CHALMERS
UNIVERSITY OF TECHNOLOGY

AEBS Event Classification Under Data Heterogeneity and Limited Trusted Labels

Master's thesis in Systems and Control

Yuchen Liu
Yiming Li

DEPARTMENT OF ELECTRICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

AEBS Event Classification Under Data Heterogeneity and Limited Trusted Labels

YUCHEN LIU
YIMING LI

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Division of System and Control
Chalmers University of Technology
Gothenburg, Sweden 2026

AEBS Event Classification Under Data Heterogeneity and Limited Trusted Labels
YUCHEN LIU
YIMING LI

Advisor: Matilda Hanes, Volvo Technology AB
Supervisor and Examiner: Jonas Sjöberg, Department of Electrical Engineering

Master's Thesis 2026
Department of Electrical Engineering
Division of System and Control
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Illustration of the AEBS system in Volvo heavy-duty trucks, showing the environment from which driving data are collected for AEBS event classification. The image is obtained from Volvo Trucks Images and Film Gallery (internal source).

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

AEBS Event Classification Under Data Heterogeneity and Limited Trusted Labels
YUCHEN LIU
YIMING LI
Department of Electrical Engineering
Chalmers University of Technology

Abstract

Advanced Emergency Braking System (AEBS) interventions can be classified according to whether the braking response was justified by the surrounding traffic situation. Accurate event classification is important for vehicle safety and driver trust, while fleet-scale assessment is challenging because most events are labeled by deterministic signal-based scripts and trusted video labels are scarce, costly, and privacy-sensitive. In addition, event data may exhibit distributional heterogeneity across operating contexts. This thesis investigates AEBS full-braking event classification under these challenges of data heterogeneity and limited trusted labels.
The study uses logged full-braking events from Volvo heavy-duty vehicles from different countries under scarce trusted video labels. Within a Machine Learning framework, the task is formulated as binary multivariate Time-Series Classification. A one-dimensional CNN is used for supervised learning, unsupervised clustering is used to examine event structure, and Mean Teacher is used for semi-supervised learning. Federated Learning is evaluated with FedAvg and FedMeanTeacher to assess country-partitioned collaboration.
Country partitions were strongly imbalanced, with the two largest accounting for 62% of the selected events and the three smallest accounting for 9%. Script-generated outcomes agreed with 75% of the trusted FP events, while agreement with trusted TP events was 94%. Supervised classification achieved a Macro-F1 of 0.9995 on script-generated outcomes and 0.9374 on trusted video labels. FedMeanTeacher reduced the cross-country Macro-F1 standard deviation to 0.0343 compared with 0.1796 for local semi-supervised learning, while its trusted-label Macro-F1 remained lower than that of centralized learning, 0.7867 versus 0.8523. Overall, script-generated outcomes enabled near-perfect reproduction of the deterministic labeling procedure, while scarce trusted labels, especially for FP events, limit correctness-oriented conclusions.

Keywords: Advanced Emergency Braking System (AEBS), Machine Learning, Time-Series Classification, Data Heterogeneity, Federated Learning.

Acknowledgements

We would like to express our sincere gratitude to our advisor at Volvo Technology AB, Matilda Hanes, for her support and valuable suggestions throughout this thesis project.
We also thank our supervisor and examiner at Chalmers University of Technology, Jonas Sjöberg, for his constructive feedback and valuable comments on the thesis structure and report.
Finally, we are grateful to David Öst at Volvo Technology AB for his support and input, and to our other colleagues for their kindness and encouragement during the project.

Yuchen Liu and Yiming Li, Gothenburg, August 2026

List of Acronyms

AEBS	Advanced Emergency Braking System
CNN	Convolutional Neural Network
EMA	Exponential Moving Average
FedAvg	Federated Averaging
FL	Federated Learning
FP	False Positive
FPR	False Positive Rate
JSD	Jensen-Shannon Divergence
MTSC	Multivariate Time-Series Classification
Non-IID	Non-Independent and Identically Distributed
PR-AUC	Precision-Recall Area Under the Curve
TP	True Positive

Contents

List of Figures

List of Tables

1

Introduction

This thesis investigates the classification of recorded Advanced Emergency Braking System (AEBS) full-braking interventions according to the observed traffic situation. The study focuses on fleet-scale event classification under limited trusted supervision and potential data heterogeneity across countries, and examines federated learning (FL) as a distributed training strategy for geographically partitioned event data.

1.1 Motivation

Every year, approximately 1.19 million people lose their lives in road traffic crashes [1]. AEBS is one of the active safety technologies designed to reduce collision risks. It monitors the road ahead and intervenes when a potential collision is detected [2, 3]. As collision risk increases, often characterized using indicators such as Time-To-Collision (TTC) [4], the AEBS may progress through visual and acoustic pre-warnings, partial pre-braking, and full autonomous emergency braking, as illustrated in Figure 1.1. However, in highly urgent situations, the system may bypass the preceding warning and partial-braking phases and initiate the AEBS full-braking event directly [2]. The full-braking stage applies the strongest braking response and represents the most safety-critical intervention among these stages.

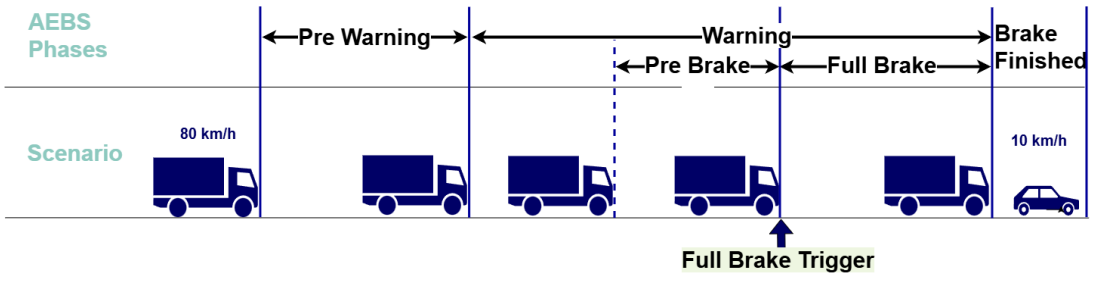

Figure 1.1: Illustration of AEBS phases and intervention progression.

However, an AEBS full-braking event may also be recorded when the observed traffic situation does not require such a strong braking response. Such unnecessary events can introduce secondary safety risks, such as rear-end collisions following unexpected braking [5, 6]. Identifying them is therefore important for evaluating

AEBS performance. In this thesis, AEBS full-braking events are classified using their event signals according to whether the braking response was justified by a genuine collision risk in the observed traffic situation. A justified event is classified as a True Positive (TP), whereas an unjustified event is classified as a False Positive (FP) [7, 8].

Fleet-scale TP/FP classification presents two main challenges. First, event signals collected under different operating conditions may follow different distributions because driving environments, traffic conditions, and other contextual factors can vary across data sources. Such distributional differences may limit the extent to which a model trained on one data source generalizes to events from other data sources [9]. Second, reliable trusted video labels for driving events are often limited, as determining whether an event is truly safety-critical may require manual review of recorded video, which is costly and time-consuming at fleet scale [7].

In addition to these data-related challenges, fleet-scale learning may also need to consider how event data are accessed and shared. When raw event data remain distributed across different data sources, collaborative learning without directly pooling the underlying data becomes relevant [10, 11].

The following section introduces the data structure through which these issues are investigated.

1.2 AEBS Event Data Overview

This study uses historical AEBS full-braking events collected from heavy-duty vehicles across multiple countries. Each event consists of a multivariate signal window centered on the AEBS full-brake trigger. In addition to the event signals, each event is associated with a country identifier and label information from two sources. The overall structure is illustrated in Figure 1.2.

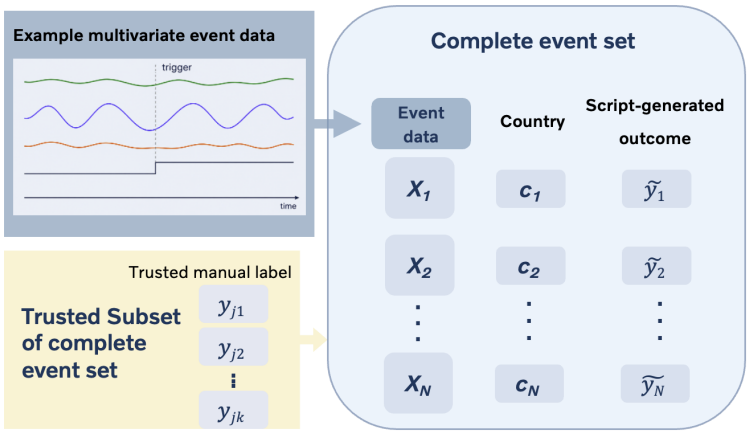

Figure 1.2: Structure of the AEBS event set, country associations, and the two label sources.

The complete event set has N events and is denoted by

$$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N),$$

where each event is represented as a fixed multivariate time window centered on the AEBS full-brake trigger, $\mathbf{X}_i \in \mathbb{R}^{T \times d}$. Here, T denotes the number of time steps in each event window and d is the number of recorded signal channels. Each event is additionally associated with the country in which it was recorded, denoted by $c_i \in \{1, \ldots, m\}$, where m is the number of countries. Events recorded in different countries may reflect different operating conditions and may therefore follow different data distributions. The country-based partitions provide a practical basis for examining such cross-country distributional differences in the event data.

The script-generated outcome associated with event i is denoted by $\tilde{y}_i$ and is available for every event.

In addition, a trusted subset

$$\mathcal{G} = \{j_1, \ldots, j_k\} \subset \{1, \ldots, N\} \tag{1.1}$$

contains k events for which manual video annotations are available, where $k \ll N$. For each $i \in \mathcal{G}$, the corresponding trusted video labels are denoted by y_i.

Together, these data characteristics link the available data to the challenges introduced above and motivate the problem formulation that follows. A detailed description of the dataset and its preparation is provided in Chapter 3.

1.3 Problem Formulation

Using the notation introduced, the primary task of this thesis is to determine whether an AEBS full-braking event represented by $\mathbf{X}_i$ constitutes a TP or an FP. Each event is assigned to one of the two classes, which makes the task a binary multivariate time-series classification (MTSC) problem. Formally,

$$f_\theta : \mathbb{R}^{T \times d} \to \{\mathrm{TP}, \mathrm{FP}\}, \qquad \hat{y}_i = f_\theta(\mathbf{X}_i). \tag{1.2}$$

where f_θ denotes a classification model with parameters θ, and $\hat{y}_i$ is the predicted class of event i.

This classification problem is studied under two conditions arising from the available data: potential heterogeneity across country-based partitions and limited trusted supervision. Specifically, each event is associated with a country c_i, while trusted video labels y_i are available only for $i \in \mathcal{G}$, whereas script-generated outcomes $\tilde{y}_i$ are available for the complete event set.

Beyond these two data conditions, the thesis also considers a distributed data-access setting. In applications where raw event data cannot be directly pooled, model training may need to operate on separate data partitions. This motivates the comparison

of pooled centralized learning, isolated local learning, and simulated federated learning, in which model updates are aggregated without directly pooling the underlying event data.

Accordingly, the analysis proceeds in three stages: first assessing whether and how the event data differ across countries, then evaluating the reliability of the available supervision, and finally examining learning strategies under the resulting data conditions. Based on this formulation, the thesis addresses the following research questions:

- **RQ1:** To what extent do the selected country partitions exhibit differences in event volume and signal distributions?
- **RQ2:** To what extent do script-generated outcomes agree with trusted video labels, and what are the implications for model evaluation and supervision?
- **RQ3:** Based on the data characteristics and label-reliability findings obtained in RQ1 and RQ2, how can the available event data be used for TP/FP classification under limited trusted supervision, and what are the effects of centralized, local, and federated training settings?

1.4 Contributions

The main contributions are:

1. The thesis characterizes how AEBS event data vary across selected countries using differences in event volume and signal distributions. This assesses the extent to which country-based partitions reflect variation in operating context.
2. The thesis quantifies agreement between script-generated outcomes and trusted video labels, establishing the implications for model supervision and evaluation.
3. The thesis evaluates supervised, unsupervised, and semi-supervised learning under the identified data conditions, comparing centralized, isolated local, and federated training while highlighting the limitations imposed by scarce trusted video labels.

1.5 Thesis Outline

Chapter 2 introduces the background for AEBS event classification, data heterogeneity, label reliability, limited trusted labels, and federated learning. Chapter 3 describes the AEBS event data, label sources, country-based partitions, and data preparation. Chapter 4 analyzes cross-country heterogeneity, while Chapter 5 evaluates agreement between script-generated outcomes and trusted video labels. Chapter 6 presents supervised, unsupervised, and semi-supervised learning under central-

ized, isolated local, and federated settings. Finally, Chapter 7 summarizes the main findings, implications, limitations, and future work.

2

Background

This chapter presents the background for the concepts and methods used in this thesis. It introduces AEBS functions and event classification, explains how the task is treated as a machine-learning problem, and discusses potential differences across country-based partitions and the availability and reliability of event labels. It then introduces learning approaches under limited trusted supervision, federated learning for distributed data, and the evaluation metrics used in the thesis.

2.1 AEBS Context

For the heavy-duty vehicle application considered in this thesis, AEBS primarily addresses potential rear-end conflicts involving objects ahead of the vehicle, using information from radar and camera sensors [3]. Based on these inputs, the system selects the forward object most relevant to the intervention logic and evaluates whether it represents a collision risk. Once an AEBS full-braking event has been recorded at the full-brake trigger, the subsequent assessment task is to determine whether the event was justified under the observed traffic situation, following the TP/FP formulation established in the Introduction.

Recorded AEBS interventions may be assessed manually or through automated methods. Automated assessment can be implemented using predefined rules or machine-learning models. Rule-based approaches are transparent and efficient but remain constrained by the signals, thresholds, and assumptions encoded in their decision logic [8]. Machine-learning methods provide a data-driven alternative by learning classification patterns from recorded events.

2.2 Machine-Learning Models and Settings

Machine-learning methods can be used for AEBS event classification by learning relationships between recorded event signals and TP/FP outcomes. In this setting, the distinction between TP and FP events may depend on complex patterns across vehicle dynamics, driver responses, system states, and target information.

2.2.1 Machine Learning Models

The model choice follows the event representation:

1. **Aggregated event representation.** Summarizing the temporal signals into event-level features produces a fixed-length feature vector for each event. This representation is used in the unsupervised clustering analyses to examine whether AEBS events form distinguishable groups in the selected feature space.

2. **Sequential event representation.** Retaining the complete signal window preserves the MTSC formulation established in the Introduction. This representation requires the model to learn patterns across signal channels and time [12, 13]. A one-dimensional CNN is suitable for the sequential setting because it can learn local temporal patterns while processing multiple signal channels jointly [12, 14].

2.2.2 Centralized Supervised Learning Settings

Centralized supervised learning provides the baseline setting for the learning methods considered in this thesis. It combines two aspects of model training: labeled examples provide the learning signal, while all available training samples are pooled into a single dataset for model development.

In supervised learning, a model is trained using paired inputs and labels to predict whether an AEBS full-braking event corresponds to a TP or FP event [15]. Its effectiveness depends on both the availability and the reliability of the training labels. In the centralized setting, the pooled dataset is used to train a single global model. Pooling allows the model to learn from the vehicles, operating conditions, and event distributions represented in the dataset. This setting provides a reference for evaluating the subsequent learning methods. Centralized training requires the relevant data to be accessible within a common training environment, an assumption that may not hold when raw data cannot be directly pooled [10, 11]. Subsection 2.5.3 introduces training settings that relax this assumption.

2.3 Data Heterogeneity

A common assumption in machine learning is that training and test data are sampled from the same underlying distribution. However, data collected under different operating conditions may follow different distributions because factors such as road infrastructure, traffic conditions, fleet usage, and driving behavior can vary across operating contexts [8].

In this thesis, AEBS events are partitioned according to the country in which they were recorded. These country-based partitions are used as a practical proxy for variation in operating context. Country itself does not identify the underlying causes of any observed differences; rather, the partitions provide a basis for examining whether and how the event data vary across countries.

Two forms of potential cross-country variation are particularly relevant:

1. **Quantity imbalance.** The number of recorded events may differ across country-based partitions. If substantial imbalance is present, larger partitions can have a stronger influence on pooled estimates and on sample-size-weighted federated aggregation.

2. **Feature distribution shift.** The distribution of the event signals may differ across country-based partitions, such that

$$p(\mathbf{X} \mid c = r) \neq p(\mathbf{X} \mid c = s) \tag{2.1}$$

for some countries r and s. Such differences may occur, for example, in vehicle speed, braking behavior, yaw rate, or target-object motion [11, 16].

If present, these forms of variation can affect both model training and evaluation. When data from multiple countries are pooled, larger partitions may have a stronger influence on the learned model, while patterns associated with smaller partitions may be underrepresented. Differences in feature distributions may also reduce how consistently a model trained primarily on one operating context generalizes to others [11, 16].

Aggregate performance may therefore conceal differences in model behavior across country-based partitions. Examining performance separately across countries helps assess whether a model generalizes consistently across the represented operating contexts.

2.4 Label Sources and Reliability

The script-generated outcomes and trusted video labels introduced in the Introduction provide two sources of supervision with different levels of semantic reliability. Whether an intervention should be classified as TP or FP depends on the actual traffic situation, whereas vehicle logs provide only indirect and potentially incomplete evidence of that situation [7, 8]. Agreement with signal-based rules therefore does not necessarily establish that the braking was justified.

Rule-based scripts usually draw on vehicle kinematics, driver inputs, system states, and target-related quantities, such as speed, deceleration, steering or yaw behavior, braking input, target distance, relative speed, and time-to-collision [8, 17]. The resulting script-generated outcomes are consistent with the predefined decision logic but remain constrained by the selected signals, thresholds, and assumptions.

Manual video review provides more direct scene-level evidence. Reviewers can examine the relevant object, its motion relative to the ego vehicle, whether the relative trajectories indicate a plausible conflict, and contextual factors such as lane geometry, cut-in or crossing behavior, roadside objects, and surrounding traffic responses [7, 17]. Video-based judgments may still be uncertain because of occlusion, insufficient recording quality, or changes caused by the intervention itself. Manual

review is also time-consuming and costly, which limits the number of events that can be annotated with trusted scene-level labels.

These differences affect model training and evaluation as well. Training primarily on script-generated outcomes may reproduce systematic assumptions or errors in the labeling procedure. Evaluation against these outcomes mainly measures consistency with the encoded decision logic. It provides limited evidence about whether the braking was justified. Evaluation against trusted video labels is more closely aligned with the intended TP/FP distinction, although its reliability is limited by the coverage of the annotated subset and possible annotation ambiguity. The two label sources differ in both coverage and semantic reliability.

2.5 Learning Settings under Limited Trusted Labels and Distributed Data

In the distributed-learning settings considered in this thesis, each country-based partition is treated as one client. When only a small subset of events has trusted video labels, supervised learning based solely on this subset may be limited by its size and coverage.

This problem has two dimensions in the present thesis. The first is label availability: the large event pool is weakly labeled, while trusted video labels are scarce. This motivates unsupervised and semi-supervised learning. The second is data location: event data may be distributed across clients and may not always be pooled freely. This motivates isolated local and federated learning settings.

2.5.1 Unsupervised Learning

Unsupervised learning analyzes input data without using class labels during training and aims to identify structure in the data [15]. A common approach is clustering, in which samples are grouped according to similarity in a selected feature space [18]. In this thesis, clustering is used to explore whether AEBS events form distinguishable groups based on their feature representations, as illustrated in Figure 2.1.

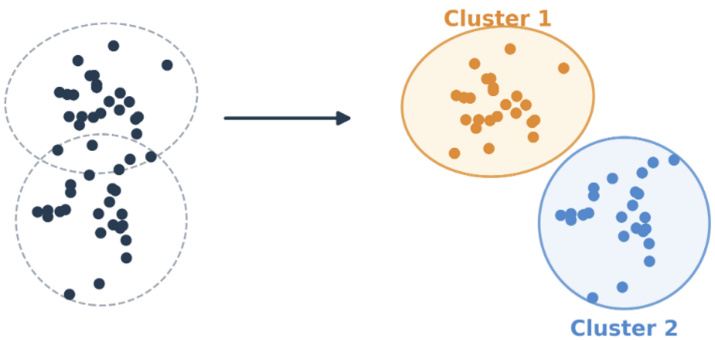

Figure 2.1: Illustration of unsupervised clustering

Because class labels are not used during clustering, the resulting clusters do not directly represent TP and FP classes. Their relationship to the classification target must therefore be examined after clustering using the available labels. Clustering is used here to explore data structure rather than as a direct TP/FP classification method.

2.5.2 Semi-Supervised Learning

Semi-supervised learning combines a limited set of reliably labeled samples with a larger set of unlabeled or weakly labeled data [19, 20]. In the AEBS setting considered here, trusted video labels provide reliable supervision, while script-generated outcomes are treated as weaker auxiliary labels.

Mean Teacher is a consistency-based semi-supervised approach consisting of student and teacher models [21]. As illustrated in Figure 2.2, the student is trained using supervised loss on trusted labeled events, together with auxiliary script loss on events without trusted video labels. In addition, a consistency loss encourages the student and teacher to produce similar predictions for the same event under perturbation. The teacher is not optimized directly; instead, its parameters are updated as an exponential moving average (EMA) of the student parameters.

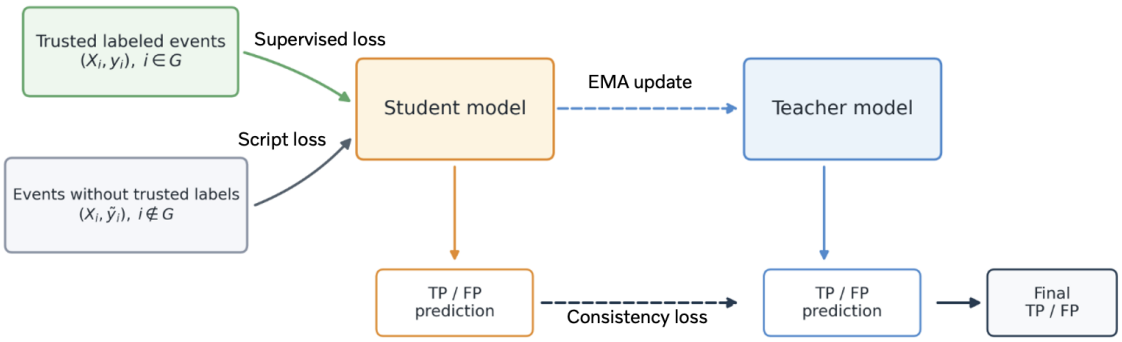

Figure 2.2: Conceptual illustration of Mean Teacher learning with trusted supervision.

This framework allows the larger event set to contribute to training without treating script-generated outcomes as equivalent to trusted video labels. The availability of trusted video labels and the location of the training data remain separate aspects of the learning problem; the semi-supervised setting can therefore be applied under centralized, isolated local, or federated training.

2.5.3 Federated Learning

The data examined in this thesis could be pooled within the restricted research environment. Federated learning is therefore evaluated as a simulated prospective setting in which similar fleet data are assumed to remain distributed across geographical or

organizational boundaries. Under isolated local learning, each client trains its own model using only locally available data, without data pooling or model aggregation across clients. This preserves data locality but prevents clients from benefiting from information available in other client datasets. Federated learning provides a collaborative alternative in which multiple clients train a shared model without directly pooling their raw data [10, 11]. As illustrated in Fig. 2.3, the server distributes a global model to the clients, each client updates the model using its local data, and the resulting model updates are returned to the server for aggregation.

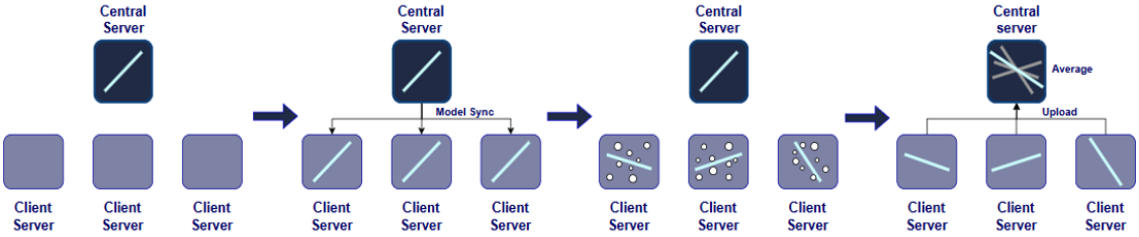

Figure 2.3: Illustration of a federated learning pipeline.

A widely used aggregation method is Federated Averaging (FedAvg) [10]. In the federated setting considered here, each country-based partition is treated as one client. Let n_r denote the number of training samples available at client r, where $r \in \{1, \ldots, K\}$, and let

$$n = \sum_{r=1}^{K} n_r.$$

Here, K is the number of selected country clients. Assuming that all K clients participate in communication round t, the global model parameters are updated after local training as

$$\boldsymbol{\theta}^{(t+1)} = \sum_{r=1}^{K} \frac{n_r}{n} \boldsymbol{\theta}_r^{(t+1)}, \tag{2.2}$$

where $\boldsymbol{\theta}_r^{(t+1)}$ denotes the locally updated parameters of client r. Thus, clients with larger local datasets contribute more strongly to the aggregated global model. When the local training procedure is semi-supervised, each client applies the supervised objective to its locally available trusted video labels and the consistency objective to its local events without trusted video labels before the resulting model parameters are aggregated.

2.6 Evaluation Metrics

AEBS trigger-event classification is treated as a binary classification problem. Following the TP/FP definitions established in the Introduction, AEBS-TP is treated as the positive class and AEBS-FP as the negative class. Classification performance can therefore be evaluated using both class-specific and class-balanced metrics.

A confusion matrix summarizes classification outcomes using four counts [22]. To distinguish these outcomes from the AEBS event classes, N_{TP}, N_{FP}, N_{TN}, and N_{FN} denote true-positive, false-positive, true-negative, and false-negative prediction counts, respectively. Under this class convention, N_{FP} denotes an actual AEBS-FP event incorrectly predicted as AEBS-TP, whereas N_{TN} denotes an AEBS-FP event correctly classified as such.

Accuracy measures the overall proportion of correctly classified events:

$$\mathrm{Accuracy} = \frac{N_{\mathrm{TP}} + N_{\mathrm{TN}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}} + N_{\mathrm{TN}} + N_{\mathrm{FN}}}.$$

Because accuracy assigns equal weight to every sample, it can obscure poor performance on a less frequent class when the class distribution is imbalanced [23].

For the positive AEBS-TP class, precision and recall are defined as [22]

$$\mathrm{Precision}_{\mathrm{TP}} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}}, \qquad \mathrm{Recall}_{\mathrm{TP}} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}}.$$

The AEBS-TP F_1-score is their harmonic mean:

$$F_{1,\text{AEBS-TP}} = \frac{2 \cdot \mathrm{Precision}_{\mathrm{TP}} \cdot \mathrm{Recall}_{\mathrm{TP}}}{\mathrm{Precision}_{\mathrm{TP}} + \mathrm{Recall}_{\mathrm{TP}}}.$$

This metric summarizes the model's ability to identify justified AEBS activations while accounting for both missed AEBS-TP events and incorrect AEBS-TP predictions.

To reflect performance on both event classes, macro-averaged F_1 (macro-F_1) is also reported [22]. It is obtained by calculating the F_1-score separately for AEBS-TP and AEBS-FP and assigning equal weight to the two classes:

$$\text{Macro-}F_1 = \frac{1}{2}\left(F_{1,\text{AEBS-TP}} + F_{1,\text{AEBS-FP}}\right).$$

Unlike accuracy, macro-F_1 is not dominated by the more frequent class. A high AEBS-TP F_1-score may coexist with poor recognition of AEBS-FP events, whereas macro-F_1 reflects performance on both classes.

Under the class convention used here, the false-positive rate (FPR) is the proportion of actual AEBS-FP events that are incorrectly classified as AEBS-TP [24]:

$$\mathrm{FPR} = \frac{N_{\mathrm{FP}}}{N_{\mathrm{FP}} + N_{\mathrm{TN}}}.$$

A lower FPR therefore indicates better recognition of unjustified or unnecessary activations. Equivalently, FPR is one minus the recall of the AEBS-FP class.

3

AEBS Event Data

This chapter describes the AEBS full-braking event data used in this thesis. It provides the detailed data description that was introduced at a higher level in Section 1.2, including event structure and representations, label sources, dataset selection, usage, and data splits.

3.1 Event Data Structure and Representations

The complete dataset contains $N = 41{,}947$ logged AEBS full-braking events. Each event is stored as a fixed window of 150 time steps, with the AEBS full-brake trigger aligned near the center to separate the pre-trigger and post-trigger portions. Each window contains 21 anonymized signal channels covering vehicle, driver, target, AEBS-state, and event-context information. Accordingly, event (i) is represented sequentially as $\mathbf{X}_i \in \mathbb{R}^{150\times 21}$. Due to industrial confidentiality, exact proprietary signal names are not disclosed in this public report. Instead, the signals are described using the functional groups summarized in Table 3.1.

Table 3.1: Anonymized signal groups in the sequential event representation.

Signal group	**Channels**	**Representative information**
Vehicle motion	3	Ego speed, acceleration, and yaw-related motion
Driver input	3	Accelerator, brake, and steering input
Target information	6	Relative longitudinal and lateral position, relative velocity, acceleration, and target motion state
AEBS warning state	4	AEBS warning and intervention-related states
Event context	5	System status, trailer information, and event-level attributes
Total	**21**	

Unlike the four time-varying signal groups, the event-context channels are constant within each event and are repeated across the 150 time steps when constructing $\mathbf{X}_i$ to preserve a fixed sequential representation.

In addition to the sequential representation, an aggregated event representation,

$$\mathbf{z}_i \in \mathbb{R}^{96},$$

is constructed from the sequential representation using summary statistics and trigger-relative features. Specifically, these include minimum, maximum, and mean values, separate pre- and post-trigger means, and slope-related features around the trigger. Figure 3.1 summarizes the relation between the sequential and aggregated representations.

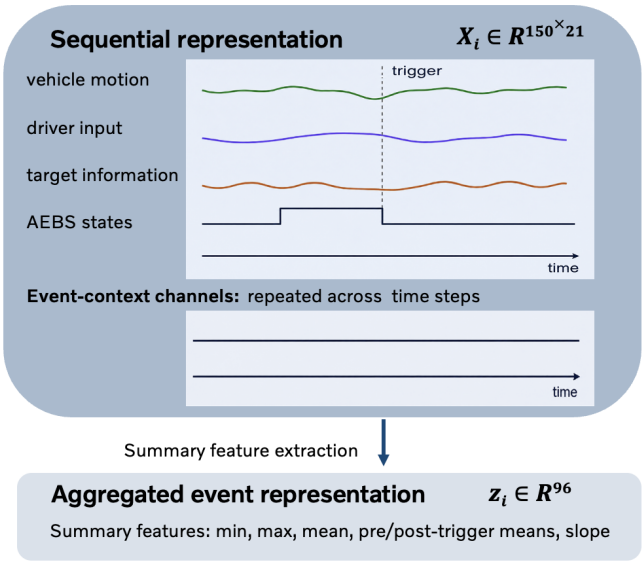

Figure 3.1: Sequential and aggregated representations of an AEBS full-braking event.

In the subsequent analyses, $\mathbf{X}_i$ is used when the temporal structure of the event is retained, whereas $\mathbf{z}_i$ provides an event-level tabular representation when the full signal sequence is not required.

3.2 Label Sources

Two label sources are associated with the AEBS events: script-generated outcomes and trusted video labels.

A script-generated outcome is available for every event,

$$\tilde{y}_i \in \{\mathrm{TP}, \mathrm{FP}, \mathrm{neither}\}.$$

These outcomes are produced by an existing deterministic rule-based procedure that applies predefined thresholds and logical conditions to recorded driver-braking and target-persistence signals. TP is assigned when the recorded signals indicate sustained driver braking together with persistence of the relevant target after the

AEBS intervention. FP is assigned when the recorded signals satisfy a separate set of predefined FP conditions associated with insufficient driver-braking evidence or insufficient target persistence. Events that satisfy neither the TP nor FP rule set receive the *neither* outcome, which is used only in the label-reliability analysis and is excluded from the main binary TP/FP event set. The exact thresholds and detailed decision logic are not disclosed in this public report.

Trusted video labels are available for 73 events,

$$y_i \in \{\mathrm{TP}, \mathrm{FP}\}, \qquad i \in \mathcal{G},$$

These labels were obtained through manual review of synchronized front-facing video. Two annotators independently assessed whether each AEBS full-braking event was justified by the observed traffic situation, and disagreements were resolved through review and consensus. The resulting labels are treated as the trusted reference for the TP/FP classification task. The videos are used only for manual assessment and are not reproduced in this thesis.

The two label sources therefore differ in both coverage and assessment procedure. Script-generated outcomes are available for the complete 41,947 events, whereas trusted video labels are available only for the 73 manually reviewed events.

3.3 Dataset Selection, Usage, and Data Splits

The complete event pool contains 41,947 AEBS full-braking events, including the 73 events in the trusted subset $\mathcal{G}$. For the classification analyses, the data were restricted to 8 selected countries and to events with binary script-generated outcomes, i.e., TP or FP.

Among the m countries represented in the complete event pool, 8 were selected in consultation with the industrial partner based on practical relevance and data coverage. Their identities are not disclosed in this public report and are denoted as c_1 to c_8. After the country selection and exclusion of events with the script-generated *neither* outcome, 21,740 events remain, including 40 events from $\mathcal{G}$. The resulting dataset composition is summarized in Table 3.2.

Table 3.2: Dataset composition before and after country and binary-outcome selection.

Dataset stage	Events	Trusted subset	Description
Complete event pool	41,947	73	All AEBS full-braking events
Selected binary event set	21,740	40	Events from the 8 selected countries with TP/FP script-generated outcomes

The selected binary event set of 21,740 events is used for the cross-country heterogeneity analysis, while the label-reliability analysis uses all 73 events in $\mathcal{G}$.

For the learning experiments, the 40 events with trusted video labels contained in the selected binary event set are separated from the remaining 21,700 events. The 21,700 events are used for the supervised and unsupervised learning experiments. The trusted events are excluded from the supervised diagnostic baseline so that the trusted-label evaluation subset is not also used as ordinary training data. They are retained in the semi-supervised setting, where their trusted labels provide the scarce supervision that the method is designed to exploit. Thus, the semi-supervised setting combines the 21,700 events with the 40 trusted video-labelled events. Dataset usage for each subsequent task is summarized in Table 3.3.

Table 3.3: Dataset usage across the subsequent analyses and learning settings.

Analysis or learning setting	**Events**	**Data used**	**Chapter**
Cross-country heterogeneity analysis	21,740	Selected binary event set across eight countries	Chapter 4
Label-reliability analysis	73	Events in $\mathcal{G}$ with both $\tilde{y}_i$ and y_i	Chapter 5
Supervised learning	21,700	Binary events with script-generated TP/FP outcomes	Chapter 6
Unsupervised learning	21,700	Binary events without using labels during clustering	Chapter 6
Semi-supervised learning	21,740	21,700 events together with 40 trusted video labels from $\mathcal{G}$	Chapter 6

To support independent model training, validation, and final evaluation, the datasets were divided into separate training, validation, and test subsets. Splitting was performed at the event level using a fixed random seed, ensuring that all representations derived from the same event remained in the same subset.

The 21,700 binary events were split 60/20/20, while the 40 events with trusted video labels were split separately 50/15/35 because of their limited number.

4

Cross-Country Heterogeneity Analysis

This chapter analyzes heterogeneity across the selected country partitions. The analysis addresses RQ1 by examining country-partition size imbalance and differences in signal distributions.

4.1 Heterogeneity Assessment

Following the dataset selection described in Section 3.3, this analysis uses the selected binary event set containing $N_{\text{sel}} = 21{,}740$ AEBS full-braking events from $K = 8$ selected countries.

Let

$$\mathcal{D}_r = \{i : c_i = r\}, \qquad r \in \{1, \ldots, K\},$$

denote the index set of events associated with country r. The number of events in country r is

$$n_r = |\mathcal{D}_r|.$$

Its proportion of the selected event set is

$$q_r = \frac{n_r}{N_{\text{sel}}}, \qquad \sum_{r=1}^{K} q_r = 1. \tag{4.1}$$

Country partition size imbalance: The values of n_r and q_r were used to describe the event-volume distribution across the K country partitions.

Signal distribution differences: Let $x_{i,t,f}$ denote the value of signal channel f at time step t for event i, where $t \in \{1, \ldots, 150\}$ and $f \in \{1, \ldots, 21\}$. For each country partition, the mean and standard deviation were calculated at each time step as

$$\mu_{r,t,f} = \frac{1}{n_r} \sum_{i \in \mathcal{D}_r} x_{i,t,f}, \tag{4.2}$$

and

$$\sigma_{r,t,f} = \sqrt{\frac{1}{n_r - 1} \sum_{i \in \mathcal{D}_r} (x_{i,t,f} - \mu_{r,t,f})^2}. \tag{4.3}$$

These statistics describe the temporal trajectories and within-partition variability of the signal channels. The representative signal channel reported in this chapter is vehicle speed.

Jensen–Shannon divergence (JSD) was used to quantify differences between the signal distributions of each country partition and the pooled event set. For each country r and signal channel f, the values across all events and all 150 time steps were combined to construct a country-specific distribution $P_{r,f}$. The corresponding pooled distribution $P_{\text{pool},f}$ was constructed using all N_{sel} events. Continuous values were discretized using common histogram bins determined from the pooled data.

Let

$$M_{r,f} = \frac{1}{2} \left(P_{r,f} + P_{\text{pool},f} \right). \tag{4.4}$$

The JSD between country r and the pooled distribution for signal channel f is defined as

$$\begin{aligned} J_{r,f} &= \text{JSD} \left(P_{r,f} \parallel P_{\text{pool},f} \right) \\ &= \frac{1}{2} \text{KL} \left(P_{r,f} \parallel M_{r,f} \right) + \frac{1}{2} \text{KL} \left(P_{\text{pool},f} \parallel M_{r,f} \right). \end{aligned} \tag{4.5}$$

The maximum divergence for each country partition was calculated across all 21 signal channels:

$$J_r^{\max} = \max_{f \in \{1,\ldots,21\}} J_{r,f}. \tag{4.6}$$

A larger value of $J_r^{\max}$ indicates a stronger divergence between the corresponding country partition and the pooled signal distribution.

4.2 Results

Figure 4.1 shows the event-volume distribution across the eight anonymized country partitions. The distribution is highly imbalanced: c_1 and c_2 together account for 62% of all events, whereas c_6, c_7, and c_8 together account for only 9%.

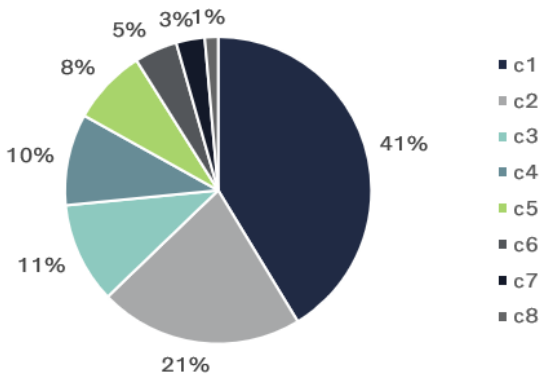

Figure 4.1: Event-volume distribution across eight anonymized country partitions.

Beyond event-volume imbalance, the selected country partitions also differ in their signal distributions. Figure 4.2 shows the temporal mean and variability of vehicle speed across the selected country partitions. Each solid line represents the mean vehicle speed of one country partition at each time step, while the shaded region represents the corresponding standard deviation.

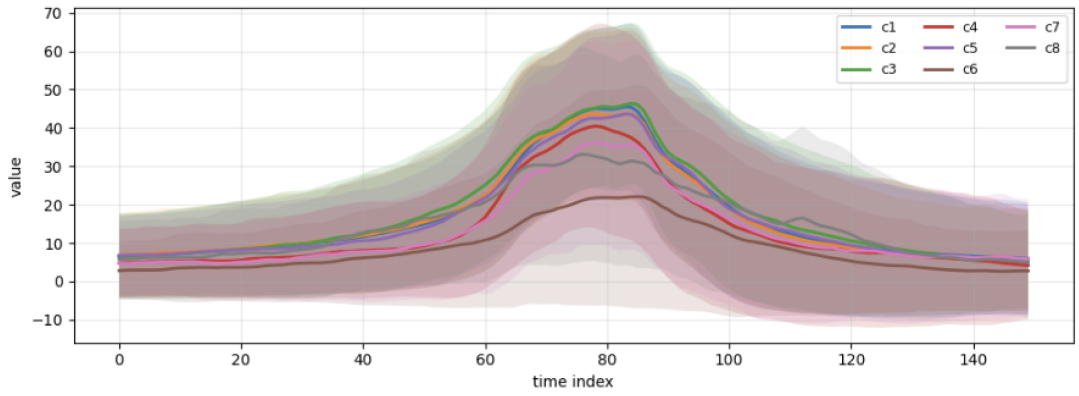

Figure 4.2: Distribution of vehicle speed across selected country partitions.

The vehicle-speed trajectories are relatively stable during the first 40 time steps. They then increase toward the center of the event window, where the AEBS trigger is aligned, and reach their highest levels around time indices 75–85. The country partitions c_1 to c_5 reach higher peak values, approximately between 40 and 47 in the plotted values. In comparison, c_6 remains substantially lower, with a peak close to 22. The trajectories of c_7 and c_8 show intermediate behavior. These differences indicate that the selected country partitions do not share identical vehicle-speed distributions.

Figure 4.3 summarizes the maximum JSD across the 21 signal channels for each country partition.

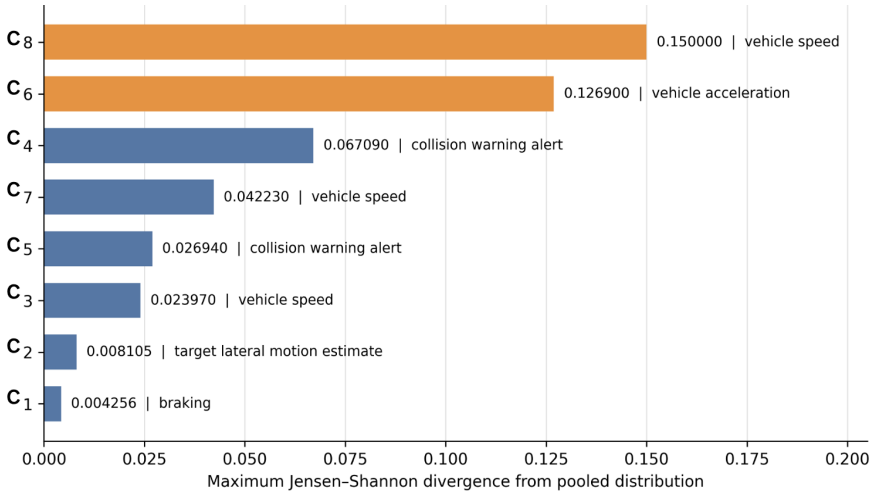

Figure 4.3: Maximum Jensen–Shannon divergence across the 21 signal channels for each anonymized country partition.

Most country partitions exhibit relatively small maximum JSD values, whereas c_8 and c_6 show the largest deviations from the pooled distribution. The largest divergence for c_8 is 0.1500, associated with vehicle speed, while the largest divergence for c_6 is 0.1269, associated with vehicle acceleration. The JSD results therefore provide quantitative evidence of signal-distribution differences across the selected country partitions.

4.3 Discussion

The results demonstrate measurable cross-country heterogeneity in both partition size and signal distributions. In particular, the selected country partitions differ substantially in event volume, while the temporal and JSD analyses show that their signal distributions are not identical.

Country is used as a proxy for operating context. The observed differences may reflect factors such as road infrastructure, traffic conditions, fleet usage, and driving behavior, but the present analysis does not isolate their individual contributions. The results should therefore be interpreted as descriptive evidence of cross-country heterogeneity rather than as evidence of causal country effects.

Together, the partition-size imbalance and signal-distribution differences establish the data conditions considered in the subsequent learning experiments. Retaining the country partitions allows model performance to be evaluated across differently sized and distributed data sources, while also providing the basis for the distributed learning settings examined in Chapter 6.

5

Label Reliability Analysis

This chapter analyzes the reliability of the script-generated outcomes by comparing them with trusted video labels on the trusted subset. The analysis addresses RQ2 and motivates evaluation against trusted video labels in the learning experiments.

5.1 Label Assessment

Following the dataset description in Section 3.3, this analysis uses the trusted subset $\mathcal{G}$, which contains 73 AEBS full-braking events with both script-generated outcomes and trusted video labels. The notation $\tilde{y}_i$ for the script-generated outcome and y_i for the trusted video label is defined in Section 3.2.

The agreement between the two label sources was assessed by comparing $\tilde{y}_i$ with y_i for all $i \in \mathcal{G}$. Agreement was summarized using row-normalized percentages of the three script-generated outcomes within each trusted video label class.

The row-normalized percentage is defined as

$$\begin{aligned} p_{a,b} &= 100\,\frac{|\{i \in \mathcal{G} : y_i = a,\ \tilde{y}_i = b\}|}{|\{i \in \mathcal{G} : y_i = a\}|}, \\ a &\in \{\mathrm{TP}, \mathrm{FP}\}, \\ b &\in \{\mathrm{TP}, \mathrm{FP}, \mathrm{neither}\}. \end{aligned} \tag{5.1}$$

To examine the nature of disagreements beyond the aggregate percentages, the disagreement cases were defined as

$$\mathcal{A} = \{i \in \mathcal{G} : \tilde{y}_i \neq y_i\}\,.$$

All events in $\mathcal{A}$ were reviewed using the synchronized front-facing video and the corresponding event signals. The qualitative review focused on scene context, target relevance, driver braking, and target persistence.

5.2 Results

Agreement was high for trusted TP events, with 94% receiving a script-generated TP outcome. Trusted FP events showed a less consistent pattern: only 75% received a script-generated FP outcome, while 13% received a script-generated TP outcome and 13% received a script-generated *neither* outcome.

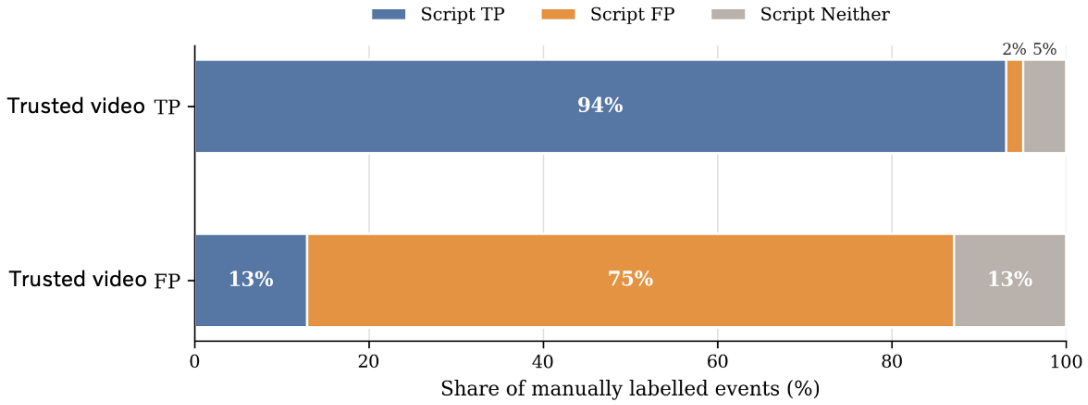

Figure 5.1: Trusted video labels and script-generated outcomes, reported as rounded row-normalized percentages.

The qualitative review of disagreement cases identified different patterns for the two trusted video label classes. Among trusted TP events that did not receive a script-generated TP outcome, the videos often showed safety-relevant situations involving vulnerable road users, such as pedestrians or cyclists, as well as intersections or cut-ins. These cases could be missed when recorded driver braking was short, weak, or partly completed before the AEBS full-brake trigger, or when the relevant target was present only briefly.

Among trusted FP events, disagreements were more often associated with geometrically complex or cluttered scenes, including yard areas, curved roads, overpasses, stationary objects, and adjacent-lane targets. In these cases, driver braking and target-related signals could produce patterns resembling TP events, while the video context did not show a clearly relevant in-path collision threat.

5.3 Discussion

The rule-based procedure and the trusted video labeling procedure use different types of evidence. Script-generated outcomes are derived from recorded signal conditions, whereas trusted video labels are obtained through manual review of synchronized front-facing video to judge whether the intervention was warranted in the observed traffic scene. Disagreement can therefore occur when similar signal patterns correspond to different scene-level interpretations.

This distinction affects how model performance should be interpreted. High performance against script-generated outcomes mainly reflects consistency with the deterministic labeling procedure. It does not establish whether an AEBS full-braking event was justified in the observed traffic scene. Correctness-oriented evaluation should therefore be based on trusted video labels.

The stronger disagreement observed for trusted FP events suggests that additional trusted video annotation should prioritize FP-related and ambiguous cases. For RQ2, the results show that script-generated outcomes can provide scalable weak supervision, while correctness-oriented TP/FP assessment requires trusted video labels. This motivates evaluation against trusted video labels and the experiments with limited trusted supervision in Chapter 6.

6

Learning with Limited Trusted Labels

This chapter studies AEBS full-braking event classification with limited trusted video labels and a larger pool of script-generated outcomes. The analysis addresses RQ3 by comparing supervised and unsupervised baselines and by evaluating semi-supervised learning under centralized, isolated local, and federated training settings. The analysis progresses from simple learning baselines to methods that use additional events and distributed collaboration under limited trusted video labels.

6.1 Learning Approaches

The data selection, label sources, event representations, and data splits used in this chapter are described in Section 3.3. Building on the heterogeneity and label-reliability findings from Chapters 4 and 5, this chapter focuses on how the available data are used under limited trusted supervision.

Supervised learning is first used to establish a baseline based on script-generated outcomes. Unsupervised structure analysis then examines the event data without using class labels. Semi-supervised learning subsequently combines the limited trusted video labels with the larger set of script-generated outcomes. Finally, the semi-supervised approach is evaluated under centralized, isolated local, and federated training settings.

6.1.1 Supervised Diagnostic Baselines

The supervised diagnostic baseline used a centralized one-dimensional CNN with the sequential event representation $\mathbf{X}_i$. The model was trained to reproduce the script-generated TP and FP outcomes.

The CNN architecture used in the sequence-based settings is shown in Figure 6.1. Exact hyperparameters are reported in Appendix A.

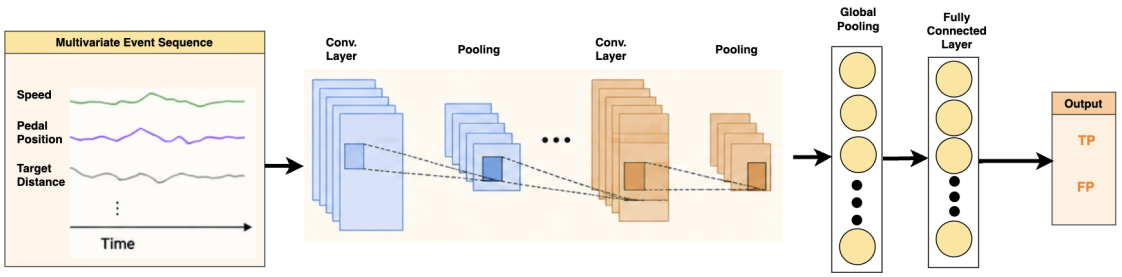

Figure 6.1: Schematic architecture of the sequence-based 1D CNN used for AEBS event classification.

6.1.2 Unsupervised Structure Discovery

K-means and hierarchical clustering were fitted centrally using the standardized aggregated event representations $\mathbf{z}_i \in \mathbb{R}^{96}$. Let $\mathcal{U}$ denote the 21,700 events used for clustering, and let $\hat{c}_i \in \{1, 2\}$ denote the cluster assignment of event $i \in \mathcal{U}$. Labels were not used during clustering.

The script-generated outcome composition of cluster j was summarized as

$$\pi_{j,a} = \frac{|\{i \in \mathcal{U} : \hat{c}_i = j,\ \tilde{y}_i = a\}|}{|\{i \in \mathcal{U} : \hat{c}_i = j\}|}, \qquad a \in \{\mathrm{TP}, \mathrm{FP}\}, \quad j \in \{1, 2\}.$$

For the trusted video label comparison, the 40 events with trusted video labels contained in $\mathcal{G}$ were assigned to the nearest cluster using the fitted clustering solution. Each cluster was then associated with the TP or FP class that was dominant among the assigned trusted video labels. The resulting post hoc cluster-to-class alignment was used as a descriptive measure of correspondence between the discovered clusters and the trusted TP/FP distinction. Trusted video labels were used only after clustering for cluster mapping and alignment evaluation.

6.1.3 Semi-Supervised Mean Teacher

Semi-supervised learning used the 40 events with trusted video labels together with the remaining 21,700 events in the selected binary event set. Table 6.1 summarizes how the Mean Teacher setup used trusted labels, script-generated outcomes, and unlabeled structure.

Table 6.1: Mean Teacher semi-supervised experiment settings.

Component	Setting
Trusted supervision	Trusted video labels y_i were used in $\mathcal{L}_{\text{trusted}}$.
Weak supervision	Script-generated TP/FP outcomes were used in $\mathcal{L}_{\text{script}}$ only for events outside the trusted subset.
Consistency contribution	Events without trusted video labels contributed through the consistency loss.
Input perturbations	Noise injection and channel dropout were applied to create perturbed views.

The student-model objective was

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{trusted}} + \gamma_{\text{script}}\mathcal{L}_{\text{script}} + \mathcal{L}_{\text{cons}}, \tag{6.1}$$

where $\mathcal{L}_{\text{trusted}}$ is the trusted video label loss, $\mathcal{L}_{\text{script}}$ is the script-generated outcome loss, and $\mathcal{L}_{\text{cons}}$ is the consistency loss.

The teacher parameters were updated after each student update using

$$\boldsymbol{\theta}_{\text{teacher}} \leftarrow \alpha\boldsymbol{\theta}_{\text{teacher}} + (1-\alpha)\boldsymbol{\theta}_{\text{student}}, \tag{6.2}$$

where α is the exponential-moving-average decay rate.

6.2 Training Settings

The centralized, isolated local, and federated supervised settings used the same sequence-based CNN architecture described in Subsection 6.1.1. In the centralized setting, data from all selected clients were pooled for training. In isolated local learning, a separate model was trained for each client. In federated learning, the clients trained local models and their parameters were aggregated using sample-size-weighted FedAvg.

The federated semi-supervised setting additionally requires coordinating local Mean Teacher training with federated aggregation and is therefore specified separately below.

6.2.1 Federated Semi-Supervised Learning

The semi-supervised framework was extended to the federated setting by applying the Mean Teacher objective defined in Section 6.1.3 within each client and aggregating the resulting model parameters using sample-size-weighted FedAvg. This method is referred to as *FedMeanTeacher*. Table 6.2 summarizes the settings specific to the federated extension.

Table 6.2: FedMeanTeacher settings.

Part	Setting
Local training	Each client used the Mean Teacher objective defined in Section 6.1.3. If no trusted video labels were available in the local training subset, $\mathcal{L}_{\text{trusted}} = 0$.
Aggregation	Student and teacher parameters were aggregated separately using sample-size-weighted FedAvg, and both aggregated models initialized the next communication round.
Evaluation model	The aggregated teacher after the final communication round was used for evaluation.

6.3 Experimental Protocol

6.3.1 Experimental Setup and Evaluation Protocol

Experimental reproducibility. Exact CNN, Mean Teacher, and federated training configurations are reported in Appendix A. Hyperparameters and model-selection decisions were determined using the corresponding validation sets defined in Section 3.3. Semi-supervised experiments were repeated using five random seeds because the trusted subset was small and imbalanced. Performance for these experiments is reported as the mean across the five runs, with variation summarized where applicable.

Evaluation scope and label sources. All classification models were evaluated at the event level using the training, validation, and test splits defined in Section 3.3.

Supervised and semi-supervised models were first evaluated against script-generated outcomes $\tilde{y}_i$, measuring consistency with the deterministic labeling procedure. Correctness-oriented evaluation was then performed using trusted video labels y_i from the 40 selected events. Because these events provide limited coverage across the clients, this evaluation was treated as a trusted reference for correctness rather than as a comprehensive assessment across all clients.

Evaluation metrics. Clustering results were summarized using cluster purity under script-generated outcome and trusted video label references, together with post hoc cluster-to-class alignment.

Supervised and semi-supervised models were evaluated using Macro-F1, AEBS-TP F1, and false positive rate (FPR), following the class convention defined in Chapter 2. The trusted video label test set contained only a small number of actual AEBS-FP events. FPR was therefore interpreted together with the number of actual AEBS-FP test events and the test-set size.

Decision threshold selection. Threshold selection was label-source specific. For script-generated outcome evaluation, every model used a threshold selected on the

script validation set. For trusted video label evaluation, every compared model used a threshold selected on the same trusted video label validation set. The test sets were not used for threshold selection. Local trusted-label results were reported only when the corresponding trusted validation and test samples were available.

For FedMeanTeacher, client-side updates did not use client-specific thresholds. One operating threshold for the aggregated global teacher was selected using the common trusted video label validation set. For the isolated local semi-supervised baseline, a client with trusted video label validation data used a locally selected threshold. Client without trusted video label validation data used the default threshold of 0.5.

Client evaluation. For distributed experiments, performance was reported on both the pooled test set and the individual clients. Each selected client was treated as one client. Clients without trusted video label test samples were excluded from the client-level trusted evaluation summary.

Let F_r denote the Macro-F1 score for client r, where $r = 1, \ldots, K$ and $K = 8$. Cross-client performance was summarized using

$$\bar{F} = \frac{1}{K} \sum_{r=1}^{K} F_r, \qquad \sigma_F = \sqrt{\frac{1}{K} \sum_{r=1}^{K} \left(F_r - \bar{F}\right)^2}. \tag{6.3}$$

The coefficient of variation and the performance gap were defined as

$$\mathrm{CV} = \frac{\sigma_F}{\bar{F}}, \qquad \mathrm{Gap} = \max_r F_r - \min_r F_r. \tag{6.4}$$

The minimum and maximum client scores were also reported where applicable. The standard deviation σ_F uses the population form because it summarizes all K clients included in the experiment.

6.4 Results

6.4.1 Supervised Diagnostic Baseline

Table 6.3 reports supervised performance on the script-generated outcome test set and, for the two global models, on the trusted video label test set.

Table 6.3: Supervised diagnostic performance on script-generated outcomes and trusted video labels.

Setting	Evaluation label	Macro-F1	AEBS-TP F1	FPR
Centralized / FedAvg	Script-generated outcome	0.9995	0.9999	0.0000
Local (own client)	Script-generated outcome	0.9828	0.9953	0.0069
Centralized / FedAvg	Trusted video label	0.9374	0.9859	0.2000

Figure 6.2 shows the variation in Macro-F1 across clients for the supervised models.

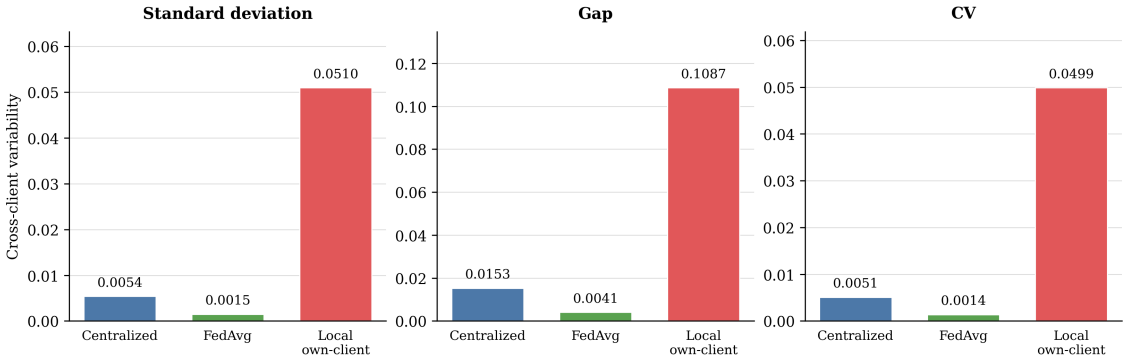

Figure 6.2: Cross-client Macro-F1 variation for supervised models on the script-generated outcome test set.

On the script-generated outcome test set, centralized training and FedAvg achieved identical aggregate performance. The Local model showed slightly lower performance. On the trusted video label test set, the two global models again achieved identical performance, with lower Macro-F1 and higher FPR than on the script-generated outcome test set.

FedAvg showed the lowest client-level variation across all three measures, with a standard deviation of 0.0015, a gap of 0.0041, and a CV of 0.0014. Local own-client training showed the largest client-level variation, with corresponding values of 0.0510, 0.1087, and 0.0499.

6.4.2 Unsupervised Structure Discovery

Figure 6.3 compares cluster purity under script-generated outcomes and trusted video labels.

Under the script-generated outcome reference, K-means produced the slightly purer TP-mapped cluster, 99.7% compared with 98.3% for hierarchical clustering. Hierarchical clustering produced the purer FP-mapped cluster, 78.0% compared with 65.4% for K-means.

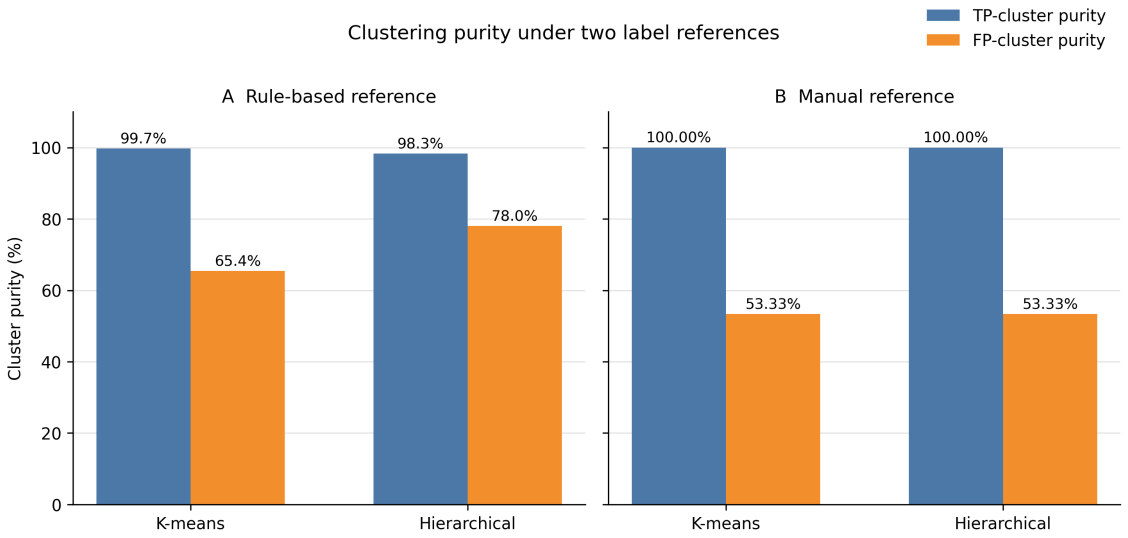

Figure 6.3: Cluster purity under script-generated outcome and trusted video label references.

Under the trusted video label reference, both methods produced the same post hoc cluster-to-class alignment score of 0.9041. Both methods achieved 100.0% TP-cluster purity and 53.33% FP-cluster purity. Thus, both clustering methods isolated a TP-dominant cluster more effectively than a distinct FP-dominant cluster in the selected representation.

6.4.3 Semi-Supervised Results

Table 6.4 reports the mean semi-supervised performance across five random seeds. Local SSL performance on the trusted video label test set is not reported because some clients contained no trusted video label test samples.

Table 6.4: Mean semi-supervised performance across five random seeds.

Setting	Evaluation label	Macro-F1	AEBS-TP F1	FPR
Centralized SSL	Script-generated outcome	0.4664	0.8158	0.8794
FedMeanTeacher	Script-generated outcome	0.7164	0.9170	0.5479
Local SSL (own client)	Script-generated outcome	0.4191	0.4926	0.3430
Centralized SSL	Trusted video label	0.8523	0.9713	0.2000
FedMeanTeacher	Trusted video label	0.7867	0.9733	0.4000

On the script-generated outcome test set, FedMeanTeacher achieved the highest Macro-F1 and AEBS-TP F1 among the semi-supervised models, with a lower FPR than Centralized SSL. Local SSL achieved the lowest FPR, together with lower Macro-F1 and AEBS-TP F1. All semi-supervised models obtained lower Macro-F1 scores than their corresponding supervised baselines in Table 6.3.

On the trusted video label test set, Centralized SSL achieved higher Macro-F1 and lower FPR than FedMeanTeacher. FedMeanTeacher achieved a slightly higher AEBS-TP F1. Local SSL was not included because trusted video label test samples were unavailable for some client.

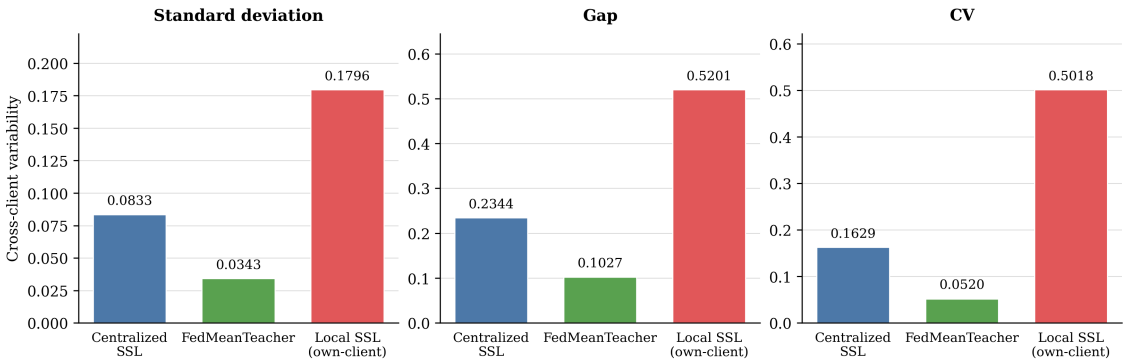

Figure 6.4: Cross-client Macro-F1 variation for semi-supervised models on the script-generated outcome test set.

Figure 6.4 shows that FedMeanTeacher had lower variation than Local SSL across all three measures. Its standard deviation, gap, and CV were 0.0343, 0.1027, and 0.0520, respectively, compared with 0.1796, 0.5201, and 0.5018 for Local SSL.

6.5 Discussion

The supervised diagnostic results show that the centralized and FedAvg models reproduced the script-generated outcomes almost perfectly, while Local own-client performance was slightly lower. These results indicate consistency with the deterministic labeling procedure rather than scene-level correctness.

The unsupervised analysis showed that TP-related structure was clearer than FP-related structure in the selected aggregated representation. The two-cluster structure therefore isolated a TP-dominant group more clearly than an FP-dominant group.

The semi-supervised results showed that the larger event pool and script-generated outcomes did not replace trusted video labels for correctness-oriented evaluation.

Federated learning provided a data-local collaboration setting. FedMeanTeacher achieved higher Macro-F1 and lower client-level variation than Local SSL on the script-generated outcome test set, although its FPR was higher. On the trusted video label test set, it showed no overall advantage over Centralized SSL.

For RQ3, script-generated outcomes supported large-scale script-oriented learning, while the scarcity of trusted video labels remained the main limitation for correctness-oriented TP/FP classification.

7

Conclusions and Future Work

This chapter summarizes the main findings, practical implications, limitations, and future work.

7.1 Main Findings

This thesis investigated AEBS full-braking event classification under cross-country data heterogeneity and limited trusted supervision.

For RQ1, the selected countries showed differences in event volume, vehicle-speed distributions, and JSD-based feature divergence, providing evidence of measurable cross-country heterogeneity. Country is used as a practical proxy for operating context, and the analysis does not identify the physical causes of these differences.

For RQ2, the script-generated outcomes showed substantially higher agreement with trusted TP labels than with trusted FP labels. Performance against script-generated outcomes should therefore be interpreted primarily as consistency with the deterministic labeling procedure. Correctness-oriented TP/FP evaluation requires trusted labels.

For RQ3, supervised models reproduced the script-generated outcomes almost perfectly. FedMeanTeacher provided more consistent client-level performance than isolated local SSL on the script-oriented evaluation, but no corresponding advantage was established on the limited trusted-label evaluation. The scarcity and imbalance of trusted labels, particularly trusted FP evidence, remain the main constraint on correctness-oriented learning.

7.2 Practical Implications

For AEBS event analysis, script-generated outcomes can support scalable diagnostics, weak supervision, and reproducibility checks, but should not replace trusted scene-level assessment.

Trusted annotation should be used strategically. Since FP-related and ambiguous cases showed the largest disagreement between the two label sources, annotation

effort is particularly valuable for these cases. A broader trusted subset would also enable more reliable evaluation of local and federated models across clients.

Federated learning provides data-local collaboration when raw event data cannot be directly pooled, but this benefit must be balanced against client-level performance.

7.3 Limitations and Future Work

The trusted subset is small and imbalanced, with limited country coverage. Future work should expand trusted annotation across operating environments, particularly for FP-related and ambiguous events, enabling more reliable client-level comparisons and uncertainty estimates.

The script-generated outcomes show structured disagreement with trusted labels, especially for FP-related cases. Future work could further characterize recurring disagreement scenarios, refine the deterministic labeling procedure, and investigate methods for structured label noise or confidence-weighted weak supervision. A larger trusted FP dataset could also support analysis of recurring FP subtypes and their signal patterns.

Country is used as a proxy for operating context and does not identify the underlying causes of heterogeneity. Future studies could examine alternative client definitions based on factors such as vehicle configuration, duty cycle, road environment, sensor setup, or data-collection conditions.

The federated experiments used simulated clients, standard aggregation, and a shared global model.Future work should evaluate practical deployment factors such as training time, communication conditions, and client availability.

The study is restricted to recorded AEBS full-braking events from Volvo heavy-duty vehicles. Future work could extend the analysis to other fleets, event definitions, and intervention stages, including pre-brake events.

7.4 Concluding Remarks

Reliable AEBS full-braking event classification requires both country-aware evaluation and trusted-label-aware learning. Cross-country analysis reveals measurable heterogeneity, while comparison with trusted labels shows that script-generated outcomes cannot support correctness-oriented conclusions on their own. A direct next step is to expand trusted annotation, especially for FP-related events, while continuing to use large-scale event data and federated learning as tools for scalable and data-local analysis.

Bibliography

[1] World Health Organization. (2023). *Global status report on road safety 2023*. World Health Organization.

[2] United Nations Economic Commission for Europe. (2023). Uniform provisions concerning the approval of motor vehicles with regard to the Advanced Emergency Braking Systems (AEBS), UN Regulation No. 131. *UNECE*.

[3] Volvo Trucks. (2017). Emergency Brake: A system that saves lives. *Volvo Trucks Magazine*.

[4] Vogel, K. (2003). A comparison of headway and time to collision as safety indicators. *Accident Analysis & Prevention, 35*(3), 427–433.

[5] National Highway Traffic Safety Administration. (2024). Federal Motor Vehicle Safety Standards; Automatic Emergency Braking Systems for Light Vehicles. *Federal Register*.

[6] Lubkowski, S. D., Demir, M., Duley, A. R., Cicchino, J. B., & McCartt, A. T. (2021). Driver trust in and training for advanced driver assistance systems. *Transportation Research Part F: Traffic Psychology and Behaviour, 80*, 540–556.

[7] Dozza, M., & González, N. P. (2013). Recognising safety critical events: Can automatic video processing improve naturalistic data analyses? *Accident Analysis & Prevention, 60*, 298–304.

[8] Guyonvarch, L., Lecuyer, E., et al. (2020). Evaluation of safety critical event triggers in the UDrive data. *Safety Science, 132*, 104937.

[9] Sun, T., Segu, M., Postels, J., Wang, Y., Van Gool, L., Schiele, B., Tombari, F., & Yu, F. (2022). SHIFT: A synthetic driving dataset for continuous multi-task domain adaptation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21371–21382.

[10] McMahan, H. B., Moore, E., et al. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 1273–1282.

[11] Kairouz, P., McMahan, H. B., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning, 14*(1–2), 1–210.

[12] Fawaz, H. I., Forestier, G., et al. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery, 33*, 917–963.

[13] Foumani, N. M., Miller, L., et al. (2024). Deep learning for time series classification and extrinsic regression: A current survey. *ACM Computing Surveys*.

[14] Ismail Fawaz, H., Lucas, B., et al. (2020). InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery, 34*, 1936–1962.

[15] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[16] Li, T., Sahu, A. K., et al. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems, 2*, 429–450.

[17] Barnard, Y., Utesch, F., et al. (2016). The study design of UDRIVE: The naturalistic driving study across Europe for cars, trucks and scooters. *European Transport Research Review, 8*(2), 1–10.

[18] Jain, A. K., Murty, M. N., et al. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

[19] Chapelle, O., Schölkopf, B., et al. (Eds.). (2006). *Semi-Supervised Learning*. MIT Press.

[20] Zhu, X., & Goldberg, A. B. (2009). *Introduction to Semi-Supervised Learning*. Morgan & Claypool Publishers.

[21] Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems, 30*, 1195–1204.

[22] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

[23] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[24] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[25] European Data Protection Board. (2021). Guidelines 01/2020 on processing personal data in the context of connected vehicles and mobility related appli-

cations. *European Data Protection Board*.

[26] Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National Science Review, 5*(1), 44–53.

[27] Frénay, B., & Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems, 25*(5), 845–869.

[28] Volvo Trucks. (2018). Collision Warning with Emergency Brake. In *25 ways Volvo FH keeps the roads safe*. Retrieved May 12, 2026, from `https://www.volvotrucks.com/en-en/news-stories/stories/2018/jul/25-volvo-fh-safety-features.html`

[29] Sohn, K., Berthelot, D., et al. (2020). FixMatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems, 33*, 596–608.

[30] Ratner, A., Bach, S. H., et al. (2017). Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment, 11*(3), 269–282.

A

Model and Training Configurations

This appendix reports the model architectures, training hyperparameters, decision-threshold settings, and federated-learning configurations used in the reported experiments.

A.1 CNN Architecture

The sequential CNN received an input tensor with 150 time steps and 21 input channels. The CNN contained three 1D convolutional blocks without intermediate pooling or temporal downsampling. Each block consisted of a convolutional layer, batch normalization, and a ReLU activation. After the final convolutional block, masked global average pooling aggregated the valid time steps before the fully connected classification head.

Table A.1: Architecture of the sequence-based 1D CNN.

Layer	Type	Output channels	Kernel	Stride	Other setting
1	Convolutional block	64	5	1	padding=2; BatchNorm1D; ReLU
2	Convolutional block	64	5	1	padding=2; BatchNorm1D; ReLU
3	Convolutional block	128	5	1	padding=2; BatchNorm1D; ReLU
4	Global pooling	–	–	–	Masked global average pooling
5	Fully connected	128	–	–	ReLU activation
6	Dropout	–	–	–	$p = 0.2$
7	Output	1	–	–	Binary logit

Table A.2: Summary of the CNN input and model size.

Property	Value
Input shape	150×21
Number of trainable parameters	85,569
Parameter data type	float32
Output convention	One TP logit; sigmoid applied for evaluation

A.2 Supervised Training Configuration

Table A.3: Supervised CNN training configuration.

Parameter	Value
Loss function	Binary cross-entropy with logits
Optimizer	AdamW
Learning rate	10^{-3}
Weight decay	10^{-4}
Batch size	256
Maximum epochs	50
Learning-rate scheduler	None
Class weighting	Positive-class weighting with `pos_weight`
Sampling strategy	Shuffled mini-batch sampling
Parameter initialization	PyTorch default initialization
Early stopping	None
Model state used for evaluation	Final epoch

A.3 Mean Teacher Configuration

Table A.4: Mean Teacher configuration.

Parameter	Value
Student architecture	CNN in Table A.1
Teacher architecture	Same as student
Teacher initialization	Copy of the initialized student model
EMA decay α	0.99
Trusted video label loss	Binary cross-entropy with logits
Script-generated outcome loss	Binary cross-entropy with logits
Consistency loss	Mean squared error between class probabilities
Selected script-loss weight γ_{script}	0.1
Candidate γ_{script} values	$\{0.0, 0.1\}$
Consistency-loss weight	1
Input-noise distribution	Gaussian noise
Input-noise magnitude	0.01 for weak augmentation; 0.02 for strong augmentation
Channel-dropout probability	0.1
Perturbation applied to	Teacher uses weak augmentation; student uses strong augmentation
Batch size	256 for script-generated outcome batches; 16 for trusted video label batches
Trusted video label samples per batch	16
Events without trusted video labels per batch	256
Optimizer	AdamW
Learning rate	10^{-3}
Weight decay	10^{-4}
Training epochs	30
Model used for test evaluation	Teacher

For clients without trusted video label training samples, $\mathcal{L}_{\text{trusted}}$ was set to zero. For isolated-local models without trusted video label validation samples, the final teacher state was used.

A.4 Federated-Learning Configuration

Table A.5: FedAvg and FedMeanTeacher configurations.

Parameter	Value
Number of clients K	8
Communication rounds R	10 for FedMeanTeacher; 30 for FedAvg
Local epochs per round E	3 for FedMeanTeacher; 1 for FedAvg
Client participation fraction	1.0
Clients sampled per round	8
Aggregation weighting	Number of local training samples for FedAvg and FedMeanTeacher
Local optimizer	AdamW
Local learning rate	10^{-3}
Local batch size	256, with 64 or 128 for smaller clients in FedMeanTeacher
Optimizer state across rounds	Reset
FedAvg aggregated parameters	Student model
FedMeanTeacher aggregated parameters	Student and teacher separately
Teacher initialization at each round	Received global teacher
Model state used for evaluation	Model state after the final communication round

A.5 Threshold Selection and Evaluation Model States

For every model, the decision threshold was selected using the validation set corresponding to the evaluation label. Script-generated outcome evaluation used script validation data, whereas trusted video label evaluation used trusted video label validation data. No test data were used for threshold selection. Trusted-label results for local models were reported only when the corresponding trusted validation and test samples were available.

No intermediate checkpoint selection or early stopping was applied. Centralized and isolated-local models used the model state after the final training epoch, whereas federated models used the model state after the final communication round.

Table A.6: Threshold-selection settings and model states used for evaluation.

Setting	Value
Search range	0.05 to 0.95
Search step	0.05 increments
Minimum AEBS-TP recall	0.6
Primary criterion	Lowest FPR
Secondary criterion	Highest macro-F_1
Tie-breaker	None
Fallback rule	Select the threshold with the highest recall
Centralized supervised CNN	Label-matched validation; final epoch; validation-selected threshold
Supervised FedAvg	Label-matched validation; final round; validation-selected threshold
Centralized Mean Teacher	Label-matched validation; final epoch; validation-selected threshold
FedMeanTeacher	Label-matched validation; final round; validation-selected threshold
Local supervised CNN	Label-matched validation where available; final epoch; validation-selected threshold
Local Mean Teacher	Label-matched validation where available; final epoch; validation-selected threshold

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY