 ± 0.9
256	84.2 ± 0.4	40.9 ± 0.6	37.1 ± 2.3	9.6 ± 3.1	2.0 ± 1.3

Table 5
Performance of MultiFusion compared to SotA fusion techniques based on experiments on the simulated dataset (best in **bold**).

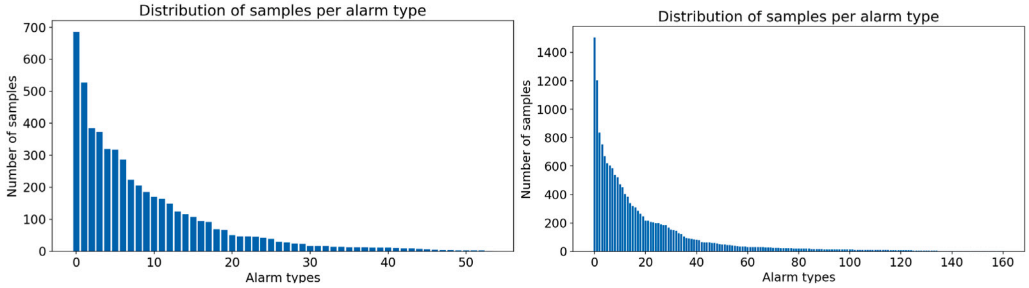
	F1 Bin ↑	F1 MC ↑	BLEU ↑	Params (M)	Time
Skip Fusion	90.7 ± 0.2	74.3 ± 0.6	75.5 ± 1.7	0.4 ± 0.0	0.8 ± 0.3
Early Fusion	90.7 ± 0.2	74.1 ± 0.3	79.3 ± 0.4	0.7 ± 0.0	2.2 ± 0.3
Late Fusion	90.5 ± 0.2	74.3 ± 0.6	79.4 ± 0.4	1.0 ± 0.0	4.2 ± 0.8
Deep Late	90.7 ± 0.3	74.4 ± 0.5	79.5 ± 1.1	1.3 ± 0.0	4.9 ± 1.0
Hybrid Fusion	90.7 ± 0.2	74.5 ± 0.4	79.3 ± 0.8	1.3 ± 0.0	4.0 ± 0.8
MultiFusion	**90.8 ± 0.3**	**74.7 ± 0.7**	**79.6 ± 0.5**	0.9 ± 0.0	2.9 ± 0.7

Table 6
Effects of dimensionality on predictive performance averaged across baseline fusion types based on experiments on the simulated dataset.

	F1 Bin ↑	F1 MC ↑	BLEU ↑	Params (M)	Time
32	90.7 ± 0.2	74.3 ± 0.5	78.6 ± 1.8	0.9 ± 0.3	4.1 ± 1.5
64	90.7 ± 0.3	74.3 ± 0.5	77.6 ± 2.8	1.8 ± 0.7	3.2 ± 1.7
128	90.7 ± 0.2	74.5 ± 0.5	77.4 ± 3.0	3.8 ± 1.4	2.5 ± 1.5
256	90.7 ± 0.3	74.6 ± 0.7	77.9 ± 2.6	8.4 ± 3.1	3.6 ± 2.5

Role of the dataset. Overall, the scores highly depend on the dataset, with the most accurate predictions on the simulated dataset, which contains relatively few samples (7440) and has much less noise and fewer classes (18), and lowest for the real dataset from plant A, which is both highly noisy, unbalanced and relatively small (with 10,600 samples vs. 33,496 samples for plant B). The same pattern can be observed not only in the absolute scores but also their variation across different models, which is fairly low for the more tractable simulated dataset, but higher on the real dataset from plant B. This is even more pronounced on the dataset from plant A, highlighting the importance of the choice of the fusion technique, especially when dealing with small and unbalanced datasets.

Role of the fusion method. It is important to observe that, depending on the dataset and the number of classes, different fusion techniques may perform better on different metrics. For instance, the least complex skip fusion models, although inferior in terms of BLEU score on all datasets, outperform all other baselines w. r. t. binary F1 score on the



(a) Real data (plant A)

(b) Real data (plant B)

Fig. 6. Alarm type distribution: alarm types are highly unbalanced with ca. 20% of the alarm types accounting for 90% of the samples.

data from plant A. On plant A data, late fusion is better than early fusion in binary F1 but slightly worse in multiclass F1 and BLEU scores. Deep late fusion, although scoring higher than the two in most cases, is outperformed by the late fusion on the plant A data. Likewise, the hybrid fusion reaches higher scores then the alternatives in most cases, yet sometimes is slightly inferior despite the higher model complexity, such as in multiclass F1 and BLEU scores on the plant B data. This finding supports previous work (Snoek et al., 2005; Perez-Rua et al., 2019; Boulahia et al., 2021; Ma et al., 2022) highlighting the importance of choosing the right fusion method in each individual case and the value of learning the best fusion automatically.

In contrast to baselines, MultiFusion Transformer, which intrinsically learns the optimal representation of the data from multiple fusion techniques, consistently matches the best fusion in each setting and even outperforms it. This applies to all datasets across all baselines both in terms of binary and multiclass predictions, and even more so on the BLEU score, which is the prevalent metric used for the evaluation of language models on multiple tasks such as machine translation or text generation. In the given setup, it provides a more fine-grained measurement of model accuracy by accounting for partially correct tag predictions (such as X0HNA 70 CP001 XH52 instead of X0HNA 70 FQ013 XH52). The advantage of MUST is more pronounced when the task is more complex, e. g. on the dataset from plant A, the binary F1 score and BLEU are, respectively, 4% and 8% higher for the MultiFusion Transformer than the average across the baselines, which is a non-trivial improvement given the number of classes, the task complexity and the data limitations.

A more detailed insight into the performance of MUST across classes can be gained from the confusion matrices for the simulated dataset in Fig. 7 (chosen due to the more manageable number of classes compared to the real datasets). The proposed model makes reliable binary predictions of an upcoming phase change in the process (Fig. 7(a)). MUST is also remarkably robust in the multiclass scenario: 11 out of 18 phases are predicted correctly more than 70% of the time, out of which 8 are successfully predicted in 80% of cases and 4 classes even surpassing the 90% performance. This demonstrates the high accuracy of the proposed model while indicating a promising potential in other applications.

Overall, it is important to stress that MUST not only performs on par with the best of the baseline fusion methods, thus eliminating the need for preliminary architecture search, but also surpasses it in almost all scenarios across all evaluation metrics. The consistently higher scores attained by MUST on simulated and real datasets demonstrate that the proposed MultiFusion method is capable of learning patterns in multimodal industrial data and generating more robust early alarm predictions regardless of the data size and the distribution of classes and task complexity.

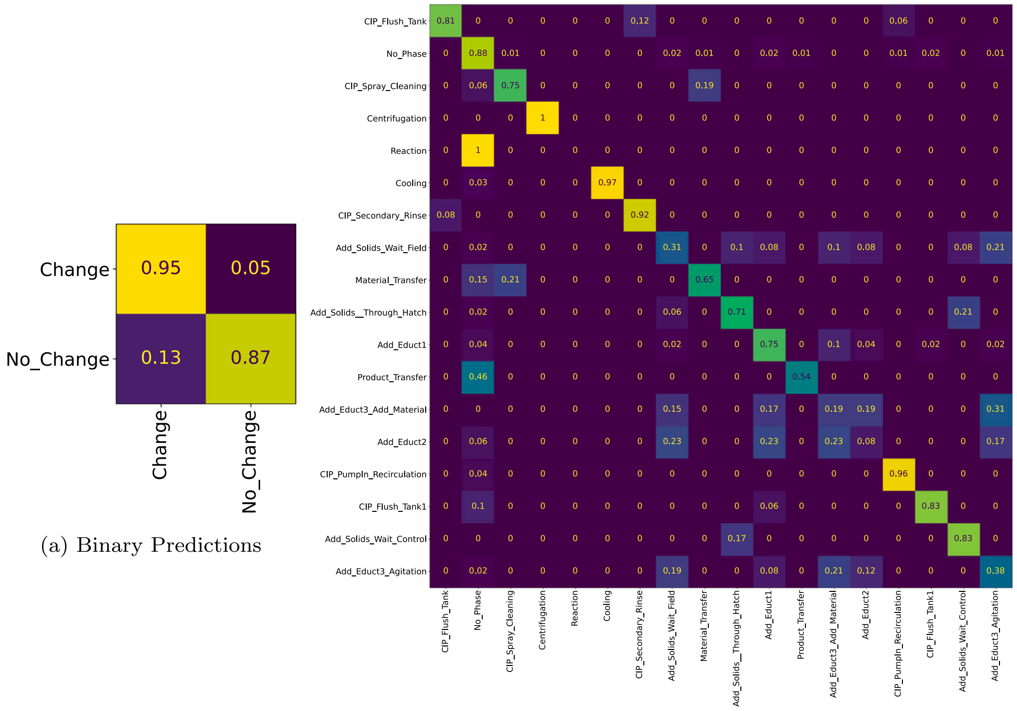
Role of the model size. One important point to address is model size: since MultiFusion incorporates other fusion types, the number of parameters is higher (except for the model trained on the simulated dataset, since the predicted sequence length is lower in this case). Therefore, one could raise a question that the model's success is due to its larger size. To eliminate this concern, we ran experiments for all the baseline fusion types with different hidden dimensionality values: 32, 64, 128 and 256, resulting in model sizes ranging from ∼1 to 2 million parameters to ∼8.5 to 9.5 million (depending on the number of classes and output sequence length per dataset). The averaged results are shown in Tables 2 (plant A), 4 (plant B) and 6 (the simulated dataset). As can be seen, the increase of dimensionality either brings negligible improvement (e. g., in binary F1 on the plant A dataset) or even impairs model performance (especially the multiclass F1 score and BLEU), with the multiclass F1 and BLEU scores dropping for all models with dimensionality higher than 32 on real data and increasing very slightly on the simulated data. At the same time, the overall model size increases significantly, while all the average metrics across all datasets remain below the performance of MUST, which is only slightly bigger than the baseline models. Thus, the advantage of MultiFusion Transformer cannot be attributed to the model size.

It can also be observed that MultiFusion Transformer converges faster than the average of the baselines, which results in a faster training time and saves a lot of time overall considering that tuning, training and comparing other models can be eliminated.

As described above, the task at hand is complex due to multiple factors such as sparse and noisy input data, a high number of classes and the highly unbalanced distribution of labels (as illustrated in Figs. 6(a) and 6(b)). Nonetheless, the experiments demonstrate competitive performance of our model and the advantage of the MultiFusion method, which is especially pronounced on the smaller and noisier dataset. MUST surpasses baseline models on all datasets and across all metrics, while remaining efficient w. r. t. model size and training time.

5. Conclusion

As repeatedly demonstrated in existing research, combining multiple modalities enables deep learning models to be more powerful predictors compared to unimodal approaches. At the same time, the choice of a fusion method is not trivial and highly depends on the data and use case. To this end, we introduced MUST, a multimodal Transformer-based model which learns the optimal representation of the data from multiple fusion strategies automatically, thereby eliminating the requirement of extra manual tuning. The model has been applied to solve a highly complex problem in the industrial domain, where, to the best of our knowledge, multimodal Transformers have not yet been used. We have shown that MUST can predict, based on the combination of recent events and signal data as input, whether an alarm is going to be triggered after the given forecasting horizon and, if yes, it also predicts an alarm location. In a series of experiments, our model not only matched, but even outperformed state-of-the-art fusion baselines such as early and late fusion, as well as the competitive novel skip, deep late and hybrid fusion strategies. Our experimental evaluation on two real world industrial datasets and a simulated dataset



(a) Binary Predictions

(b) Multiclass Predictions

Fig. 7. Confusion matrices for operation predictions: MultiFusion Transformer can successfully learn the difference between operations even in the presence of large number of classes on the simulated dataset (with true labels along the y axis and predicted labels along the x axis).

demonstrates that the proposed MultiFusion method yields state-of-the-art predictive performance while eliminating the need to implement and choose among conventional fusion techniques, thus reducing the tuning costs and the GPU runtime.

One potential direction for further exploration would be incorporating the plant topology to take into consideration the causal relationships between the individual units. Further, other applications of MultiFusion in industry can be explored, such as prediction of process KPIs and what-if scenarios. Other interesting applications would be quality assurance and fault detection, which would naturally imply including more modalities such as image data in the MultiFusion architecture.

CRediT authorship contribution statement

Nika Strem: Conceptualization, Data curation, Methodology, Visualization, Writing – original draft. **Devendra Singh Dhami:** Writing – review & editing. **Benedikt Schmidt:** Data curation. **Kristian Kersting:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by ABB AG Forschungszentrum Mannheim, the Collaboration Lab ‘AI in Construction‘ (AICO), the ICT-48 Network of AI Research Excellence Center ‘TAILOR‘ (EU Horizon 2020, GA No 952215) and the HMWK cluster project ‘The Third Wave of AI‘. The Eindhoven University of Technology authors received support from their Department of Mathematics and Computer Science and the Eindhoven Artificial Intelligence Systems Institute.

Data availability

The data that has been used is confidential.

References

Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., Gong, B., 2021. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems, Vol. 34. Curran Associates, Inc., pp. 24206–24221.

Boulahia, S.Y., Amamra, A., Madi, M.R., Daikh, S., 2021. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. Mach. Vis. Appl. 32 (6), 121. http://dx.doi.org/10.1007/s00138-021-01249-8.

Cai, S., Palazoglu, A., Zhang, L., Hu, J., 2019. Process alarm prediction using deep learning and word embedding methods. ISA Trans. 85, 274–283. http://dx.doi.org/10.1016/j.isatra.2018.10.032.

Chatterjee, J., Dethlefs, N., 2020. A dual transformer model for intelligent decision support for maintenance of wind turbines. In: 2020 International Joint Conference on Neural Networks. IJCNN, pp. 1–10. http://dx.doi.org/10.1109/IJCNN48605.2020.9206839.

Chen, S., Guhur, P.-L., Schmid, C., Laptev, I., 2021a. History aware multimodal transformer for vision-and-language navigation. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems, Vol. 34. Curran Associates, Inc., pp. 5834–5847.

Chen, M., Peng, H., Fu, J., Ling, H., 2021b. AutoFormer: Searching transformers for visual recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 12270–12280.

Chitty-Venkata, K.T., Emani, M., Vishwanath, V., Somani, A.K., 2022. Neural architecture search for transformers: A survey. IEEE Access 10, 108374–108412. http://dx.doi.org/10.1109/ACCESS.2022.3212767.

Cho, J., Lei, J., Tan, H., Bansal, M., 2021. Unifying vision-and-language tasks via text generation. In: Meila, M., Zhang, T. (Eds.), Proceedings of the 38th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 139, PMLR, pp. 1931–1942.

Datong, L., Yu, P., Xiyuan, P., 2009. Fault prediction based on time series with online combined kernel svr methods. In: 2009 IEEE Instrumentation and Measurement Technology Conference. pp. 1163–1166. http://dx.doi.org/10.1109/IMTC.2009.5168630.

Di Lello, E., Klotzbücher, M., De Laet, T., Bruyninckx, H., 2013. Bayesian time-series models for continuous fault detection and recognition in industrial robotic tasks. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 5827–5833. http://dx.doi.org/10.1109/IROS.2013.6697200.

Fadzail, N.F., Zali, S.M., Mid, E.C., Jailani, R., 2022. Application of automated machine learning (AutoML) method in wind turbine fault detection. J. Phys. Conf. Ser. 2312 (1), 012074. http://dx.doi.org/10.1088/1742-6596/2312/1/012074.

Feng, C.-M., Yan, Y., Chen, G., Xu, Y., Hu, Y., Shao, L., Fu, H., 2023. Multimodal transformer for accelerated MR imaging. IEEE Trans. Med. Imaging 42 (10), 2804–2816. http://dx.doi.org/10.1109/TMI.2022.3180228.

Giurgiu, I., Schumann, A., 2019. Explainable failure predictions with RNN classifiers based on time series data. arXiv:1901.08554.

Helbing, G., Ritter, M., 2018. Deep learning for fault detection in wind turbines. Renew. Sustain. Energy Rev. 98, 189–198. http://dx.doi.org/10.1016/j.rser.2018.09.012, URL https://www.sciencedirect.com/science/article/pii/S1364032118306610.

Hendricks, L.A., Mellor, J., Schneider, R., Alayrac, J.-B., Nematzadeh, A., 2021. Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers. Trans. Assoc. Comput. Linguist. 9, 570–585.

Hu, R., Singh, A., 2021. UniT: Multimodal multitask learning with a unified transformer. In: Proceedings of the IEEECVF International Conference on Computer Vision. ICCV, pp. 1439–1449.

IEC, 2014. IEC 62682 Management of Alarm Systems for the Process Industries. Standard IEC 62682, International Electrotechnical Commission, IEC, Geneva, Switzerland.

Ijaz, M., Diaz, R., Chen, C., 2022. Multimodal transformer for nursing activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 2065–2074.

Inceoglu, A., Aksoy, E.E., Cihan Ak, A., Sariel, S., 2021. FINO-net: A deep multimodal sensor fusion framework for manipulation failure detection. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, pp. 6841–6847. http://dx.doi.org/10.1109/IROS51168.2021.9636455.

Jabeen, S., Li, X., Amin, M.S., Bourahla, O., Li, S., Jabbar, A., 2023. A review on methods and applications in multimodal deep learning. ACM Trans. Multimed. Comput. Commun. Appl. 19 (2s), http://dx.doi.org/10.1145/3545572.

Just, G., Khaydarov, V., Urba, L., Klöpper, B., Bähner, F.D., 2022. Hidden Markov models und active learning zur automatisierten kennzeichnung von batchphasen in der prozessindustrie. VDI-Ber. 2022 (2399), 615–624.

Kingma, D.P., Ba, J., 2017. Adam: A method for stochastic optimization. arXiv:1412.6980.

Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.M., 2017. OpenNMT: Open-source toolkit for neural machine translation. arXiv:1701.02810.

Koltsidopoulos Papatzimos, A., Thies, P.R., Dawood, T., 2019. Offshore wind turbine fault alarm prediction. Wind Energy 22 (12), 1779–1788. http://dx.doi.org/10.1002/we.2402.

Langone, R., Alzate, C., Bey-Temsamani, A., Suykens, J.A.K., 2014. Alarm prediction in industrial machines using autoregressive LS-SVM models. In: 2014 IEEE Symposium on Computational Intelligence and Data Mining. CIDM, pp. 359–364. http://dx.doi.org/10.1109/CIDM.2014.7008690.

Langone, R., Cuzzocrea, A., Skantzos, N., 2020. Interpretable Anomaly Prediction: Predicting anomalous behavior in industry 4.0 settings via regularized logistic regression tools. Data Knowl. Eng. 130, 101850. http://dx.doi.org/10.1016/j.datak.2020.101850.

Li, H., Huang, J., Huang, J., Chai, S., Zhao, L., Xia, Y., 2021. Deep multimodal learning and fusion based intelligent fault diagnosis approach. J. Beijing Inst. Technol. 30 (2), 172–185. http://dx.doi.org/10.15918/j.jbit1004-0579.2021.017.

Li, H., Qian, B., Parikh, D., Hampapur, A., 2013. Alarm prediction in large-scale sensor networks — A case study in railroad. In: 2013 IEEE International Conference on Big Data. pp. 7–14. http://dx.doi.org/10.1109/BigData.2013.6691771.

Li, Z., Wang, Y., sheng Wang, K., 2017. Intelligent predictive maintenance for fault diagnosis and prognosis in machine centers: Industry 4.0 scenario. Adv. Manufact. 5, 377–387.

Limoyo, O., Ablett, T., Kelly, J., 2023. Learning sequential latent variable models from multimodal time series data. In: Intelligent Autonomous Systems 17. Springer Nature Switzerland, pp. 511–528.

Lomov, I., Lyubimov, M., Makarov, I., Zhukov, L.E., 2021. Fault detection in Tennessee Eastman process with temporal deep learning models. J. Ind. Inf. Integr. 23, 100216.

Lu, J., Batra, D., Parikh, D., Lee, S., 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc.

Lucke, M., Stief, A., Chioua, M., Ottewill, J.R., Thornhill, N.F., 2020. Fault detection and identification combining process measurements and statistical alarms. Control Eng. Pract. 94, 104195. http://dx.doi.org/10.1016/j.conengprac.2019.104195.

Ma, M., Ren, J., Zhao, L., Testuggine, D., Peng, X., 2022. Are multimodal transformers robust to missing modality? In: Proceedings of the IEEECVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 18177–18186.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318. http://dx.doi.org/10.3115/1073083.1073135.

Perez-Rua, J.-M., Vielzeuf, V., Pateux, S., Baccouche, M., Jurie, F., 2019. MFAS: Multimodal fusion architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Pramanik, S., Agrawal, P., Hussain, A., 2020. OmniNet: A unified architecture for multi-modal multi-task learning. arXiv:1907.07804.

Proto, S., Ventura, F., Apiletti, D., Cerquitelli, T., Baralis, E., Macii, E., Macii, A., 2019. PREMISES, a scalable data-driven service to predict alarms in slowly-degrading multi-cycle industrial processes. In: 2019 IEEE International Congress on Big Data. BigDataCongress, pp. 139–143. http://dx.doi.org/10.1109/BigDataCongress.2019.00032.

Psomopoulos, C., Bourka, A., Themelis, N., 2009. Waste-to-energy: A review of the status and benefits in USA. Waste Manage. 29 (5), 1718–1724. http://dx.doi.org/10.1016/j.wasman.2008.11.020, URL https://www.sciencedirect.com/science/article/pii/S0956053X08004066.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (Eds.), Proceedings of the 38th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 139, PMLR, pp. 8748–8763, URL https://proceedings.mlr.press/v139/radford21a.html.

Rahman, W., Hasan, M.K., Lee, S., Zadeh, A., Mao, C., Morency, L.-P., Hoque, E., 2020. Integrating multimodal information in large pretrained transformers. In: Proceedings of the Conference. Association for Computational Linguistics. Meeting, Vol. 2020. NIH Public Access, p. 2359.

Reinartz, C., Kulahci, M., Ravn, O., 2021. An extended Tennessee Eastman simulation dataset for fault-detection and decision support systems. Comput. Chem. Eng. 149, 107281. http://dx.doi.org/10.1016/j.compchemeng.2021.107281.

Roy, S.K., Deria, A., Hong, D., Rasti, B., Plaza, A., Chanussot, J., 2023. Multimodal fusion transformer for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 61, 1–20. http://dx.doi.org/10.1109/TGRS.2023.3286826.

Snoek, C.G.M., Worring, M., Smeulders, A.W.M., 2005. Early versus late fusion in semantic video analysis. In: Proceedings of the 13th Annual ACM International Conference on Multimedia. MULTIMEDIA '05, Association for Computing Machinery, New York, NY, USA, pp. 399–402. http://dx.doi.org/10.1145/1101149.1101236.

Song, X., Sun, P., Song, S., Stojanovic, V., 2023. Finite-time adaptive neural resilient DSC for fractional-order nonlinear large-scale systems against sensor-actuator faults. Nonlinear Dyn. 111 (13), 12181–12196.

Stauffer, T., Clarke, P., 2016. Using alarms as a layer of protection. Process Saf. Prog. 35 (1), 76–83. http://dx.doi.org/10.1002/prs.11739.

Strem, N., Dhami, D.S., Schmidt, B., Klöpper, B., Kersting, K., 2025. APT: Alarm Prediction Transformer. Expert Systems with Applications 261, 125521. http://dx.doi.org/10.1016/j.eswa.2024.125521.

Sun, C., Baradel, F., Murphy, K., Schmid, C., 2019. Learning video representations using contrastive bidirectional transformer. arXiv:1906.05743.

Tan, R., 2022. Datasets from Multiple Cycles. Root, http://dx.doi.org/10.57826/KEEN/ODU6MA.

Tan, H., Bansal, M., 2019. LXMERT: Learning cross-modality encoder representations from transformers. arXiv:1908.07490.

Tsai, Y.-H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.-P., Salakhutdinov, R., 2019. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the Conference. Association for Computational Linguistics. Meeting, Vol. 2019. NIH Public Access, p. 6558.

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A.N., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., Sepassi, R., Shazeer, N., Uszkoreit, J., 2018. Tensor2Tensor for neural machine translation. arXiv:1803.07416.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc.

V.G.B. Kraftwerkstechnik GmbH Essen, 2021. KKS Kraftwerk-Kennzeichensystem. VGB Kraftwerkstechnik GmbH Essen.

Villalobos, K., Suykens, J., Illarramendi, A., 2021. A flexible alarm prediction system for smart manufacturing scenarios following a forecaster–analyzer approach. J. Intell. Manuf. 32 (5), 1323–1344.

Wang, J., Hu, X., Gan, Z., Yang, Z., Dai, X., Liu, Z., Lu, Y., Wang, L., 2021. UFO: A UniFied Transformer for vision-language representation learning. arXiv:2111.10023.

Wang, X., Liang, D., 2020. LSTM-based alarm prediction in the mobile communication network. In: 2020 IEEE 6th International Conference on Computer and Communications. ICCC, pp. 561–567. http://dx.doi.org/10.1109/ICCC51575.2020.9344951.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H., 2022. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (Eds.), Proceedings of the 39th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 162, PMLR, pp. 23318–23340, URL https://proceedings.mlr.press/v162/wang22al.html.

Wang, R., Zhuang, Z., Tao, H., Paszke, W., Stojanovic, V., 2023. Q-learning based fault estimation and fault tolerant iterative learning control for MIMO systems. ISA Trans. 142, 123–135.

Wissing, F., Wirtz, S., Scherer, V., 2017. Simulating municipal solid waste incineration with a DEM/CFD method – influences of waste properties, grate and furnace design. Fuel 206, 638–656. http://dx.doi.org/10.1016/j.fuel.2017.06.037.

Yang, Z., Baraldi, P., Zio, E., 2021. A multi-branch deep neural network model for failure prognostics based on multimodal data. J. Manuf. Syst. 59, 42–50. http://dx.doi.org/10.1016/j.jmsy.2021.01.007.

Yazdani, S., Salimipour, E., Moghaddam, M.S., 2020. A comparison between a natural gas power plant and a municipal solid waste incineration power plant based on an emergy analysis. J. Clean. Prod. 274, 123158. http://dx.doi.org/10.1016/j.jclepro.2020.123158.

Ye, B., Shi, B., Shi, M., Zhang, L., Zhang, R., 2021. Process simulation and comprehensive evaluation of a system of coal power plant coupled with waste incineration. Waste Manag. Res. 39 (6), 828–840.

Zhang, W., Qiu, F., Wang, S., Zeng, H., Zhang, Z., An, R., Ma, B., Ding, Y., 2022. Transformer-based multimodal information fusion for facial expression analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 2428–2437.

Zhao, H., Hu, Y., Ai, X., Hu, Y., Meng, Z., 2018. Fault detection of Tennessee eastman process based on topological features and SVM. IOP Conference Series: Materials Science and Engineering, IOP Conference Series: Materials Science and Engineering, vol. 339 (1).http://dx.doi.org/10.1088/1757-899X/339/1/012039,

Zhu, J., Wang, C., Li, C., Gao, X., Zhao, J., 2016. Dynamic alarm prediction for critical alarms using a probabilistic model. Chin. J. Chem. Eng. 24 (7), 881–885. http://dx.doi.org/10.1016/j.cjche.2016.04.017.