INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL




MFFER: Multimodal Federated-Learning based Facial Emotion Recognition

JOÃO CRUZ DA SILVA
Setembro de 2025

P.PORTO

MFFER: Multimodal Federated-Learning based Facial Emotion Recognition

João Cruz da Silva

Student n.:1150425

Dissertation to obtain a Master's degree in Artificial Intelligence Engineering

Supervisor: **Maria Goreti Carvalho Marreiros, Full Professor, Institute of Engineering, Polytechnic of Porto**

Jury

President:

Ana Maria Neves Almeida Baptista Figueiredo, Associate Professor, Institute of Engineering, Polytechnic of Porto

Members:

Dalila Alves Durães, Assistant Professor, School of Engineering, University of Minho
Maria Goreti Carvalho Marreiros, Full Professor, Institute of Engineering, Polytechnic of Porto

Porto, September 2025

Resumo

O sector retalhista é um pilar das economias modernas, exigindo inovação contínua para se adaptar às tendências em sustentabilidade, digitalização e envolvimento do cliente. Neste contexto, diversos estudos exploraram a aplicação de técnicas computacionais avançadas para melhorar a experiência do cliente e a eficiência operacional.

A implementação de assistentes virtuais é uma das propostas para a inovação do sector, combinando-os com diferentes métodos de aprendizagem automática, como sistemas de recomendação, reconhecimento facial de emoções, entre outros. No entanto, o processamento de dados pessoais e subjetivos dos utilizadores pode levar a falhas de privacidade. Para enfrentar estas questões, a implementação de uma solução baseada em Federated Learning revela-se uma alternativa viável, permitindo usufruir dos benefícios das restantes técnicas enquanto se preserva a privacidade dos dados.

O objetivo desta dissertação é apresentar, implementar e testar um novo sistema automático de recomendação multimodal, suportado por métodos de reconhecimento facial de emoções, num ambiente de Aprendizagem Federada, capaz de lidar com problemas relacionados com a privacidade e armazenamento de dados, enquanto atualiza continuamente os modelos do sistema, aumentando a sua qualidade.

O desenvolvimento da dissertação demonstra a viabilidade do sistema proposto, apresentando resultados iniciais promissores em algumas das métricas avaliadas, uma arquitetura escalável e um desempenho global aceitável quando adaptado à estrutura federada proposta, em comparação com uma estrutura centralizada tradicional.

Palavras-chave: Aprendizagem Federada, Aprendizagem de Máquina, Reconhecimento Facial de Emoções, Privacidade de dados, Sistemas de Recomendação

Abstract

The retail sector is a cornerstone of modern economies, requiring continuous innovation to adapt to evolving trends in sustainability, digitalization, and customer engagement. In this context, numerous studies have explored the application of advanced computational techniques to enhance customer experience and operational efficiency.

The implementation of virtual assistants is one of the suggestions for the innovation of the sector, combining them with different machine learning methods, such as recommendation systems, facial emotion recognition, amongst others. However, the processing of personal and subjective data of the users leads to a data privacy breach. To address these issues, the implementation of a Federated Learning solution becomes a viable answer to the presented flaw, while being able to utilize the benefits of the other methods.

The objective of this dissertation is to present, implement and test a new automatic recommendation multimodal system with the support of facial emotion recognition methods, in Federated Learning environment, being able to tackle the problems related to privacy and storage of data, while constantly updating the models of the system, increasing its quality.

The development of the dissertation showcases the viability of the proposed system by displaying some early good results across some of the evaluated metrics, a scalable architecture, and an overall acceptable performance when adapted to proposed federated framework when compared to a traditional centralised structure.

Keywords: Federated Learning, Machine Learning, Facial Emotion Recognition, Data Privacy, Recommendation Systems

Acknowledgements

This research work was developed under the project CAPE (22017 CAPE), funded by the European Regional Development Fund (ERDF) within the project number NORTE2030-FEDER-01241200 - 17576, and funded by National Funds through the Portuguese FCT - Fundação para a Ciência e a Tecnologia under the R&D Units Project Scope, UIDB/00760/2020 (https://doi.org/10.54499/UIDB/00760/2020).

Table of contents

List of Figures

List of Tables

Acronyms and Symbols

List of Acronyms

AI	Artificial Intelligence
ANN	Artificial Neural Network
CK+	Extended Cohn-Kanade Dataset
CNN	Convolutional Neural Network
CREMA-D	Crowd-sourced Emotional Multimodal Actors Dataset
DNN	Deep Neural Network
DDQN	Double Deep Q-Network
EDA	Electrodermal Activity
EEG	Electroencephalography
EMG	Electromyography
FACS	Facial Action Coding System
FCFS	First-Come, First-Served
FCT	Fundação para a Ciência e Tecnologia
FER	Facial Emotion Recognition
FL	Federated Learning
GA	Genetic Algorithm
GDPR	General Data Protection Regulation
GECAD	Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development
GSR	Galvanic Skin Response
HFL	Hierarchical Federated Learning
HMM	Hidden Markov Model
HRI	Human-Robot Interaction

I/O	Input/Output
IoT	Internet of Things
IFedAvg	Federated Averaging Interoperable Element-wise Affine Layers
KNN	K-Nearest Neighbours
LPC	Linear Predictive Coding
LSTM	Long Short-Term Memory
M2M	Machine-to-Machine
MER	Multimodal Emotion Recognition
MFCC	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
Non-IID	Independent and identically distributed
PAD	Pleasure-Arousal-Dominance
RAF-DB	Real-world Affective Face Database
RGB	Red, Green, Blue
RNN	Recurrent Neural Network
SVM	Support Vector Machine
VFL	Vertical Federated Learning
TFL	Transfer Federated Learning
SGD	Stochastic Gradient Descent
SSL	Self-Supervised Learning

List of Symbols

1 Introduction

To begin this dissertation, a small contextualization of its context will be provided such as its insertion as well as some of the scientific contributions it provided in the project it was inserted in.

1.1 Contextualization

The retail sector is one of the most crucial sectors in any country's economy, therefore becoming a subject of multiple studies such as the (*Retail and Wholesale Transformation May Require up to 600 Billion in Investments to Future-Proof the Sector - EuroCommerce*, no date) in which it was concluded that the European Union's retail and wholesale sectors need to undergo transformation for sustainability, digitization as well as skills and talent.

With this, different efforts are being made to see how this can better be achieved by exploring new technological approaches, including the incorporation of AI to increase both the interaction with customers, as well as the optimization of store management, like sales, stocks and distribution logistics.

As such, some of the techniques being explored are the implementation of recommendation systems and emotion analysis of customers' opinions of the products, including the ones suggested by AI models, through a multimodal detection channel, such as the combination, of audio, video, and text, it's possible to create a profile of the customer's preferences to increase the efficiency of the recommendation system. The user engagement with the application may drive it to a higher degree of trust towards the system efficacy, however due to the sensibility of the subject, this being, the user's own preferences, and personal data which is captured during the usage of the system, making it necessary for the system to be in compliance with the GDPR (*EUR-Lex - 32016R0679 - EN - EUR-Lex*, no date), to also keep the user's trust in the safety of the system.

One such alternative that has garnered considerable attention is Federated Learning. This emerging technique diverges from the conventional centralized model by emphasizing edge computing and harnessing data islands for training. Federated Learning allows AI models to be trained collaboratively across a network of decentralized devices while preserving data privacy and security. This approach aligns with the principles of data minimization and local processing, which are essential components of GDPR compliance.

The primary objective of this dissertation is to investigate the feasibility and efficacy of implementing a multimodal Facial Emotion Recognition system, which can support recommendation systems by evaluating the user's feelings and opinions towards a recommended product, while using Federated Learning to protect the data necessary for the system's performance.

This research endeavours to bridge the gap between data privacy concerns and the need for alternative AI training methodologies. Through empirical studies and data-driven analyses, this dissertation aims to contribute valuable insights into the practical application of Federated Learning within the affective computing field, in order to create a new tool for retail, which can tackle the different challenges that are currently being faced in a technological level.

The development of this dissertation was made as part of the CAPE project on ITEA, which is being developed within an international consortium aiming to build a scalable platform. The work developed here is primarily focused on the Portuguese use case, that has as a goal, the development of a platform to assist sales in a store for materials for house decoration by having kiosks with a recommendation system and chatbot that are powered by emotion recognition model to provide more sensible recommendations and answers while interacting with users to promote engagement and trust.

1.2 Research Questions and Objectives

This dissertation was written within the scope of the CAPE project; however scientific considerations were taken on how this project can bring not only economic benefits but also new scientific contributions. The main questions that arose during the development of the project were the following:

- RQ1: The capability of the system to respond based on the user's emotions
    - RQ1.1: Would the subsequent NLP model be able to generate answer that could match the detected feelings?
    - RQ1.2: Would the Recommendation model be capable of adequately answering the recognized emotions (keep the same trend, adjust recommendations properly, etc)
- RQ2: How can data be protected and kept in compliance with GDPR and other protocols

    - RQ2.1: During the interaction with the system (Such as capturing the user's face)
    - RQ2.2: For model training and analytics (For both Recommendation and Facial Emotion Recognition models)

A list of objectives has been formulated according to the list of research questions that are to be explored and to layout the route to be made throughout the development of this dissertation. The objectives are the following:

- O1 – Exploring the state of the art of federated learning and emotion recognition
- O2 – Exploring the integration of federated learning with emotion recognition
- O3 – Exploring the open-source federated learning frameworks that are available and their utilization
- O4 – Conceptualization and implementation of a solution that integrates federated learning with emotion recognition, and its validation.

1.3 Scientific Contributions

The prototype being developed within the scope of this dissertation is a pivotal component of the CAPE project, which has received funding from the Portuguese Foundation of Science and Technology.

During the development of the associated project, a conference paper under the title "Enhancing Personalized Recommendations with Federated Learning and Multimodel Emotion Recognition" was submitted, accepted, and presented in the 2025 Sixth International Conference on Intelligent Data Science Technologies and Applications (IDSTA2025) that happened in Varna, Bulgaria between 1st to 4th of September.

2 State of the Art

2.1 Federated Learning

Federated Learning (FL), first introduced in 2016 (Brendan McMahan *et al.*, 2016) represents a paradigm shift in machine learning (ML) methodology by enabling the decentralized and collaborative training of models across a large number of edge devices. These devices, often referred to as "clients" in the FL ecosystem, store and process their own local data. The data remains on the device, obviating the need for data to be transferred to a central server for model training. This decentralized approach addresses critical concerns surrounding data privacy, security, and ownership, particularly in domains where sensitive or personal information is involved, such as healthcare, finance, and IoT.

The primary motivation behind Federated Learning is to mitigate the privacy risks associated with traditional centralized machine learning approaches, where raw data is often transferred to central servers for processing. In centralized ML, data sharing introduces significant privacy and security vulnerabilities, as sensitive information can be exposed to malicious entities or unauthorized users. FL provides a framework for overcoming these limitations by ensuring that data remains on the edge device, thereby protecting the user's privacy while still allowing for collaborative model improvement.

In this decentralized setup, each device is responsible for processing its local data and training its model on the data it possesses, which is inherently more secure than moving raw data to a central server. Moreover, FL can be integrated with differential privacy techniques to further enhance security by ensuring that the model updates sent to the central aggregator are sanitized to prevent any leakage of sensitive information. This allows organizations to maintain high standards of data privacy while still benefiting from collective knowledge embedded in the distributed data.

2.1.1 Federated Learning Process

The process of Federated Learning involves several key stages, which are repeated iteratively, leading to the continuous refinement of the global model. These stages facilitate the collaborative learning process, where the model evolves based on data from diverse edge devices, ultimately enabling the global model to generalize across various environments.

1. **Initialization**: The FL process begins with the initialization of a global model. Typically, this model is initialized with random parameters or pre-trained weights from an existing model. Once initialized, the global model is sent to all participating edge devices, where each device holds its own data, and is ready to perform local updates based on its unique data distribution. The initialization phase is crucial as it sets the

foundation for the entire training process, providing the baseline from which model updates will occur.

2. **Local Training**: Once the global model is downloaded, each participating device begins the process of local training using its own local data. Local training involves iteratively adjusting the model's parameters through an optimization technique such as stochastic gradient descent (SGD), typically using the loss function associated with the model's objective. Importantly, during this phase, the local data remains on the device and is not shared with the central server or other participants, ensuring privacy and compliance with data protection regulations. Local training can take place in multiple iterations, where the model's weights are gradually refined to fit the local data distribution. The advantage of local training is that it allows the model to benefit from the data heterogeneity present across edge devices. This diversity in the data enables the model to learn from a wide range of scenarios, improving its generalization capabilities without needing access to centralized data. Additionally, local training can also be tailored to the specific needs or constraints of each device, such as energy efficiency or memory limitations.

3. **Model Update**: After completing the local training phase, each device computes the gradients of the loss function with respect to the model parameters. These gradients represent the updates made to the model as a result of the local training. The gradient computation typically involves evaluating the difference between the model's predictions and the actual data points in the local dataset, followed by an optimization step to adjust the model's parameters. These computed gradients are not the model's raw data, but rather abstract representations that capture the learning performed on the local data. The key advantage of this step is that the devices share only these gradient updates (not raw data) with the central server. This ensures that sensitive information never leaves the device, making Federated Learning an effective mechanism for maintaining data privacy. Moreover, these updates are often compressed and aggregated before transmission to reduce the risk of exposing sensitive information through inference attacks.

4. **Aggregation**: Once the local updates have been computed, they are sent to a central aggregator (also called a federated server), which is responsible for combining these updates into a single global model. The aggregation process typically involves a weighted averaging of the model parameters or gradients from each device, where the weights correspond to the amount of data or the training quality on each device. Common aggregation methods include FedAvg (Federated Averaging), where each device’s update is weighted by the number of data points it has. The aggregation step is crucial for ensuring that the global model accurately reflects the collective learning of all participating devices. The server does not access any individual device’s data but instead aggregates the gradients in a way that accounts for the varying sizes and characteristics of the local datasets.

5. **Global Model Update**: Once the local updates are aggregated, the central server generates an updated global model by applying the combined gradients to the global model parameters. This new global model is then distributed back to all edge devices. Each device replaces its current local model with the updated global model, which now incorporates the learning from all other devices. This process of distributing the new global model ensures that all devices benefit from the collective knowledge gained across the federation, without the need for raw data sharing. Over time, as more rounds of local training and aggregation occur, the model's performance improves, and it becomes increasingly robust to the diversity of data encountered across devices.

6. **Iterative Process**: The steps outlined above (local training, gradient computation, aggregation, and model update) are repeated over several rounds, allowing the global model to gradually refine and improve based on the insights gained from the distributed data sources. This iterative process continues until a specified convergence criteria is met—either when the model performance reaches an acceptable level or when the model's parameter updates become sufficiently small, indicating that further training would not yield significant improvements. The number of rounds required for convergence can vary depending on factors such as the complexity of the model, the heterogeneity of the data, the computational resources available on the devices, and the size of the training dataset. Convergence can also be influenced by the choice of optimization algorithm and hyperparameters, such as the learning rate or the batch size.

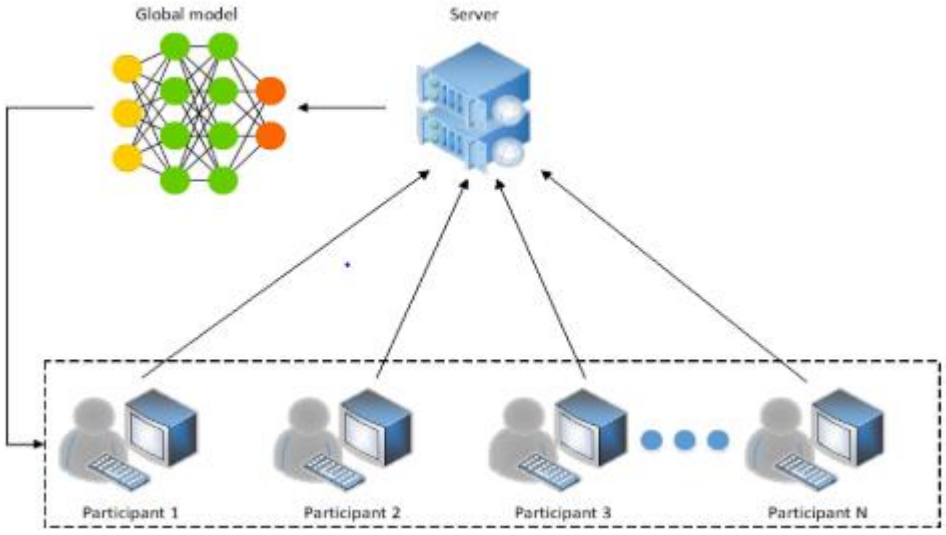

Figure 1 - Generic Framework of Federated Learning. (Wen *et al.*, 2022)

2.1.2 Types and Variants of Learning Process

The usage of Federated Learning can be implemented using different architectures depending on the communication and coordination paradigm among participating nodes. In a centralized approach, a server orchestrates the federation process by selecting a sample of clients to perform local model training on their private datasets. The clients then send their locally trained model updates to the server, which aggregates these updates—commonly using Federated Averaging (McMahan *et al.*, 2017)—to generate a global model. This global model is then redistributed to the clients for further local updates, iteratively improving the shared model without requiring direct access to raw data.

Alternatively, a decentralized approach removes the reliance on a central server. Communication occurs in a peer-to-peer manner, where clients exchange model updates directly with one another or in small clusters. Decentralized FL can enhance fault tolerance and scalability but introduces challenges in ensuring convergence and consistency of the global model (Yang *et al.*, 2019).

A heterogeneous or hybrid approach has also been proposed, where local models may differ in architecture or capacity, and a heterogeneous global model accommodates these variations. Such designs aim to improve performance in environments where clients have different computational resources or feature spaces (Li et al., 2020).

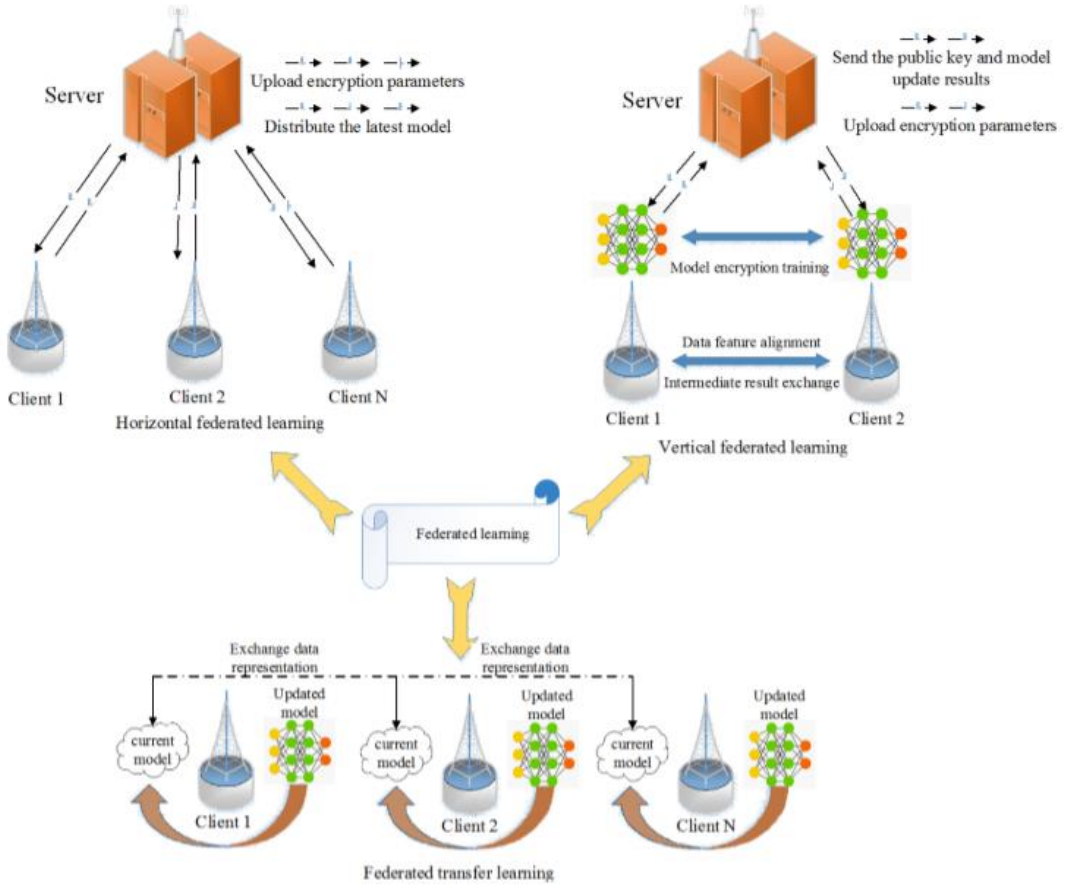

Figure 2 - Categorization of Federated Learning. (Wen *et al.*, 2022)

FL can further be categorized based on the distribution of data across clients:

- **Horizontal FL (sample-partitioned):** Clients share a similar feature space but have different data samples. This is typical in scenarios like mobile keyboard prediction or sensor networks, where the same features are collected across multiple users, but each user contributes unique records. Horizontal FL leverages overlapping features for collaborative training while preserving data privacy across disjoint datasets (Yang *et al.*, 2019).
- **Vertical FL (feature-partitioned):** Clients share some overlapping samples (i.e., common IDs), but the feature spaces differ. Vertical FL is common in financial and healthcare applications where multiple institutions have complementary features for the same set of users or patients. Due to the heterogeneous feature spaces, vertical FL often relies on simpler models or specialized protocols for secure gradient computation, leaving room for methodological improvements in handling more complex models (Yang *et al.*, 2019).
- **Federated Transfer Learning (sample and feature-partitioned):** In scenarios where neither the sample space nor the feature space overlaps, transfer learning techniques are integrated with FL to enable knowledge transfer across domains. Here, patterns learned in one domain can be applied to a different, non-overlapping domain, extending FL to highly heterogeneous datasets (Pan *et al.*, 2010), where knowledge acquired in on domain is applied to a different domain, is used with FL (Liu *et al.*, 2018). This approach is particularly useful for cross-institutional collaborations in healthcare or multi-enterprise settings, where data is highly distributed and heterogeneous.

These categorizations provide a structured understanding of FL paradigms, clarifying the design choices based on data characteristics, privacy requirements, and computational constraints. Recent research continues to explore adaptive aggregation methods, model personalization strategies, and privacy-preserving enhancements to expand the applicability of FL in complex real-world environments.

2.1.3 Challenges of Federated Learning

- **Data heterogeneity and non-IID data:** While Federated Learning offers numerous benefits, including improved data privacy and the ability to learn from distributed datasets, there are several challenges that need to be addressed to fully realize its potential. These include issues related to data heterogeneity, communication efficiency, and the handling of non-IID data. Furthermore, the security of the aggregation process is critical, as adversarial attacks on the federated server could potentially compromise the integrity of the global model. Techniques such as secure multi-party computation (SMPC) and homomorphic encryption are actively being explored to mitigate these risks and ensure robust model aggregation.

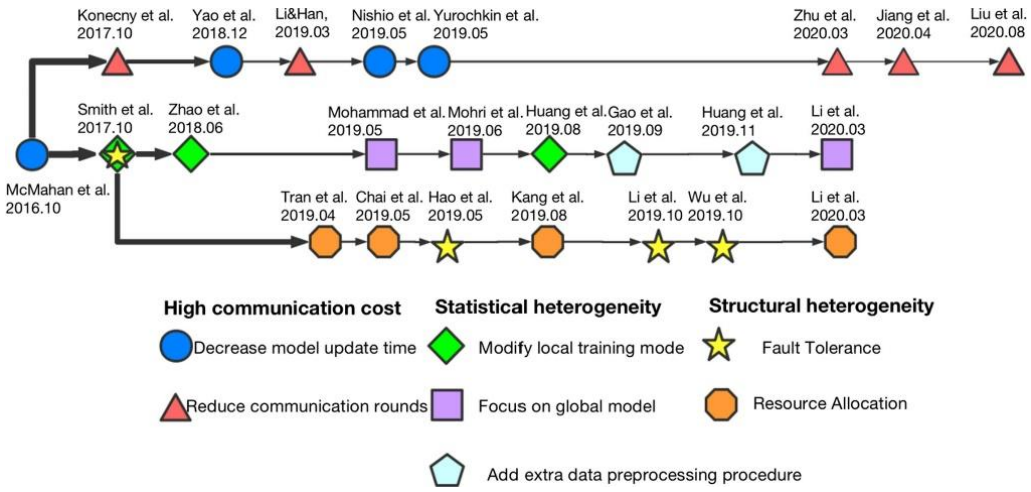

Figure 3 - Optimization Path to solve open problems (Li *et al.*, 2020)

- **Regular data curations:** One of the current limitations of FL is the regular curations required in each node's dataset. This requirement poses challenges in terms of maintaining data quality and ensuring the accuracy of the global model.(Liu *et al.*, 2021) have addressed this issue by proposing a novel protocol aimed at reducing the frequency of necessary curations. Their protocol not only focuses on improving the accuracy of the curator but also enhances privacy preservation without the reliance on trusted third parties. Furthermore, the protocol emphasizes privacy preservation, a fundamental concern in FL. By leveraging cryptographic techniques and secure aggregation methods, the authors ensure that sensitive data remains protected during the curation process. This approach minimizes the risk of data leakage or privacy breaches, aligning FL with strict privacy requirements.

- **Practical scalability of FL**: Importantly, their proposal offers a promising avenue for advancing FL's practicality and scalability in real-world applications. By reducing the overhead associated with regular curations, improving curator accuracy, and enhancing privacy, their protocol addresses critical limitations in current FL systems, paving the way for more efficient and secure decentralized machine learning.

- **Bias during global updates:** There's also the problem of bias in FL during the global update as a critical concern that stems from the random selection of nodes for training within most FL frameworks. Since these frameworks are designed to be agnostic to the data distribution of each client, there is a risk of introducing bias into the global model. This bias can significantly impact the overall performance and fairness of the federated learning system. Addressing this issue, (Cho, Wang and Joshi, 2022) have introduced a novel framework that offers a convergence analysis of biased clients. Their framework exhibits enhanced efficiency in terms of convergence speed

when compared to traditional FL methods. To delve into the scientific aspects of this proposal, it is imperative to understand the mechanisms and techniques employed to mitigate bias and expedite convergence in FL.

- **Fairness-aware aggregation:** (Ezzeldin *et al.*, 2021) introduces an algorithm tailored for the purpose of fairness-aware aggregation within the context of federated learning. Federated learning is a decentralized machine learning paradigm that enables the training of models on distributed data sources, typically located on client devices or edge servers, while preserving data privacy and security. In federated learning, the primary goal is to collaboratively train a global model by aggregating locally trained models from multiple clients. One of the key challenges in federated learning is ensuring fairness across different groups of clients or data contributors, particularly when there are imbalances or biases present in the data. The concept of group fairness refers to the equitable treatment of different groups or subsets of data, such as demographic groups, in the learning process. This is crucial to prevent the propagation of biases and to ensure that the resulting model is fair and unbiased when applied to real-world scenarios.

Communication bottlenecks and node failures: One notable challenge is the bottleneck created by communication among the participating nodes, which can lead to node failures and device dropouts, thereby hindering the overall effectiveness of FL systems. Addressing this limitation, (Lim *et al.*, 2022) introduced the Hierarchical Federated Learning (HFL) framework as a potential solution. In the HFL framework, cluster heads are designated to support data owners by facilitating intermediate model aggregation within their respective clusters. This hierarchical and decentralized approach aims to alleviate the communication bottleneck and reduce the reliance on a central controller, thereby enhancing the robustness of FL systems. HFL offers promising advantages, it also introduces its own set of challenges. Resource allocation, for instance, becomes a critical consideration in this framework.

Efficiently distributing computational resources among cluster heads and data owners to ensure fair and effective model training remains an area of active research.

Furthermore, incentive design is another key aspect that requires careful attention within the HFL framework. Incentives play a pivotal role in motivating data owners to actively participate in FL and contribute their data for model training. Designing effective incentive mechanisms that align the interests of all parties involved while preserving data privacy is a complex problem that demands further investigation.

2.1.4 Current Advances and Future Directions

- **Surveys and evolution of FL:** Multiple surveys have been done about Federated Learning through the years since its introduction to explain not only its inline workings and advantages, but also to describe the current situation of state of the art as well as the challenges in the open to further develop the framework. (Wen *et al.*, 2022) wrote one of these surveys in which they go over the general overview of the workings of Federated Learning, their categorization, concerns about privacy and security challenges, going over the main techniques being used to ensure data communication privacy as well as the most common types of attacks and methods to counter them.
  In a different review proposed by (Li *et al.*, 2020), they mention the evolution of FL since it was first proposed. In this evolution chart, it mentions the progresses made in various topics such as minimization of high communication costs and the statistical and structural heterogeneity of data. Another of their summaries is related to security improvements made. In this section a summary of the progresses made regarding privacy risk minimization, preservation at both server and client side and creating more secure frameworks.
  They also delve into a comprehensive exploration of FL and its multifaceted applications across various domains, including but not limited to mobile devices, industrial engineering, and healthcare. This emergent paradigm of machine learning has garnered substantial attention due to its potential to facilitate collaborative model training while preserving data privacy and security.
  In their final chapter, they present an extensive roadmap for advancing Federated Learning, encompassing asynchronous training, gradient aggregation, incentive mechanisms, model verification, blockchain integration, and the extension of FL to unsupervised learning, thereby underscoring the rich and evolving landscape of this transformative machine learning paradigm.

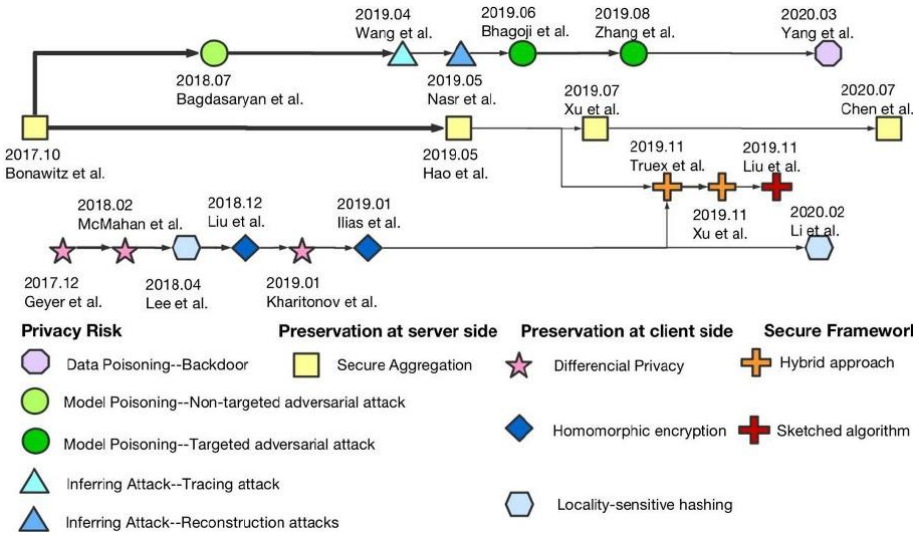

Figure 4 - Evolution of privacy in FL (Li *et al.*, 2020)

- **Transformers for heterogeneity:** (Qu *et al.*, 2022) have raised valid concerns regarding the data heterogeneity observed across different islands, and their proposal to address this issue through the utilization of self-attention-based architectures, particularly Transformers, is a notable approach in the realm of machine learning and deep learning. One key advantage of using self-attention-based architecture like Transformers is their ability to accelerate convergence during training. This is crucial, especially when dealing with heterogeneous data, as quicker convergence can reduce the risk of overfitting to specific subpopulations within the dataset. Additionally, Transformers have been shown to exhibit better generalization performance when faced with distribution shifts, as their inherent flexibility enables them to capture and adapt to intricate patterns present in diverse data sources. Furthermore, the notion of achieving a "better global model" aligns with the goal of developing machine learning models that perform well across all islands, irrespective of the variations in data characteristics. By leveraging self-attention mechanisms, researchers aim to enhance the model's capacity to generalize effectively, thereby reducing the impact of data heterogeneity on its performance.

- **FedAlign:** (Mendieta *et al.*, 2022) introduced a novel method, referred to as FedAlign, which represents a promising advancement in the realm of FL. In the contemporary landscape of machine learning, FL has garnered substantial attention due to its potential to enable collaborative model training across decentralized devices or servers while preserving data privacy. The primary objective of FedAlign, is to not only achieve competitive performance comparable to state-of-the-art methods but also to mitigate the computational demands and memory overhead typically associated with such endeavours. Federated learning entails training machine learning models on data distributed across multiple local devices or servers, thereby bypassing the need for centralized data aggregation. This decentralized approach is particularly pertinent in scenarios where data privacy and security concerns are paramount, such as in healthcare, finance, and edge computing. However, federated learning often confronts challenges in terms of communication efficiency, model convergence, and scalability. FedAlign addresses these challenges through a set of innovative techniques. The method's core proposition is the efficient synchronization of model updates across participating devices or servers, which enables faster convergence. By meticulously optimizing communication protocols and aggregating model updates in a more resource-efficient manner, FedAlign reduces the demand for extensive computational power and memory resources, making it a pragmatic choice for resource-constrained environments. Moreover, FedAlign incorporates adaptive strategies for model selection and parameter tuning, tailoring its behaviour to the specific characteristics of the distributed data. This adaptability enhances its ability to compete with state-of-the-art methods while mitigating the risk of overfitting or underfitting on local data distributions. The interoperability between nodes in a distributed computing system is a critical concern that continues to be the subject of

ongoing research and development efforts, as it represents a fundamental prerequisite for the seamless and efficient execution of such systems. One promising solution that has emerged to address this challenge is the Federated Averaging with Interoperable Element-wise Affine Layers (IFedAvg), as proposed by (Roschewitz *et al.*, 2021). IFedAvg is an innovative approach designed to enhance the coordination and communication between nodes in federated learning settings, facilitating the convergence of global model parameters while accommodating local variations.

- **IFedAvg:** IFedAvg builds upon the conventional Federated Averaging (FedAvg) algorithm, which is a widely adopted federated learning technique. FedAvg involves aggregating local model updates from individual nodes and computing a global update to refine the model. However, the standard FedAvg approach may encounter difficulties when dealing with nodes that exhibit significant heterogeneity in their data distributions or model architectures. In contrast, IFedAvg introduces the concept of local element-wise affine layers, which serve to address the challenges of interoperability and data heterogeneity among participating nodes. These affine layers are applied to the local model updates generated by individual nodes before they are incorporated into the global model. By introducing this additional layer of adaptability, IFedAvg allows nodes to adapt their updates to the global model in a more context-aware manner. The incorporation of local element-wise affine layers in IFedAvg effectively provides a mechanism for nodes to fine-tune their contributions to the global model, aligning them with the overarching objectives of the federated learning system. This adaptability enables nodes to account for variations in their local data distributions and model architectures, ultimately leading to more efficient model convergence and improved overall system performance.

- **Curator accuracy:** The need for regular dataset curations arises from the decentralized nature of FL, where individual nodes contribute local updates to a global model. These local updates may contain noisy or biased data, leading to model performance degradation if not carefully curated. Traditional approaches to data curation often involve centralizing data, which may not align with the privacy-preserving goals of FL. This protocol introduces innovations to mitigate these challenges. One key aspect of their approach is the enhancement of the curator's accuracy. By improving the quality of data selection and aggregation, the global model can benefit from cleaner and more informative updates, reducing the need for frequent curations. This improvement in curator accuracy contributes to the overall robustness and efficiency of the FL system.

- **Feature Selection:** (Feng, 2022) presents an innovative feature selection mechanism designed specifically for vertical Federated Learning setups. Vertical Federated Learning involves scenarios where data sources possess different sets of features, and the goal is to collaboratively train a model across these distributed datasets. However, the inherent heterogeneity in features across data sources can pose challenges,

including communication overhead and model convergence issues. His proposed mechanism aims to address these challenges by optimizing feature selection within the FL framework. The proposed feature selection mechanism operates at the intersection of data privacy, communication efficiency, and model performance. It leverages techniques such as differential privacy to ensure that sensitive information is not exposed during the feature selection process. By enabling each local node to autonomously decide on feature inclusion or exclusion, the mechanism contributes to improved data privacy preservation within the FL ecosystem. Furthermore, the feature selection mechanism enhances the performance of local nodes. By allowing nodes to select the most relevant features for their local dataset, it can lead to more efficient and accurate model training. This is particularly beneficial when dealing with data sources that have varying degrees of data quality and relevance. In addition to the performance gains at the local level, his approach also contributes to better curation of the datasets used in federated learning. By focusing on selecting features that are informative and relevant to the specific task, it reduces noise and redundancy in the aggregated dataset, ultimately leading to improved model generalization and predictive accuracy.

- **Backdoor mitigation:** Furthermore, the mitigation of backdoor attacks in FL has emerged as a crucial research area due to its inherent vulnerability. Backdoors, when surreptitiously injected into the global model by malicious clients, pose a significant threat to the integrity of the FL process, potentially compromising the collective model's performance and confidentiality. In response to this threat, researchers have introduced various methodologies aimed at enhancing the security and robustness of FL systems. One notable contribution in this context is the framework proposed by (Andreina *et al.*, 2021). Their approach is founded on the concept of leveraging feedback mechanisms within the FL system, harnessing data emanating from individual client nodes to detect and mitigate the presence of poisoned or backdoored models. The underlying premise of this framework revolves around the idea that the data generated by benign clients can serve as a valuable resource for identifying anomalous behavior and isolating nodes with malicious intentions. The effectiveness of their framework is demonstrated through rigorous experimentation, highlighting its superior performance in countering state-of-the-art backdoor attacks. Notably, the reported accuracy of 100% suggests that the proposed approach successfully identifies and neutralizes backdoor-infected models with a high degree of certainty. Moreover, the low false positive rate, documented as being below 5%, underscores the framework's ability to minimize the likelihood of misclassifying benign models as malicious, thus ensuring the reliability of the detection process. This innovative framework not only showcases promising results in bolstering FL security but also represents a significant step forward in the ongoing efforts to fortify FL against adversarial threats. By harnessing client-generated data and leveraging sophisticated feedback mechanisms, this approach holds the potential to enhance the resilience of FL systems and contribute to their widespread adoption in sensitive and privacy-sensitive applications. However, as the field of FL security continues to evolve, further

research and refinement of such frameworks will be essential to stay ahead of increasingly sophisticated adversarial tactics and ensure the long-term viability of FL as a robust and secure machine learning paradigm.

- (Kumari *et al.*, 2022) presents a robust defence framework designed to enhance the security of distributed systems by effectively detecting and mitigating the presence of malicious updates. The framework capitalizes on the concept of probability distributions over client updates to scrutinize and identify any potentially harmful alterations within the updates. This innovative approach involves the calculation of a probabilistic measure that evaluates the characteristics of client updates, thus enabling the monitoring of modifications made to the data. Central to this defence framework is the development of a novel detection algorithm, which leverages the computed probabilistic measure to efficiently discern and subsequently filter out any malicious updates. The algorithm employs sophisticated statistical techniques to compare the observed characteristics of incoming client updates against the expected properties of legitimate updates. By establishing a probabilistic model that captures the inherent variability in the update patterns, the algorithm can make informed decisions regarding the trustworthiness of each incoming update. Furthermore, this framework operates within the realm of distributed systems, making it particularly relevant in contexts where multiple clients contribute updates to a shared system or database. The utilization of probability distributions and probabilistic measures adds a layer of sophistication to the security apparatus, enabling the early identification and isolation of malicious updates, thus safeguarding the integrity and reliability of the distributed system. The framework's emphasis on probabilistic analysis and its novel detection algorithm offers promising prospects for improving the resilience of distributed systems against malicious threats, ultimately contributing to the overall security and trustworthiness of such systems in various applications, including but not limited to cloud computing, decentralized networks, and collaborative data sharing environments.

- (Nguyen *et al.*, 2021): One fundamental aspect of this framework involves the development of specialized client selection strategies. These strategies aim to intelligently select clients for training during the global update, considering not only their model performance but also the potential bias they may introduce. This selection process incorporates statistical techniques and machine learning algorithms to assess the bias that each client's data may exhibit with respect to the global model. The convergence analysis performed in this framework is rooted in mathematical and statistical foundations, formulating optimization objectives that balance model performance and fairness. Advanced optimization algorithms are integrated to adapt to biased clients' contributions, accelerating convergence while preserving fairness across the federated learning system.
  Another key component of the framework is the incorporation of privacy-preserving mechanisms. Given the distributed and privacy-sensitive nature of federated learning,

safeguarding client data remains paramount. The proposed framework employs state-of-the-art cryptographic and differential privacy techniques to ensure that sensitive information remains protected while enabling efficient global updates. Experimental evaluations demonstrate the efficacy of the framework, showing improvements in convergence speed, model performance, and bias reduction across real-world datasets.

Finally, the framework introduces a server-side fairness-aware aggregation algorithm, designed to operate independently of the local debiasing methods used by individual clients. Local debiasing methods can vary widely, from preprocessing techniques to model modifications, but this algorithm allows clients to choose the approach best suited for their data while still benefiting from equitable aggregation at the server. By combining client selection, convergence optimization, privacy preservation, and fairness-aware aggregation, this work provides a comprehensive solution to mitigate bias and enhance fairness in federated learning scenarios, particularly in environments with heterogeneous and distributed datasets.

2.2 Affective Computing and Emotion Recognition

Human intelligence is compromised not only by rational and logical thought processes, but also by emotional processes, and as such, to translate this part of human rationality, affective computing was introduced to expand the capabilities of machine learning models. (Shoumy *et al.*, 2020; Arya, Singh and Kumar, 2021)

Affective computing is a multidisciplinary area that involves computer science, psychology, and cognitive science with the goal of creating a more human-like machine behavior for Human-Computer Interactions.

This kind of technique allows for the exploration of areas such as the understanding of sentiments, emotions, and opinions from different sources of data, through different physiological signals, extending from facial reactions to heart rate signals, performing sentiment analysis, which focuses on a binary classification (positive vs negative) or 3-way classification (positive, neutral, negative) of the perception of sentiments, and the emotional recognition, which is a more detailed analysis, identifying detailed views on the perceived topic.

2.2.1 Emotion Models and Theories

The analysis of text thought Natural Language Processing was the main method to apply this technique, but with the ever increase of availability of data on the Web, and the development of hardware, incorporating video and audio, creating multimodal channels to increase the effectiveness of models. Affective computing was developed by integrating emotion models

into the computational fields, creating a combination of psychology and computer science. To achieve this, different Emotion models and theories were analyzed and integrated into computational system, which can be split into two different categories, named discrete theories and continuous emotional states.

The core emotional discrete models are the following:

- **Tomkin's Affective imagery consciousness** (Tomkins, 2008): Tomkins' framework emphasizes the centrality of affect as a primary motivational system in human behaviour. In contrast to models that treat emotions as secondary to biological drives, Tomkins proposed that innate affects—such as joy, fear, anger, or shame—directly shape perception, thought, and action. His theory highlights how emotions are expressed through imagery and consciousness, forming the foundation of human personality and social interaction.

- **Ekman's basic emotions** (Ekman, 2004): One of the most widely used models for the representation of discrete emotions is the Ekman's model that covers what he affirms to the be, the core human emotions, these being happiness, sadness, anger, fear, surprise and disgust. In his earlier studies, contempt was also one of the core emotions, but with the development of his studies, he focused on the first 6.
  Some models that derive from Ekman's work include neutrality within these core emotions.

In contrast to discrete emotions, a continuous emotional state represents a dimensional vector where emotions change as a reaction to different stimuli such as arousal, being an active reaction based on these stimuli, or valence, where one finds the stimuli pleasurable and/or interesting compared to more displeasing sensations. The foundational frameworks of these models were:

- **Russell's Circumplex Model** (Russell, 1980): where he organizes emotions in a multidimensional space based on the said valence and arousal being able to describe more complex emotions or mental states in a circumplex, managing to map multiple or mixed emotions simultaneously. This presented a way to represent both structured affective experiences and a representation of cognitive structures.
- **Pleasure-Arousal-Dominance (PAD) model** (Mehrabian, 1996): just like Russell's Model, represents emotions in a multidimensional view, however, introduces a new value in this scale, namely dominance that an individual as over a situation and how these stimuli can better influence emotions than just using arousal and valence, and how temperament and personality trains can further influence some of the existing emotional states

Some applications have been explored by integrating these models into affective computing to create systems capable of better identifying emotions either in a discrete or continuous representation. Some works based on continuous models such as (Di Tecco, Foglia and Prete, 2024) where they explore the use of key signals to attempt to identify the continuous states of emotions where they achieved higher accuracy results compared to discrete models.

(Sharma *et al.*, 2022) propose a model that can jointly predict discrete and continuous emotional attributes by using multi-task learning, displaying perceivable improvements in performance when comparing to single label models. Another proposal that studies the comparison of the distinct models (Kansizoglou *et al.*, 2022) where they estimate arousal and valence when tracking individual emotional variations. During their experiments they managed to observe that continuous models outperformed the discrete models.

Just like the example works above, there are more proposals that support the higher performance of continuous models over discrete, however there are data limitations when using these that will be discussed better in the challenges section.

Another method to analyse emotional models is through categories, named categorical models or dimensions, named dimensional models, can be linked to the previously categorized discrete and continuous emotions, however instead of focusing on perception and the evolution of emotions, these methods to categorize models instead focus on computational quantifications. In this representation, discrete models directly translate to categorical models and continuous models to dimensional models.

The most widely used categorial models are, once again, the Ekman model, given the way of how discrete emotions can be easily turned into categories, the model doubles down in how it can be represented, and the other model is Plutchik's wheel of emotions (PLUTCHIK, 1980), where the wheel attempts to represent a visual representation of primary and secondary emotions. Dimensional emotions can be directly traced back to the continuous models using the same Russell's Circumplex model and the PAD model as the core models. This method of categorization highlights the bridge between theory and computing by more easily representing these as algorithms (Wang *et al.*, 2022)

2.2.2 Modalities for Emotion Recognition

The core component of affective computing is emotion recognition which allows computer systems to perceive and interpret the different human emotional states through different techniques. Since emotions can manifest through a variety of behavioural and physiological signals that complement each other providing information of an individual's internal state. These signals are a core component for applications that range from a basic human-machine interaction to supporting healthcare and education.

There are several methods with which emotion can be recognized. Commonly the techniques focus on video, audio or text. These different modalities can be translated into different

components such as Facial expressions, speech and vocality, physiological signals and textual and linguistic features. Each of these components has different techniques which will be described:

- **Facial Expression**
    - **Facial Action Coding System (FACS)**: FACS is a comprehensive tool that is used to categorize facial movements based on their appearance. It's an extensively used technique for the analysis of facial expression recognition. (Ekman and Friesen, 2019)
    - **Facial Landmark Detection**: This technique focuses on the identification and analysis of key points of a face, such as eyes, nose, mouth and subtle changes in these for the progression and evolution of emotions. Recent studies are focused on the improvement of this technique(Cootes *et al.*, 1995)
    - **Static vs Dynamic Features**: This technique was developed based on the distinction between single-frames (static) and temporal progression (dynamic) which originated from the first analysis of facial expression recognition in videos (Bartlett *et al.*, 2005)

- **Speech and Vocal features**

  Speech being a form of communication, it is also a source of emotional information that can be utilized for emotion recognition analysis. Pitch, intensity, rhythm and other properties play an important part displaying the emotional state of an individual.(Scherer, 1995)
    - **Pitch**: The variation in a pitch can describe arousal and valence that was previously mentioned. While high pitched voices are more likely to indicate strong emotions, a lower pitch tends to be associated with more passive emotions. The study of the variation can represent the progress of someone's emotional state
    - **Intensity:** Intensity matches the amplitude of one's speech, and it usually reflects the emotional energy, just like with pitch, a louder speech is typically associated to stronger emotions, while a lower speech can reflect more passive emotions. Intensity patterns are often used to evaluate the overall emotional perception
    - **Speech Rate (Tempo)**: Speech rate represents emotional activity and cognitive load. Faster speed is often associated to positive emotions or emotional overload, while a lower speech rate can be linked to negative emotions or lower cognitive load. Speech patterns can also be identified in this subject such pauses or syllable duration to further categorise emotions.

- **Acoustic Feature Extraction methods**

  Prosodic features described above have broad perceptive patterns which acoustic feature extraction methods attempt to transcribe into numerical representations to enable computer models to analyse these signals more systematically. The techniques

utilised by this approach aim to capture fine-grained details in spectral, temporal and cepstral domains, which can be used by machine learning or deep learning models

- **Mel-Frequency Cepstral Coefficients (MFCC)**: MFCCs are among the most widely used method for speech-based emotion recognition. They aim to capture the short-power spectrum in a speech based on perceptual (Mel) frequency scale. This attempts to better simulate how the human auditory system would process the sound. Some subtle emotional cues such as stress or arousal can be better perceived in variations captured by MFCCs. (Davis and Mermelstein, 1980)
- **Linear Predictive Coding (LPC)**: LPC models the vocal tract as a filter to estimate spectral envelope of speech. Predicting the current samples, using previous samples as reference, it's possible to capture resonant frequencies, known as formants, that reflect both phonetic and paralinguistic information like emotion. A difference in coefficients may translate into distinctions between lower and higher arousal speeches. (Atal and Hanauer, 1971).
- **Formant Analysis**: As mentioned above, formants are resonant frequencies of the vocal tract, often associated with the production of vowels. Emotional states can shift the frequency of formants, allowing to identify different emotions through them, frequent formants are often associated with anger while a sadness lowers the frequency. (*Acoustic Theory of Speech Production: With Calculations Based on X-Ray ... - Gunnar Fant - Google Livros*, no date).
- **Spectral Features (Energy, Centroid, Flux, Entropy)**: These features are associated with the energy distribution across speech frequencies. The variation of signals in these features is usually a good indicator for emotional arousal and intensity (Mitra *et al.*, 2017).

- **Physiological Signals**

  Physiological signals are associated to autonomic, subconscious signals of the central nervous system that accompany emotional experiences. Since they're subconscious in contrast to facial or speech-based emotions, these signals can give a more accurate information on the emotional state of a person. These are more associated to healthcare and cognitive load assessment.
  - **Heart Rate (HR) and Heart Rate Variability (HVR)**: Emotional arousal can be often detected by cardiovascular activity. A higher heart rate can be a good indicator of excitement or stress, while a variety can be linked to the autonomic processes of the nervous system (Berntson *et al.*, 1997).
  - **Galvanic Skin Response (GSR)/ Electrodermal Activity (EDA)**: GSR measures the electrical conductance of the skin which may vary based on the activity of sweat glands due to the sympathetic nervous system control. It is mostly used as method for lie detects and stress detection(Dawson, Schell and Filion, 2007).

- **Electroencephalography (EEG)**: EEG captures brain activity through electrical potentials on the scalp. Emotional states can be traced to specific frequency bands and certain asymmetries on the hemispheres of the brain, presenting direct measures of cortical processes and the underlaying effect (Davidson, 1992).
- **Electromyography (EMG)**: EMG measures the electrical activity generated by the skeletal muscles, often used in affective computing to detect subtle facial muscular reactions to stimuli that are not perceived by the human eye, that are linked to changes in valence, either positive or negative (Cacioppo *et al.*, 1986).

- **Text and Linguistic features**
  Another modality used to express emotions is language, either spoken or written, where affective cues can be identified and extracted by computational methods to be analysed in either sentiment analysis or opinion mining. When comparing linguistic signals to physiological or facial signals, these are indirect but still offer important insight on one's emotional state, especially in modern digital contexts.
  - **Sentiment Analysis**: Sentiment analysis is the computational study of opinions, sentiments and emotions expressed in text. In traditional contexts, text extracts are classified based on their polarity, but currently it's possible to capture multiple degrees of the polarity and map sentiment and emotions. Through time these classifications evolved from using rule-based lexicons, into machine learning and deep learning models (Pang and Lee, 2008).
  - **Emotion Lexicons**: Emotion Lexicons are lists of annotated words linked to emotional associations, serving as base resource for text-based emotion recognition. Using the NRC, one of most widely used resource for text-based emotion recognition, it was compiled through crowdsourcing, specifically designed to support applications based on sentiment analysis, opinion mining and computational social sciences. It contains the annotations to represent the 8 basic emotions and both positive and negative polarity. Lexicons like this provide a interpretable and lightweight tool for affective computing and related fields, but they're lacking in some nuances related to context dependent language (Mohammad and Turney, 2013).

2.2.3 Techniques and Datasets

The development of affective computing systems relies on both the availability of high-quality datasets and the use of machine learning techniques capable of extracting and modelling complex emotional signals. While datasets provide the foundation for training and evaluation, techniques define how effectively these signals can be interpreted, combined, and generalized to real-world scenarios. This section introduces the core resources and methods used in

emotion recognition research, covering datasets, machine learning approaches, multimodal fusion strategies, and the architectures that integrate them.

2.2.3.1 Datasets and their Challenges

Research and development of affective computing relies on the availability of properly annotated datasets that provide the foundation for training and evaluating emotion recognition systems. There's a far wide of datasets with different modalities, for example facial images, audio, video, or even physiological signal, size of the dataset, type of annotation schemes, if its categorical or dimensional. For the correct development of a new systems, the correct selection of a dataset to achieve desired levels of performance and generalisation.

Some of the most popular datasets that are widely used by the scientific community are:

- **CK+ (Extended Cohn-Kanade Dataset):** This was one of the first datasets to be introduced, and so, turned into being one the most widely used, it contains high-resolution facial expression sequences that are labelled with basic emotions. It's regarded to be well suited for controlled laboratory conditions and posed expressions (Lucey *et al.*, 2010).

- **FER2013**: This dataset was introduced in a Kaggle competition, quickly becoming a benchmark for testing models in real-world scenarios despite the slight amount of noise in labels and its low resolution, due to the large sample of greyscale images categorised into the basic emotions (Goodfellow *et al.*, 2013).

- **AffectNet**: Currently seen as one of the largest datasets due to having over a million samples from the internet, with both categorical and dimensional classifications, it is seen has the cornerstone for deep learning training approaches for systems (Mollahosseini, Hasani and Mahoor, 2019).

- **RAF-DB (Real-world Affective Face Database):** Focused on more natural facial expressions, RAF-DB is also compromised by samples from the web, that have been labelled by trained annotators for both basic and complex emotions, increasing its validity (Li, Deng and Du, 2017).

- **CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset):** A multimodal dataset that contains audio and video clips of different actors expressing the basic emotions. It's widely used for the training of models that use multimodal fusion since it combines vocal and visual modalities (Cao *et al.*, 2014).

These datasets are differentiated by their multiple different characteristics and scopes such as:

- **Size:** Datasets can go from the thousands of samples (CK+, RAF-DB) to the millions (AffectNet). Datasets of different sizes end up having different purposes, for example, AffectNet as previously said due to its large size, it's mostly used for the training of deep neural networks, the smaller datasets are often used for controlled experiments.
- **Modalities**: Even though facial images are the most of modalities, datasets like CREMA-D introduce audio and video, aligning closer to real case scenarios.
- **Annotation scheme**: Most labels follow the categorical models, such as the Ekman basic emotions, or Plutchik Wheel, some follow dimensional models, focusing on the valence or arousal of ratings. AffectNet is an example of a hybrid categorical dataset that enables cross-model comparisons.
- **Controlled vs Wild**: CK+ and similar datasets are examples of controlled lab conditions ideal for a controlled environment, where AffectNet or RAF-DB being compiled with sources from the web, have a more natural display of emotions, managing to better reflect real-world scenarios.

However, as it is with most scenarios, both the compilation and selection of datasets is constrained due to multiple challenges that are naturally inherit to the subject.

- **Bias and Diversity:** Most datasets lack a proper representation of different age ranges, ethnicity and sometimes even gender, creating doubts on the veracity and accuracy of models trained under these datasets (Dominguez-Catena, Paternain and Galar, 2024).
- **Annotation Ambiguity:** Categorisation of emotions is often subjective, leading to the creating of noise on samples when different annotators disagree (Kim and Wallraven, 2021).
- **Context Dependence:** Due to the fact that most datasets are limited to facial images, it increases the difficulty or preparing systems to accurately evaluate emotions from different contexts, for example audio (Aguilera, Mellado and Rojas, 2023a).
- **Controlled scenarios vs Real-World generalisation:** Models trained using lab-controlled datasets tend to underperform when deployed on real-world based scenario, often due to external effects such as illumination, pose, background, sometimes even the intensity of emotions and expressions (Suresh, Yeo and Ong, 2021).

2.2.3.2 Machine Learning Techniques in Affective Computing

The analysis and recognition of human emotions have been approached through a wide variety of machine learning methods, ranging from classical statistical models to more advanced deep learning and transformer-based architectures. The choice of technique has historically been guided by the availability of annotated data, computational resources, and

the complexity of the target application. This section outlines the progression of machine learning techniques for affective computing, highlighting their main contributions and limitations.

- **Classical Machine Learning Approaches**: Early works in affective computing relied heavily on classical machine learning algorithms, for example Support Vector Machines (SVMs) (Cortes, Vapnik and Saitta, 1995), Random Forests (Breiman, 2001), and Hidden Markov Models (HMMs) (Rabiner, 1989). These methods typically required handcrafted feature extraction, where prosodic cues (e.g., pitch, intensity) or geometric facial descriptors (e.g., distances between landmarks) were first computed and then fed into classifiers. SVMs proved effective for small-scale and well-structured datasets, while HMMs were frequently applied to capture the temporal progression of emotions in sequential data such as speech or facial videos. Despite their success in controlled environments, these approaches were limited in scalability and struggled with the variability found in real-world data.
- **Deep Learning Advances**: The introduction of deep neural networks, and especially Convolutional Neural Networks (CNNs) (LeCun *et al.*, 1998), marked a paradigm shift in emotion recognition. Unlike classical methods, CNNs are capable of learning feature representations directly from raw data, significantly reducing reliance on manual feature design. For facial emotion recognition, CNNs demonstrated substantial improvements in accuracy by capturing fine-grained spatial information from images. In parallel, Recurrent Neural Networks (RNNs) (Elman, 1990) and their extensions, particularly Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks, were introduced to better model temporal dependencies in dynamic modalities such as video sequences or speech signals. These architectures enabled emotion recognition systems to move beyond static analysis, incorporating context and temporal evolution into predictions.
- **Transformer-Based Models and Attention Mechanisms**: Recent advances in natural language processing and computer vision have introduced transformer architectures into affective computing. Transformers rely on self-attention mechanisms that allow them to capture long-range dependencies across data sequences, making them particularly suitable for multimodal and temporal emotion recognition tasks. For example, transformer-based models have been successfully applied to combine facial and vocal features, attending to the most salient signals across modalities. Beyond their accuracy, transformers also offer interpretability advantages, as attention weights can provide insights into which signals contribute most to emotion prediction (Vaswani *et al.*, 2017).
- **Transfer Learning and Domain Adaptation**: Another key development in affective computing has been the application of transfer learning (Pan and Yang, 2010), where models pretrained on large-scale datasets (e.g., ImageNet for vision, or large speech corpora) are adapted to emotion recognition tasks. Transfer learning allows for effective use of smaller and domain-specific datasets, reducing the need for extensive manual annotations. Complementary to this, domain adaptation techniques (Ganin

and Lempitsky, 2014) address discrepancies between training and deployment environments, such as differences in recording conditions, demographics, or cultural contexts. These strategies are critical in bridging the gap between controlled laboratory datasets and real-world applications.

In summary, the trajectory of machine learning techniques in affective computing has progressed from handcrafted feature-based methods to deep, data-driven representations and finally to highly adaptable architectures that leverage pretraining and multimodal integration. Each stage has addressed the limitations of the previous one, moving the field closer to robust and generalizable systems capable of functioning effectively in unconstrained real-world environments.

2.2.3.3 Multimodal fusion approaches

Multimodal fusion approaches refer to the techniques introduced to integrate the previously mentioned multiple modalities of affective computing, like facial expression, speech, text, physiological signs, gestures, amongst others for emotion recognition. Single modality channels are often considered insufficient to accurately detect emotions, due to subtility or ambiguity associated with them, for example the masking of emotions through visual signs while voice or physiological inputs would detect, or sarcasm detected in tonality that a text input wouldn't detect. Fusing modalities in a multimodal approach allows the leveraging of complementary cues between the different modalities which boosts the performance of models across the multiple conditions, such as accuracy, robustness, and generalisation.

The main challenge of using these techniques is the combination of heterogenous data through different temporal, spatial and statical properties. Properly coordinating the different modalities like audio, video and text, can require some lengthy data preprocessing to have it all properly assigned and aligned to each correct interval (Li and Tang, 2024).

Currently there are three fusion approaches that are widely being used in the development of multimodal solutions.

- **Early fusion (Feature-Level):** Early fusion techniques are focused on the aggregation of features from the multiple modalities, such as MFCCs from speech, distance between facial landmarks, into single vectors that are fed to classifiers or predictive models. This allows for classifiers to directly learn the correlations between the different modalities, while turning model more expressive since all the features are presented simultaneously. This increases the potentiality of improved results for the model, due to the cross-modal interactions happening at a feature level, displaying how the multiple modalities complement each other. This also simplifies the design of the models, since the input later will contain all the information. However, this causes the issues such as large dimensionality that

increases the risk of overfitting, often needing to be reduced before being fed to the model. Temporal mismatches are also possible causing some noise in the data and this technique, since it needs to be fed all the features as an input, it may be able to not recognise modalities that aren't present, causing performance issues. Imputation techniques may circumvent the issues, but the complexity of the model is increased (Atrey *et al.*, 2010).

- **Late Fusion (Decision-Level):** Contrasting early fusion, with late fusion, each modality is processed individually, being processed on their model or classifier and the output results are combined through different methods such as voting, averaging, weighted sum or other, more complex strategies**.** Since this approach keeps the modalities separated until the final decision, it is more robust to missing or noisy data. The separation of models also allows for their individual use and optimization, leading to an improved scalability of the system. Besides the increased robustness to noisy or missing modalities, and the scalability, late fusion also allows the incorporation of uncertainty by weighting predictions according to the reliability of each modality, but the separation of the modalities, makes it that it's possible to miss the interactions of modalities, losing out on subtle emotional cues (Zeng *et al.*, 2009).
- **Hybrid fusion:** Hybrid fusion incorporates both the previous methods by having a selected number of features and modalities be fused at a feature level and having a decision level fusion at the end to ingrate the fusions of each stream. Hybrid fusion is usually implemented as a hierarchical model or neural network with modal specific minor networks that are followed by fusion layers. This approach attempts to balance the advantages of each approach, while minimising on the drawbacks that each present. While inter-modal interactions are kept and the flexibility of prioritising the most reliable modalities, the increased complexity increased the difficulty of the design of the model, tuning of hyperparameters, while demanding more computational resources. The balancing of contributions of feature-level and decision-level also have increased difficulty since that highly dimensional vectors at the early fusion may lead to overfitting (Baltrusaitis, Ahuja and Morency, 2019).

2.2.4 Frameworks, Challenges, Applications and Recent Trends in Affective Computing

The development of affective computing systems has advanced rapidly, yet several practical and theoretical challenges remain. These limitations influence the performance, generalizability, and usability of emotion recognition technologies. Alongside these challenges, the field has found applications across multiple domains and is embracing new approaches and trends to overcome current limitations. The following sections summarize the key challenges, application areas, and emerging directions in affective computing.

2.2.4.1 Multimodal Emotion Recognition Architectures

As mentioned above, multimodal emotion recognition systems can integrate multiple sources of information from the different modalities, like facial expressions, speech and physiological signals as a method to improve the robustness of emotional inference. The combination also helps addressing the ambiguity that is inherent to single channels, real-time awareness and context awareness. As such, the proper election of fusion strategies and the correct architectural design of system is necessary to create trustworthy and applicable systems to the different fields.

- **Healthcare:** Emotion recognition in healthcare is primarily used on the monitoring of mental health, stress and affective states as a supportive intervention or reinforce medical assessments. Typically, in this field facial expression ns physiological signals like heart rare or GSR are the primary modalities to be used, applying late fusion strategies to reduce asynchronous or noisy signals during the clinical assessment, ensuring correct results even if a modality is missing. (Pan *et al.*, 2023) introduce a deep learning-based multimodal emotion recognition labelled Deep-Emotion. This system integrates facial expressions, speech and EEG adaptively selecting the most discriminative features. (Alarcão and Fonseca, 2019) focused their work on EEG-based emotion recognition, emphasizing the role of EEG to identify affective disorders, introducing a benchmark dataset and algorithm for this purpose, underlining how physiological signals complement both facial and vocal subtility in healthcare contexts.

- **Education:** In education, the role of emotion recognition to enable adaptative learning environments that can dynamically evolve to adapt to student's cognitive and affective states. These intelligent tutoring systems are mainly using facial expressions, voice and their gaze to determine student's engagement, frustration or even confusion with certain subjects. A theorical basis for an affect-aware was introduced by (D'Mello and Graesser, 2014) for education. It focuses on detecting affective states like boredom, confusion or interest to improve feedback loops, managing to improve students learning outcomes. More recent works such as of (Ye, Zhou and Tao, 2023) introduced a synchronous multimodal dataset of learners watching educational or instructional videos, while capturing features such as eye movement, physiological signals, and the video. They propose a new model that integrated features related to knowledge to boost the emotional representation reporting a 10% increase of accuracy compared to uni-modal or baseline models.

- **Human-Robot Interaction (HRI):** Robots are used in different fields across society, being necessary that these are capable of detecting emotions to establish trust and facilitate their ability to maintain natural communication. Using a multimodal approach to detect affective states, robots should be able to adapt their speech, gestures and behaviour in real time to be more "humanlike". (Spezialetti, Placidi and

Rossi, 2020) introduce a complex emotional communication-based HRI system, by integrating facial expressions, speech, posture and EEG signals on NAO robots and other devices to recognise emotion emotions so they can respond accordingly. The study focuses on the importance of fusion techniques to enhance the naturality and effectiveness of the interactions of the robots. (Su *et al.*, 2023) present an extensive comprehensive review on recent advancements on HRI systems, mostly focusing on the integration of vision, audio and touch to enhance the perception and responsiveness of robots. They discuss current challenges and methodologies in synchronizing and processing multimodal data, focusing on the importance of systems that are context-aware and can adapt to human behavioural changes and environmental change.

- **Marketing and customer experience personalization:** Multimodal emotion recognition is a core feature to enhance marketing and costumer experience personalization by allowing smart systems to understand and respond to the emotional states of users. Focusing mostly on modalities like facial expressions, tonality, textual sentiment, and physiological signals, businesses can create more responsive and empathetic interactions. This leads to a higher customer satisfaction, leading to emotional connections that increases loyalty and brand perception. **Emotion-LLaMA** (Cheng *et al.*, 2024) integrates audio, visual and textual inputs through different emotion-specific encoders, using MERR dataset to enhance emotion recognition across the multiple possible scenarios. The system achieves high state-of-the-art performance with top scores in multiple competitions. Another system is introduced by (Lee *et al.*, 2024) where a multi-modal human recognition system that can efficiently comprehend emotional information by combining both verbal and non-verbal expression data. They use a system based on personalized-skin integrated facial interface that is integrates data processing circuit to wirelessly transfer data to enable real-time emotion recognition.

2.2.4.2 Challenges and Limitations

As previously mentioned, affective computing has multiple challenges and limitations that pose a roadblock to the development of new affective computing applications. Besides computational and data acquisition difficulties that are natural to the develop of AI models, there are some that are predominant in affective computing.

- **Real-time Recognition Constraints:** Emotion recognition applications tend to operate in dynamic environments where real-time processing is necessary to give instant feedback to users, but achieving a low-latency performance without dropping accuracy is one of the main challenges. Most studies focus on the necessity of optimization of hardware and algorithms to allow real-time emotion detection in practical applications (Yu *et al.*, 2023).

- **Cultural and Demographic Variability:** Emotional expressions are influence by cultural norms, age, gender, and other demographics factors. Models that are trained with homogenous datasets within these characteristics may lead to biases, which reduces the accuracy and fairness in a global perspective. Research reinforces the need of diverse datasets to develop generalised and adaptative models that can address these issues (Abbruzzese *et al.*, 2019).

- **Ambiguity and Subjectivity in Emotion Labelling:** Emotion labelling has a crucial flaw that its subjective nature, with different annotators having mismatched readings of different emotional expressions. This variability is a common reason to existence of noise in different training datasets, which affect the reliability of emotion recognition systems. The community has been appealing for the introduction of ambiguity-aware models that can handle the inconsistencies of these labelling issues (Cowen and Keltner, 2021).

- **Privacy Concerns:** The acquisition, storage and analysis of data have been a growing concern of users, and sensitive biometric data, such as facial expressions and physiological signals, raise multiple privacy concerns. The unauthorized data collection and potential misuse under the guise of surveillance purposed have created regulatory scrutiny. Many legal frameworks are evolving to address such concerns, advocating the need for transparent data practices and user consent (McStay, 2020).

2.2.4.3 Recent Trends and Future Directions

Multimodal emotion recognition being a crucial tool in multiple fields, it is quickly evolving due to advances in machine learning, modelling, an increased number of multimodal datasets. Current research shifted its focus towards core concerns related to the limitations of the existing systems, that recently have been increasing such as data privacy, generalisation of across different demographic groups, contextual boundaries and real-time conditions. The goal of this section is to highlight the current works to continue improving multimodal systems.

- **Federated Learning for Privacy-Preserving Emotion Recognition:** Due to the sensitive data that is associated with Affective Computing, usually it being biometric data that can be traced back to individuals, facial emotion recognition is an even higher challenge as the primary data it's the face of a person, making them instantly recognised. The emergence of Federated learning provided a promising solution to this issue. The decentralized nature of this method allows for conservation of data on trusted edge devices, capable of training models, therefore ensuring the privacy of data, while still allowing the performance of large-scale collaborative learning systems. Such work that displays this evolution was presented by (Gahlan and Sethia,

2024b), where they do a comprehensive review on the integration of Federated Learning in emotion recognition systems, highlighting the potential of multimodal FL architecture they dubbed as Fed-PhyERS. They aim to evaluate the feasibility but also address some issues like limitations and ethical considerations in privacy-aware systems.

- **Multimodal Transformers and Self-Supervised Learning:** Transformers and self-supervised learning (SSL) methods are being applied to evolve multimodal emotion recognition systems by reducing their dependence on annotated datasets, which are harder to acquire and might be biased in for certain demographic character. SSL allows for models to exploit large multimodal data samples, which use transformer architectures provide the means to do long-range dependencies and cross-modal relationship for models. The combination of these methods provides a robust and adaptative model. (Wu, Daoudi and Amad, 2023) present a framework that leverages modality-specific encoders that when combined with a transformer-based encoder to capture both intra and inter modal correlations. They demonstrate the ability of the system to significantly enhance the performance of wearable emotion recognition systems, especially in low-resource labelling settings.
- **Cross-cultural and cross-domain generalization:** Emotion recognition systems can demonstrate performance drops when applied in different cultures or domains than ones initially trained or deployed in. These limitations are derived from how emotions are expressed and perceived in varying cultural and situational contexts, either from culturally specific reactions or expressions, or from different demographic features that can't be transcribed through cultures. This key issue is one that needs to be tackled and addressed by system that want to have a reliable performance in a global scale. The work of (Liang *et al.*, 2019)propose an adversarial learning framework to reduce both domain and cultural discrepancy in multimodal emotion recognition. The features are learned from audio and visual inputs while at the same time an opposing culture discriminator is trained to be "confused" by the feature encoder. Because of this method, the encoder learns about features that are predictive of emotions without having cultural as a variant. This type of approach is typical of different domain-adversarial adaption applied specifically to cross-culture MER.
- **Emotion recognition in the wild (unconstrained environments):** Most emotion recognition systems are developed and evaluated in controlled and supervised conditions, like laboratories, which often underperform when deployed in real-world environments introduce variability such as background noise, occlusion and lighting issues, increasing the challenge of emotion recognition. Progress in this topic are necessary to develop more robust system in different domains like healthcare, HRI, customer service. (Aguilera, Mellado and Rojas, 2023b) assess existing in-the-wild datasets, specifically designed for multimodal emotion recognition, highlighting the challenges and opportunities of moving beyond controlled environments. Their study

focuses on the need for systems to be able to remain accurate despite noise or lack of data.

2.2.5 Chapter Conclusion

This chapter has surveyed the foundations and current directions of affective computing. It began with theories and models of emotion, which provide the conceptual basis for computational recognition, and reviewed the main modalities used to capture affective signals. Computational techniques and benchmark datasets were then discussed, highlighting the progression from single-modality approaches toward multimodal frameworks, while also exposing issues such as data alignment, generalization, and bias.

Applications in healthcare, human–computer interaction, and marketing illustrated the wide relevance of emotion-aware systems, while recent trends such as federated learning, multimodal transformers, and cross-cultural generalization pointed to ongoing efforts to make these systems more robust and ethically responsible.

In conclusion, affective computing is moving toward more accurate, generalizable, and privacy-preserving models, yet significant challenges remain in interpretability, fairness, and deployment in real-world contexts.

2.3 Recommendation Systems

Despite not being within the scope of the thesis, recommendation systems are a component of the project which this thesis is part of, so I decided to make a short review of these to better explain the architecture of the system in the chapters ahead.

Recommendation systems are software tools and techniques designed to assist users in discovering content that fits their potential taste and interest, mostly being provided by services aimed at end users. Their primary objective is to reduce information overload by providing personalized suggestions, thereby improving user experience, engagement, and decision-making efficiency, creating user trust and brand loyalty (Ricci, Rokach and Shapira, 2022). These systems are now widely used across domains such as e-commerce, digital media platforms, and social networks, where the volume of available content far exceeds the user's capacity to explore manually. By automatically identifying relevant items for each user, recommendation systems play a critical role in guiding interactions between users and the system.

At a high level, recommendation systems are commonly categorized into three main types.

- **Content-based filtering** relies on the attributes of items and a user's historical preferences to generate recommendations (Pazzani and Billsus, 2007). For example, in a movie recommendation system, the method would suggest films that share

characteristics—such as genre, director, or keywords—with movies previously consumed by the user. This approach emphasizes the features of items themselves and the alignment with user profiles, allowing personalized recommendations even when few users have interacted with the system.

- **Collaborative filtering**, in contrast, leverages the collective behaviour of users to generate suggestions (Herlocker *et al.*, 2004). Instead of analysing item attributes, this method identifies patterns in user interactions, such as ratings or purchase histories, to recommend items preferred by similar users. Collaborative filtering is particularly effective when user behaviour provides rich signals, but item features are difficult to quantify, and it often uncovers interests that the user might not have explicitly expressed.
- **Hybrid approaches** combine content-based and collaborative filtering techniques to leverage the strengths of both while mitigating their respective limitations (Burke, 2002).By integrating multiple strategies, hybrid methods can improve recommendation accuracy, address the cold-start problem, and provide more robust suggestions across diverse user populations.

Within the scope of the project that the development of this thesis was developed in, recommendation systems will interact with multimodal emotion recognition system, receiving the emotional outputs and some textual sentiment extraction of features, like requests of product characteristics, to boost the performance of the recommendation system, being more sensitive to the user's emotional state and capable of providing recommendations that are more aligned with the user's interests.

In the context of this project, the presence of recommendation systems highlights the importance of handling user data in a structured and efficient way. As these systems rely on patterns extracted from user interactions, the way data is collected, stored, and processed directly impacts their effectiveness and the overall system performance. This consideration becomes particularly relevant when exploring approaches that prioritize data privacy and decentralized learning, such as federated learning. The following chapter discusses the integration of federated learning into both sentiment analysis and recommendation modules, illustrating how these techniques can work together to enhance system capabilities while preserving user privacy.

2.4 Chatbots

In a similar fashion as the recommendation systems, chatbots aren't a core topic of this thesis development, however, they're a core component of the project where this thesis is centred in, making it in my view, a topic worth going over slightly to describe how the multimodal models will interact with the chatbots.

Chatbots are software agents designed to interact with users through natural language, providing responses and guidance in a conversational manner. Their main goal is to facilitate communication, automate tasks, and enhance user experience by simulating human-like interaction (Shum, He and Li, 2018). Chatbots are widely used across customer service, digital assistants, healthcare, and educational applications, where they can efficiently handle large volumes of queries while providing personalized interactions.

From a functional perspective, chatbots can be broadly categorized into two main types.

- **Rule-based chatbots** operate according to predefined scripts and decision trees. They are limited to scenarios anticipated by their developers but can deliver highly controlled responses in structured environments (McTear, 2002).
- **AI-based chatbots**, in contrast, leverage natural language processing (NLP) and machine learning techniques to interpret user input and generate responses dynamically. These systems can handle a wider variety of queries, learn from user interactions, and provide more flexible and context-aware responses (Adamopoulou and Moussiades, 2020).

Within the context of this project, chatbots play a pivotal role in interacting with users while integrating insights from multimodal emotion recognition and sentiment analysis. The output of the sentiment analysis model informs the chatbot about the user's current emotional state,

allowing it to tailor its responses accordingly. This integration enhances user engagement and ensures that the system provides replies that are not only contextually relevant but also emotionally appropriate.

Although the focus of this thesis is not on the development of chatbots themselves, understanding their function and interaction with other components is essential for grasping the system architecture. By situating chatbots within the broader framework of multimodal emotion recognition and recommendation systems, this chapter provides the necessary context to appreciate the design and workflow of the system modules discussed in subsequent chapters.

2.5 Integration of Federated Learning into Affective Computing and Recommendation Systems

The integration of Federated Learning, Affective Computing and its branches, and Recommendation System has the objective to develop intelligent systems that have as a focus, the capability to preserve data privacy, have personalized user experience and are emotionally aware to increase engagement. Federated Learning enables collaborative model training across multiple edge devices without transferring the raw data of user data, being able to address the critical issues that is handling sensitive multimodal data. Affective computing provides the system the capabilities to detect and interpret human emotions and

affective state, while recommendation systems focus on providing personalized content to the users. The combination of these fields makes system adaptable to different real-world environment, while being emotionally aware of the user's emotional state, being able to be more sensitive with its engagement, while conserving data privacy and providing the user with the content they prefer. There have been multiple works in the recent years that have displayed the integration of these different complementing technologies, to develop more robust systems.

- **Federated Multi-View Deep Learning for Privacy Preserving Recommendations** (Huang *et al.*, 2020): Their work introduces a federated multi-view framework to be deployed with recommendation system, focusing on privacy preservation capabilities of federated learning by keeping the personal data of users on their local devices. The multi view comes from the integration of multimodal data, this case, text, images and ratings to boost the accuracy of recommendations. Their work showcased that federated approaches are capable of matching traditional centralized framework's accuracy of recommendations.
- **FED-ReMECS** (Nandi and Xhafa, 2022): Presents a federated learning framework capable of performing real-time multimodal emotion recognition, through audio, video and physiological signals. Their framework was developed to be able to handle data heterogeneity and also personalisation at a device-level. The goal was to focus on preservation of private sensitive data, often associated to emotion recognition.
- **Federated Collaborative Recommendation Model** (Lin *et al.*, 2023): Introduced the proposal of the utilization of federated learning for a collaborative filtering model for recommendation systems, by utilising data associated with distributed user-item interactions while preserving its privacy. Their work displays that federated collaborative models can achieve comparable results to the centralized approaches, focusing on personalised recommendations without ever exposing individual data of each user.
- **Enhancing Emotion recognition through FL** (Simić *et al.*, 2024): This work utilises a federated learning framework with CNN based models for multimodal emotion recognition focusing on audio, video and textual features, displaying the intended goal of a privacy-preserving training on user devices. Their work highlights the improved accuracy on emotion recognition and personalization at a device-level.
- **AFLEMP: Attention-Based Federated Learning for Multimodal Physiological Data** (Gahlan and Sethia, 2024a): This work introduces a federated, attention-based model for emotion recognition that focus on the multiple physiological signals, like ECG, GSR, EEG. It uses attention mechanisms to weight features from different modalities, while ensuring that privacy is kept by preserving the sensitive physiological data locally, while still achieving high accuracy results.
- **ClusterFedMeta: A Federated Recommendation Algorithm Based on User Clustering and Meta-Learning** (Yu *et al.*, 2024): They propose the introduction of a federated recommendation algorithm combining user clustering with meta-learning approaches, addressing data heterogeneity and privacy concerns of federated environments. They

demonstrate effectiveness in personalised recommendation tasks while maintaining the user data on local devices.

Table 1 - Recent contributions in FL, SA and RS

Citation	Application	Data Modality	FL Method	Key Contributions
(Huang *et al.*, 2020)	Recommendation Systems	Text, images, ratings	Federated multi-view deep learning	Introduced federated multi-view learning; privacy-preserving; matches centralized performance; integrates multimodal data for improved recommendations
(Nandi and Xhafa, 2022)	Real-time Emotion Recognition	Audio, video, physiological signals	Fed-ReMECS	Federated real-time emotion recognition; handles multimodal streaming; device-level personalization; privacy-preserving
(Lin *et al.*, 2023)	Recommendation Systems	User-item interaction data	Federated collaborative filtering	Distributed recommendation with privacy; personalized recommendations; comparable performance to centralized models
(Simić *et al.*, 2024)	Emotion Recognition	Audio, video, text	CNN-based federated learning	Combines CNNs with FL for multimodal emotion recognition; improves accuracy; privacy-preserving on-device training
(Gahlan and Sethia, 2024a)	Emotion Recognition	Physiological signals (ECG, GSR, EEG)	Attention-based FL (AFLEMP)	Attention mechanism for multimodal physiological data; keeps data local; improves accuracy; privacy-preserving
(Yu *et al.*, 2024)	Recommendation System	User interaction data	Federated Meta-Learning	Proposed ClusterFedMeta, a federated recommendation algorithm combining user clustering with meta-learning, addressing data heterogeneity and privacy concerns.

2.6 Chapter conclusion

This chapter has presented a comprehensive overview of the state of the art in Federated Learning, Affective Computing, and Recommendation Systems, highlighting both foundational concepts and recent advancements. Beginning with Federated Learning, the chapter explored its operational processes, including centralized, decentralized, and heterogeneous frameworks, as well as the categorization of horizontal, vertical, and transfer learning

approaches. The discussion emphasized the inherent challenges of FL, such as data heterogeneity, communication bottlenecks, client bias, and fairness considerations, alongside emerging solutions that leverage hierarchical architectures, privacy-preserving mechanisms, and fairness-aware aggregation. Furthermore, the chapter highlighted current advances and future directions, including model alignment, self-attention architectures, and adaptive aggregation strategies, illustrating the dynamic evolution of FL research.

In parallel, the chapter examined Affective Computing and Emotion Recognition, providing an overview of prevailing emotion models and theories, multimodal recognition techniques, key datasets, and contemporary frameworks. The challenges associated with capturing, interpreting, and integrating diverse emotional signals were addressed, with emphasis on the need for robust, privacy-preserving, and adaptive solutions. Additionally, recent trends in applications of Affective Computing—ranging from healthcare to human–computer interaction—demonstrated the increasing relevance of integrating machine learning approaches, including Federated Learning, to manage distributed and sensitive data.

Finally, the chapter discussed Recommendation Systems and Chatbots, highlighting their operational principles, personalization strategies, and integration with machine learning models. The section on the convergence of Federated Learning with Affective Computing and Recommendation Systems underscored the potential benefits of combining privacy-preserving decentralized learning with adaptive, user-centric applications. Collectively, this chapter establishes a solid foundation for subsequent discussions on the design and implementation of integrated systems, providing both theoretical context and practical insights into the opportunities and challenges inherent in combining FL with emotion-aware and recommendation-driven technologies.

3 Methods and Materials

This chapter aims to describe the methods tools that will be used in the development of the proposed framework. It will also be described the ethical questions that can be posed to this solution.

3.1 Materials and Tools

In this sub-section, it will be introduced the FL framework that will be used and the datasets that will be explored for the first experiments.

3.1.1 Federated Learning Framework

The development of the proposed framework involved a thorough evaluation of existing open-source Federated Learning (FL) frameworks to determine the most suitable option for handling communication with the server and the implementation of federating strategies. Among the frameworks considered, FLOWER was chosen as the primary framework for several compelling reasons.

FLOWER stood out due to its user-friendly documentation, rapid implementation capabilities, and adaptability. Notably, FLOWER offers the flexibility to operate with either TensorFlow/Keras or PyTorch, permitting users to select the library that aligns with the specific algorithm requirements.

During the framework selection process, other FL frameworks were also explored, with specific criteria guiding the assessment. The primary selection criteria included open-source availability and ease of use. Several alternative frameworks were evaluated and subsequently excluded based on these criteria:

1. **PySyft + PyGrid**: Despite its potential, this framework was discarded due to difficulties in implementation, particularly in understanding the implementation process.
2. **Tensorflow Federated**: While initially considered, challenges arose in trying to implement this framework according to the provided documentation, leading to its exclusion from the selection process.
3. **Fate**: The Fate framework was not experientable due to its compatibility limitations, as it only functioned on Linux and Mac platforms. This limitation, combined with the extensive time required for testing, made it impractical for the current project. The framework's specifications were gathered based on collective user ratings rather than firsthand experimentation.
4. **OpenFL**: This framework was not chosen because it mandated the use of its specific communication protocols. Given that the primary goal of the prototype was to gain a fundamental understanding of FL and its implementation, OpenFL was deemed unsuitable for the project's scope.

5. **Other Complex or Domain-Specific Frameworks**: Several other open-source FL frameworks were deemed too intricate for the project or were tailored for specific domains, such as NVIDIA Clara, which is primarily designed for healthcare applications. These frameworks were considered less suitable for the initial stages of development.

It is important to note that the selection of FLOWER as the chosen framework was based on its ease of use, flexibility, and compatibility with the project's objectives. Depending on the performance and evolving requirements of the project, future reconsideration of the framework choice may be warranted.

Table 2- Comparison of Federated Learning Frameworks

Framework	Usability	Usage	OS	Open Source
Flower	Simple to use	Unrestricted	All	Yes
PsySyft + PyGrid	Difficulties implementing	Research Purposes	All	Yes
TensorFlow Federated	Difficulties implementing	Unrestricted	All	Yes
Fate	Flexible	Unrestricted	Mac, Linux	Yes
OpenFL	Requires specific certificates	Unrestricted	All	Yes
NVIDIA Clara	Not tested	Healthcare	All	Yes

3.1.2 Datasets

The only dataset currently being used in this work is the CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset), which is a publicly available and widely recognized dataset in the field of emotion recognition and affective computing. CREMA-D contains more than 7,000 short audio-visual recordings of 91 actors, each performing a set of scripted sentences with varying emotional expressions. These expressions cover a range of basic emotions, including anger, disgust, fear, happiness, neutrality, and sadness. The dataset provides synchronized audio and video data, enabling the development of models that can process and learn from multiple modalities simultaneously. The recordings were collected under controlled conditions but vary in vocal effort and intensity, offering a diverse representation of expressive behaviour.

Each clip in the dataset has been annotated by multiple raters through a crowd-sourced process, capturing both the emotion that was intended by the actor (displayed emotion) and the emotion that was interpreted by the raters (perceived emotion). This dual annotation approach allows for a more nuanced understanding of emotional communication and supports a range of research objectives, including the analysis of differences between expression and perception. The structured format and high annotation quality of the dataset make it suitable for training and evaluating machine learning models in classification, regression, and multi-task learning contexts.

Although CREMA-D is currently the only dataset being used in this work, the overall system has been designed with extensibility in mind. It has been structured to support the future integration of real-world datasets derived from experimental of the deployment scenario. These future datasets are expected to include live or recorded emotional responses collected in cyber-physical environment, allowing the proposed system to transition from controlled research setting to more practical, real-time use cases. As such, while CREMA-D serves as the foundation for initial development and evaluation, the methodology remains adaptable to accommodate additional data sources as the project progresses.

3.2 Methodologies

3.2.1 Implementation Methodologies.

In the process of implementing the framework, the initial phase involves leveraging existing tools and models to construct a prototype. This crucial step enables us to kickstart the development process efficiently. One key component of this phase is the model used for training, which is currently based on deep learning with Rectified Linear Unit (ReLU) activation layers, as well as a softmax layer on local nodes. This approach is informed by the work of (Fred Agarap, 2018) which serves as a solid foundation for our training model.

When it comes to Federated Learning (FL), we are actively exploring the possibility of implementing a variant of Federated Averaging (FedAvg) that is better suited to handle data originating from Internet of Things (IoT) devices. This adaptation is essential, as IoT devices can produce highly heterogeneous data. To simulate this diversity, we plan to consider different data loads of the same dataset for each client or even use entirely different datasets with similar data characteristics. This strategy aims to ensure that the initial trained models are well-prepared to exhibit behaviour closely aligned with our expectations during the subsequent deployment phase of the framework.

Now, regarding the aggregation strategy, we are still in the process of evaluating various options. The literature offers a wide range of possibilities, each with its own advantages and trade-offs. It is possible that we may opt for a customized aggregation strategy that builds upon the principles of a state-of-the-art approach. This flexibility allows us to tailor the strategy to the specific needs and nuances of our framework as it evolves.

As we make progress in achieving our development milestones, we remain open to new considerations on how to further enhance the proposed system. Continuous improvement

and feasibility assessments are integral to our approach, and we are committed to adapting to emerging challenges and opportunities.

In terms of communication between clients and the aggregation server, we are also exploring the incorporation of a blockchain layer. This addition is aimed at bolstering the security of transferred model weights and mitigating the risk of potential leaks. By implementing blockchain technology, we aim to fortify the integrity and confidentiality of data exchanges, ensuring that our framework adheres to the highest standards of security.

Overall, our framework is a dynamic and evolving project, driven by a commitment to innovation and excellence. We look forward to the exciting journey ahead as we work towards realizing its full potential.

3.2.2 Testing methodologies

The study outlined here is a comprehensive investigation into the effectiveness and efficiency of a novel federated learning framework compared to a conventional federated learning approach, specifically focusing on the default implementation of the FLOWER framework as described in its official documentation. This study aims to contribute to the growing body of knowledge surrounding federated learning by addressing several key aspects.

Comparative Analysis of Aggregation Strategies: One of the central focuses of this research is to compare the aggregation strategy employed in the novel framework with the widely used FedAvg strategy. The primary objective is to understand how this alternative aggregation method, if implemented, impacts various performance metrics, including accuracy, convergence speed, and model generalization. This comparison will provide valuable insights into the potential improvements or trade-offs associated with different aggregation techniques in federated learning.

Resource Utilization and Execution Time: Beyond performance metrics, the study aims to delve into the computational resources required by both the novel federated learning framework and the traditional FLOWER implementation. This analysis will include aspects such as CPU and memory consumption. Additionally, the study will measure the execution time of both frameworks to assess their efficiency and scalability. These findings will help evaluate the practical feasibility of adopting each framework in real-world scenarios.

Comparison with Traditional Machine Learning: To provide a holistic perspective, the research will also conduct a comparative analysis between federated learning and traditional machine learning approaches. This comparison will consider various factors, including model performance, training time, and data security. By juxtaposing the results obtained from these two approaches, the study will shed light on the advantages and limitations of federated learning in different contexts.

Anticipated Outcomes: It's worth noting that existing literature suggests that federated learning can sometimes yield suboptimal results compared to centralized approaches due to the inherent challenges of decentralized training. However, the study anticipates that federated learning will demonstrate clear advantages in terms of reduced training time and enhanced data privacy and security. The overarching goal of this comparative study is to

determine whether the observed differences in performance metrics between federated and traditional approaches justify the potential benefits, particularly in terms of data security and efficiency.

In summary, this comprehensive analysis seeks to provide valuable insights into the trade-offs involved in adopting federated learning in real-world applications. By examining multiple facets of performance, resource utilization, and security, the study aims to clarify when and how federated learning can strike a balance between performance, data security, and efficiency, ultimately contributing to the informed adoption of this innovative approach in various domains.

3.3 Security and Ethical questions

As previously mentioned, Federated Learning holds significant potential to address the pressing privacy and security concerns raised by various entities. In recent years, the issue of safeguarding user data has garnered substantial attention, prompting the introduction of regulations such as the General Data Protection Regulation (*EUR-Lex - 32016R0679 - EN - EUR-Lex*, no date).

FL, which is an edge computing-based AI technique, ensures that data never leaves the user's device, thereby instilling confidence in users regarding the safety of their personal information. However, there remain concerns surrounding the potential exploitation of weights transferred to the server for aggregation, which could potentially be used to glean information from a node.

In a comprehensive study by (Mothukuri *et al.*, 2021) , the authors delve into the current threats and vulnerabilities within FL, including the possibility of security attacks like inference attacks and unintentional data leakage. They also propose several solutions to mitigate the risk of these attacks and discuss the trade-offs associated with implementing these measures.

Most research in this field acknowledges that FL inherently offers data protection at the user or local node level, making it an appealing choice with minimal ethical drawbacks and vulnerabilities from a user's perspective. However, there is a unique challenge in FL's reliance on edge computing: willing participants in the system are essentially agreeing to use their own devices for training, thereby expending their own resources. This aspect might make some individuals more hesitant to participate, despite the enhanced data security compared to traditional approaches.

In the proposed framework, where training occurs on kiosks devices or devices to which IoT data is transmitted, these resources are already considered part of the system. Consequently, this concern does not apply in this scenario, and conventional security approaches typically adopted in cyber-physical systems can be utilized to address any remaining security considerations. This approach aims to strike a balance between data privacy and resource utilization, ensuring that FL remains a viable solution in a rapidly evolving landscape of privacy and security concerns.

4 MFFER: Multimodal Federated-Learning based Facial Emotion Recognition

In this chapter, a description of the proposed system will be provided. The proposed system introduces the utilization of Federated Learning to train a multimodal facial emotion recognition system deployed in Kiosks to support a chatbot agent responsible for user interaction and recommendation system.

4.1 FLOWER Framework

In the development of the system, a rigorous evaluation of open-source frameworks, as outlined in section 3.1.1, was undertaken to determine the most suitable foundation. The FLOWER framework was identified as the optimal core component, primarily due to its proven scalability and robustness. The utilization of FLOWER's core functionalities, particularly its basic aggregation functions, was deemed essential for creating the global model and performing evaluation functions within the system.

However, it is imperative to note that the FLOWER framework presented certain challenges in achieving the intended scalability of the proposed system. One notable limitation was the strategy method employed for server setup. This method stipulates a predefined number of clients required to initiate the federated process. Consequently, this approach rendered it impossible to seamlessly introduce new devices into the environment without making modifications to the strategy. This constraint posed a hindrance to the system's adaptability and extensibility.

Additionally, a critical issue arose concerning the persistence of progress within the framework. Upon termination of the federated learning process, all progress would be lost unless the model was explicitly saved. This necessitated the reloading of the model in every iteration of the process, thereby diminishing the cost-effectiveness of an iterative learning approach.

Another noteworthy characteristic of the FLOWER framework is its reliance on communication between clients and the server via the GRPC protocol. However, the set of implemented functions within this protocol did not align with the desired communication approach envisioned for the proposed system. To achieve the desired efficiency and effectiveness of the system, it became imperative to explore alternative communication layer approaches, which ultimately shaped the core design of the proposed system.

In summary, the selection of the FLOWER framework as the core component for the system was driven by its scalability and aggregation functions. However, limitations related to adaptability, progress persistence, and communication protocols necessitated innovative

solutions and adjustments in the development process to ensure the system's alignment with its intended goals and performance requirements.

4.2 Proposed System Architecture

Figure 5 shows the proposed architecture combining traditional recommendation system models with a multimodal emotion detection model to gauge the user's emotional state during the interaction with the system. This allows for a constant understanding of the user's reactions to the recommendations provided by the system, which will be based on the user's purchases history or local store trends, allowing the virtual assistant to give answers and new recommendations based on the user's reactions. However, since the system deals with sensitive data, a strong robust method to ensure user's trust into the system is necessary. The use of Federated Learning ensures data privacy during the utilisation of the system, as well as the constant training processes to keep the models updated and effective. To achieve this, the system avoids transferring sensitive information, allowing training to occur locally, ensuring greater data protection and confidentiality.

The system consists of a global federated framework, which contains its own version of all the models used by the system and its own validation dataset. This is necessary to aggregate the knowledge provided by the individual deployed nodes and validate the results of the aggregation of the model weights. The local federated learning framework represents the local deployed nodal, which enables decentralised training of recommendation and multimodal emotion detection model. Local instances also include a chatbot within their intelligent layer, which is in contact with the user interface, helping to enhance user interaction and engagement.

The recommendation component relies on three main approaches: Collaborative Filtering, Content-Based Filtering, and Hybrid Algorithm. To ensure efficient performance, the architecture also integrates a data layer, where interactions between users and the virtual assistant, as well as product information, are stored. This structure ensures that the system has all the information necessary for personalised and optimised recommendations.

The virtual assistant interface allows for direct interaction with the user, providing personalised responses and recommendations. Communication between different components is managed through web services, ensuring efficient and secure data flow.

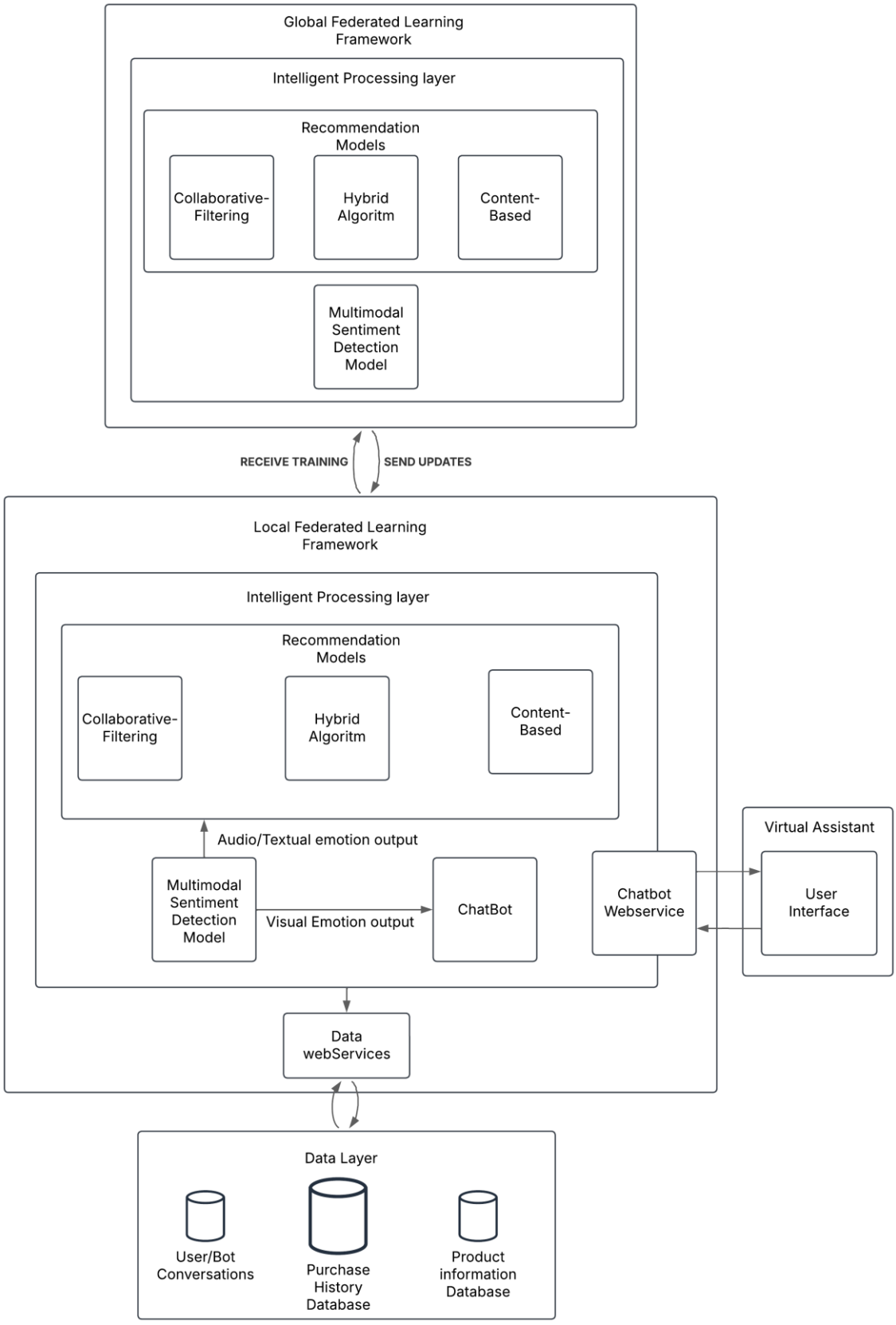

Figure 5 - MMFER Proposed Architecture

Principal Components

This section presents and details the various components that make up the architecture.

- **Global Federated Framework**: The Global Federated Framework functions as a knowledge aggregator, by averaging the weights of the locally trained models from the different nodes by performing regular updates selecting randomly selected nodes. This continuously improves the system's global performance, updating models with worse results, while ensuring that models with better performances are not overwritten with the global model. This is achieved without compromising user privacy, as sensitive data never leave the local device.
- **Local Federated Learning Framework**: Each local instance of the system trains the recommendation and sentiment detection models based on the stored data of user purchase history and store purchase history, making models prioritise user's preference and only then the purchase association that exists in the store. This federated architecture allows the local models to be constantly updated with lower training times, since all the data is stored locally. In this way, by using federated learning, the user experience can still be constantly improved while keeping data protected outside of global major updates.
- **Intelligent Processing Layer**: This layer represents the core of the intelligence of the system, including the different models that make up the system, such as the recommendation system model, the multimodal sentiment detection model, and it also contains the virtual assistant, consisting of the following modules:
    - **Recommendation Systems:**
        1. Collaborative Filtering: Analyses behaviour patterns among users to suggest relevant items based on similar preferences.
        2. Content-Based Filtering: Recommends products based on item characteristics and the user's individual preferences.
        3. Hybrid Algorithm: Combines different recommendation techniques to improve the accuracy of suggestions.
    - **Chatbot**: Enables users to interact with the system naturally using natural language. This chatbot being part of the intelligent layer, makes it capable of receiving information from the recommendation models and sentiment detection model, providing accurate recommendations based on user prompts, while being capable of reacting to emotional state
    - **Multimodal Sentiment Detection Model**: Analyse user emotions based on text, audio, and image, adjusting responses and recommendations based on their input and emotional state. The modal output will feed both the Chatbot and Recommendation systems, feeding the full output to the chatbot so it can produce the best answer, while only feeding the recommendation system with features specified by the user and the detected opinion and emotion.
    - **ChatbotWebService**: Facilitates communication between the user interface and internal services, ensuring an efficient and consistent information flow.

- **Data Layer**: The data layer is responsible for storing essential information for system operation. It includes three main databases:
    - **User/Bot Conversations**: Stores interaction history to enable continuous improvement of the user experience.
    - **Product Database**: Contains details on the products available for recommendation.
    - **Purchase History**: Contains records of user purchases and purchase association to enhance recommendation accuracy and personalization.
- **Virtual Assistant**: The virtual assistant is the interface through which users interact with the system. This component allows for the presentation of personalised recommendations, quick human-like responses, especially after detecting the user's feelings to the recommendations and queries, and a smoother interaction experience.

5 Experimentation

In this chapter, we delve into a comprehensive description of the experiments that have already been conducted. These experiments represent a significant milestone in the development of the proposed system, which was meticulously crafted during the CAPE project within GECAD. I must take this moment to extend my heartfelt gratitude once again for the support provided throughout the journey leading up to this dissertation.

The inception of this project was fuelled by a set of robust requirements, stemming from the need to work with precise and accurate information to provide users with the best experience. The aim was to devise a facial emotion recognition system that would support a recommendation system and a virtual assistant. By discerning the user emotions while they interact with the full system, the identified emotions would be used to tailor recommendations that matched the users' desires, and the virtual assistant would provide more thoughtful and delicate answers based on how users react and express themselves.

These experiments were meticulously designed and executed with a dual focus in mind. First and foremost, they aimed to meet the exacting requirements of the project, ensuring that the proposed system could seamlessly integrate into the sales assistance of a retail store and deliver tangible benefits. Secondly, they were driven by the pursuit of scientific innovation, particularly within the realm of Federated Learning. This dual commitment underscores not only the practical relevance of our work but also its contribution to the broader landscape of technological advancement.

As mentioned above, due to the hardship finding data, due to the required protection mechanisms related to sensitive data, all the models use the CREMA-D dataset for training

5.1 Experiment Report: Trial 1 — Centralized Multimodal Emotion Recognition on CREMA-D with a simplified model

This experiment aims to evaluate the performance of a prototype multimodal deep learning model for emotion recognition using the **CREMA-D** dataset. In this scenario, only the visual and audio features were taken into consideration when developing the multimodal system, with a late fusion being done as final layer to integrate the predictions into simplified outputs. The model jointly predicts:

- **Displayed emotion** (as expressed by the actor)
- **Perceived (response) emotion** (as interpreted by the listener)
- **Displayed and response emotion intensity levels** (regression)
- **Displayed valence** (positive/neutral/negative categories)

The preprocessing of data is as follows:

Total Clips Used: Filtered and aligned using clipName and HDF5 indices

Modality Inputs:

- **Audio**: MFCCs (shape: 100 × 13) normalized using training set mean and std.
- **Video**: 10 RGB frames per clip (224×224), pre-processed and normalized.

Labels:

- displayed_output: Categorical label (actor's displayed emotion)
- response_output: Categorical label (listener's perceived emotion)
- displayed_level: Float in [0, 1] (displayed emotion intensity)
- response_level: Float in [0, 1] (perceived emotion intensity)
- displayed_valence: Categorical label (valence of displayed emotion)

Data Split: 80% training / 20% testing (randomized with fixed seed)

The hardware utilised for the experiments were desktops, kindly provided by GECAD to perform these experiments. The process of visual data increases the necessary computational resources of the experiment.

Hardware:

- **CPU:** AMD Ryzen 9 7950X 16-Core Processor
- **GPU:** NVIDIA GeForce RTX 4080 (16 GB VRAM)
- **RAM:** 32 GB (2 x 16 GB) DDR5 5200 MHz
- **Hard Drive:** SSD 500 GB (181 GB free out of 487 GB)
- **OS:** Windows 11

5.1.1 Model Architecture

In this initial experiment, the model design focused only on the core multimodal fusion of audio and visual information. Audio features were extracted as MFCCs and processed through a 1D CNN, while video features were encoded via a frame-level CNN followed by a temporal LSTM. The two modalities were concatenated at a shared representation layer, from which multiple output heads performed emotion classification and intensity regression.

Table 3 - Model Components of the first experiment

Component	Description
Audio Branch	1D CNN + Global Average Pooling on MFCCs
Visual Branch	TimeDistributed 2D CNN → LSTM on video frames
Fusion	Concatenation of audio and video encodings
Shared Head	Dense + Dropout layer
Output Heads	Five parallel heads: classification × 3, regression × 2

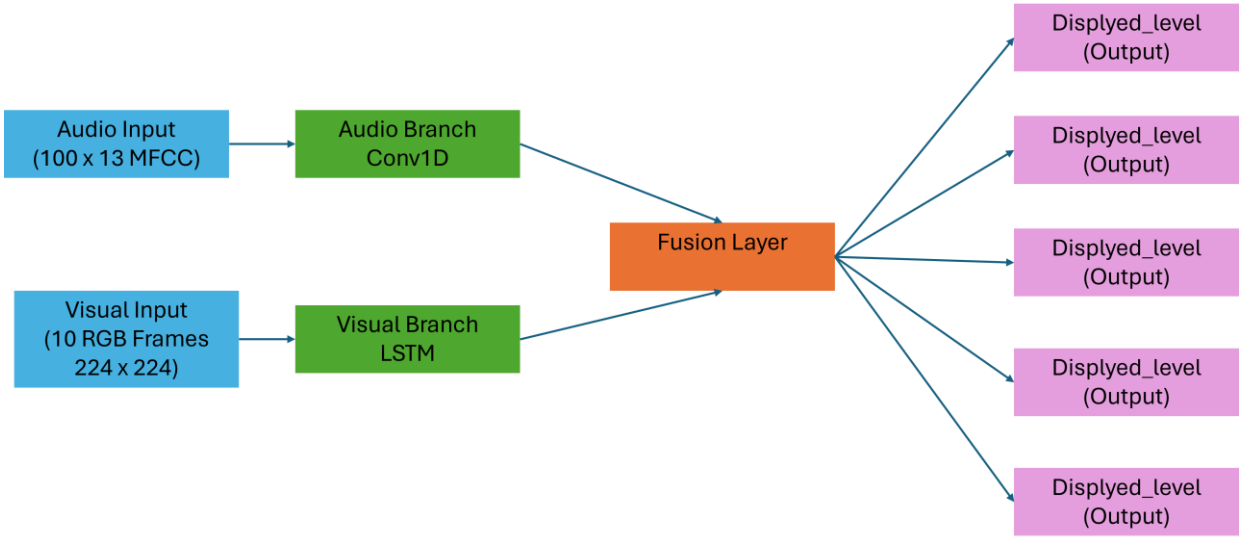

Figure 6 – Simplified flow of the First Experiment model

Loss Functions:

- Categorical Crossentropy (for classification tasks)
- Mean Squared Error (for regression tasks)

Loss Weights:

- response_output: 1.0
- displayed_output: 1.0

- response_level: 0.2
- displayed_level: 0.2
- displayed_valence: 1.0

Optimizer: Adam (learning rate = 1e-4)
Precision Policy: Mixed precision (float16) enabled on GPU
Training Duration: 10 epochs (roughly 1 minutes per epoch, for a total of 10 minutes)
Batch Size: 16

Total parameters: 12,962

Trainable parameters: 12,786

Non-trainable parameters: 176

5.1.2 Training and Evaluation Metrics

Table 4 - Classification results of the first experiment

Output	Training Accuracy	Validation Accuracy
displayed_output	68.8%	73.4%
response_output	49.6%	51.3%
displayed_valence	87.7%	89.5%

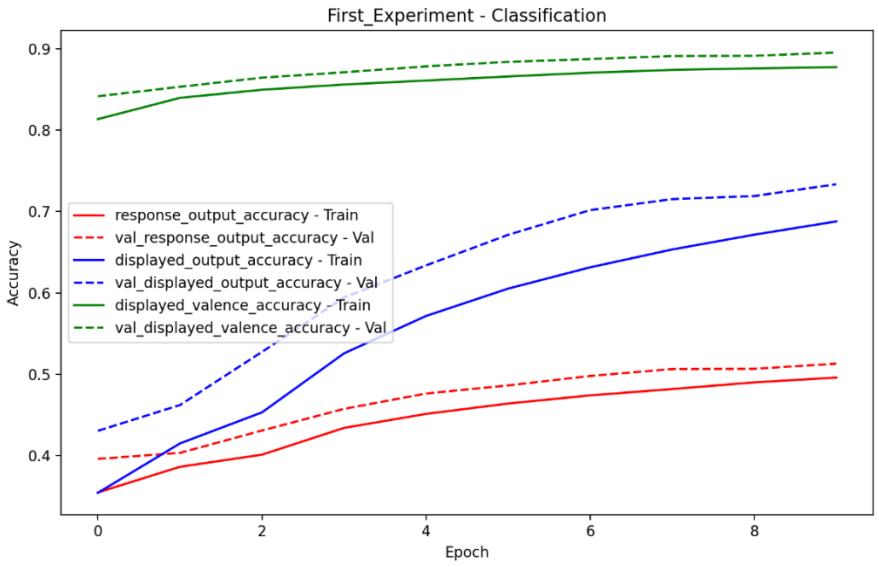

Figure 7 - Classification metrics evolution in the First Experiment

Table 5 - MAE results of the first experiment

Output	Training MAE	Validation MAE
displayed_level	0.0513	0.0474
response_level	0.2028	0.2034

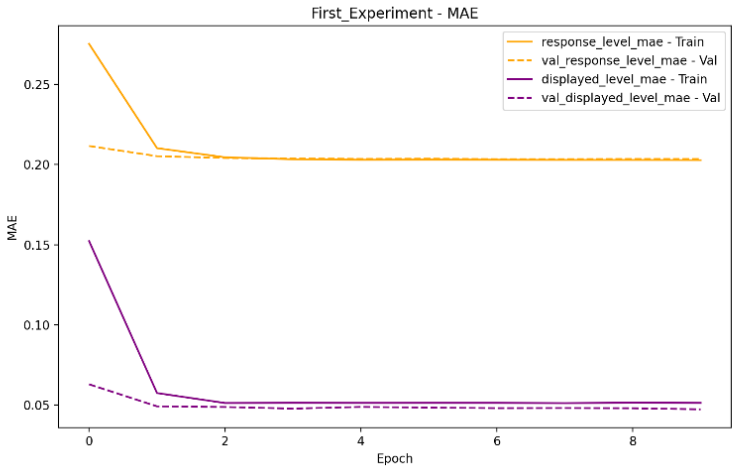

Figure 8 - MAE metrics evolution during the First Experiment

The model performs very well on tasks involving displayed emotion and valence, which aligns with the clarity of actor-performed expressions in CREMA-D.

Perceived emotions (response outputs) are harder to model, due to the subjective nature of listener interpretation. No signs of overfitting — validation accuracy and MAE remain close to training values. Regression heads for intensity (level) converge well, with displayed_level showing very low MAE.

5.2 Experiment Report: Trial 2 — Centralized Multimodal Emotion Recognition on CREMA-D with fusion methods

This experiment evaluates a multimodal deep learning architecture that integrates audio, text, video, and demographic features, with late fusion methods at key sections for joint emotion recognition on the CREMA-D dataset. The model simultaneously predicts:

- **Displayed emotion** (actor's expressed emotion)
- **Perceived (response) emotion** (listener's interpretation)
- **Displayed and response emotion intensity levels** (regression)
- **Displayed valence** (positive / neutral / negative categories)

The preprocessing of data is as follows:

Clips Used: Filtered and aligned using clipName and HDF5 indices

Modality Inputs:

- **Audio**: MFCCs, shape (100 × 13), normalized with training mean/std.
- **Text**: 50-dimensional embeddings, preprocessed and normalized.
- **Visual**: 10 RGB frames per clip, resized to 224 × 224, normalized.
- **Demographics**: Structured metadata features.

Labels:

- displayed_output: categorical (actor's displayed emotion)
- response_output: categorical (listener's perceived emotion)
- displayed_level: float in [0, 1] (displayed intensity)
- response_level: float in [0, 1] (perceived intensity)
- displayed_valence: categorical (positive/neutral/negative)

Data Split: 80% training / 20% testing (fixed seed).

The hardware utilised is the same desktop as in the previous experiment.

Hardware:

- **CPU:** AMD Ryzen 9 7950X 16-Core Processor
- **GPU:** NVIDIA GeForce RTX 4080 (16 GB VRAM)
- **RAM:** 32 GB (2 x 16 GB) DDR5 5200 MHz
- **Hard Drive:** SSD 500 GB (181 GB free out of 487 GB)
- **OS:** Windows 11

5.2.1 Model Architecture

Compared to the previous experiment, where we only had a fusion at visual and audio level, for this experiment, a new text and demographic branch were added. The text branch is compromised by a translation of the audio into text to have an independent analysis, while the demographics introduces meta data related to the actors of the dataset to attempt to generalise the contextualization of the model

Table 6 - Model Components of the Second experiment

Component	Description
Audio Branch	1D CNN + BatchNorm + Global Average Pooling + Dense(64) on MFCCs
Text Branch	Dense(64) + Dropout(0.2) on text embeddings
Demographic Branch	Dense(32) + Dropout(0.2)
Visual Branch	TimeDistributed 2D CNN → BiLSTM(32) → Dense(64) on video frames
Fusion	Hierarchical concatenation of audio–text–demo and visual–demo encodings
Shared Head	Dense(128 → 64) + Dropout
Output Heads	Five parallel heads: 3 × classification, 2 × regression

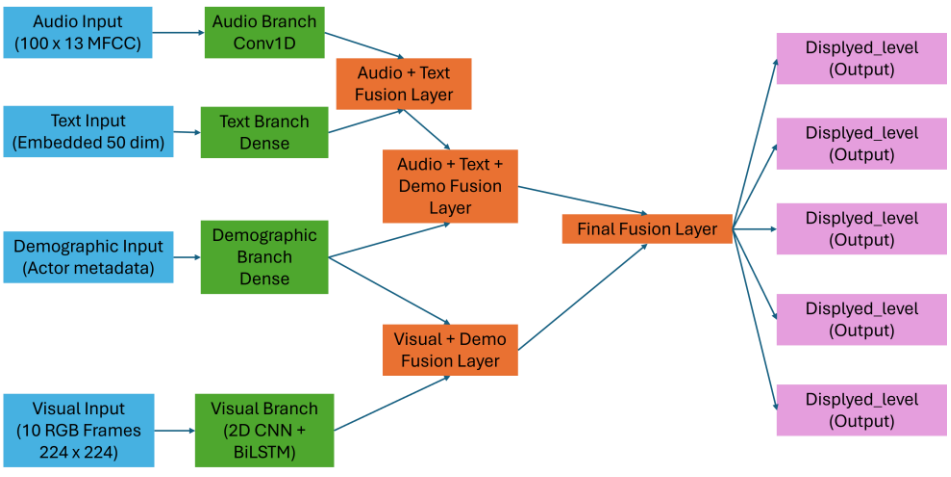

Figure 9 - Simplified flow of the Second Experiment model

Loss Functions:

- Categorical Crossentropy (classification)
- Mean Squared Error (regression)

Loss Weights:

- response_output: 1.0
- displayed_output: 1.0
- response_level: 0.2
- displayed_level: 0.2
- displayed_valence: 1.0

Optimizer: Adam (learning rate = 1e-4)
Precision Policy: Mixed precision (float16) on GPU
Batch Size: 16

Total parameters: 140,786

Trainable parameters: 140,434

Non-trainable parameters: 352

In contrast with the previous experiment, for this model, two different executions were run to better to understand the increase on the model's performance and resource exhaustion

5.2.2 Training and Evaluation Metrics

To compare the performance of this version of model, compared to the first one, the same number of epochs were executed to have a more linear degree of comparison.

Table 7 - Classification results of the Second Experiment

Output	Training Accuracy	Validation Accuracy
displayed_output	77.3%	79.9%
response_output	52.7%	53.6%
displayed_valence	91.2%	92.4%

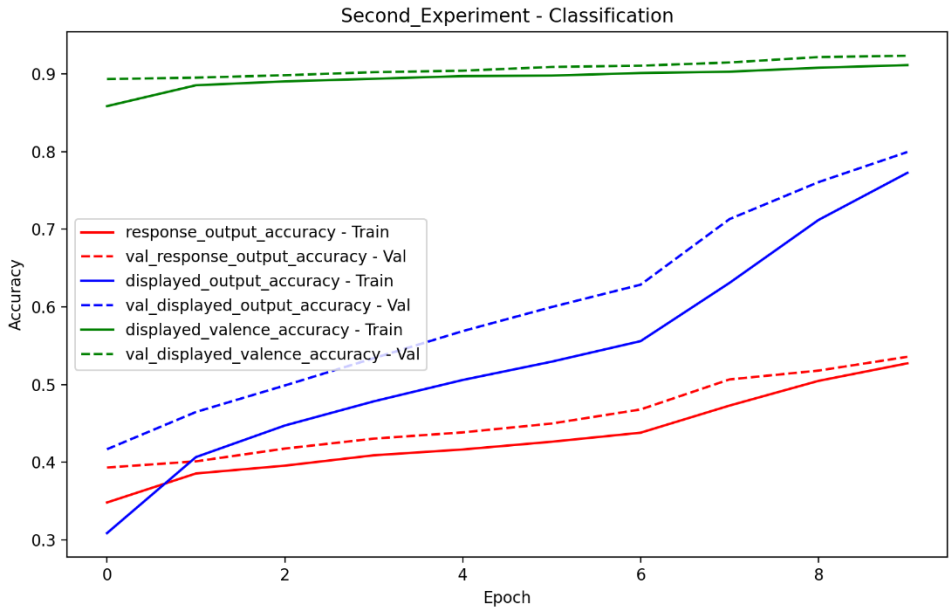

Figure 10 - Classification metrics evolution of the Second Experiment

Table 8 - MAE results of the Second Experiment

Output	Training MAE	Validation MAE
displayed_level	0.0449	0.0383
response_level	0.2025	0.2027

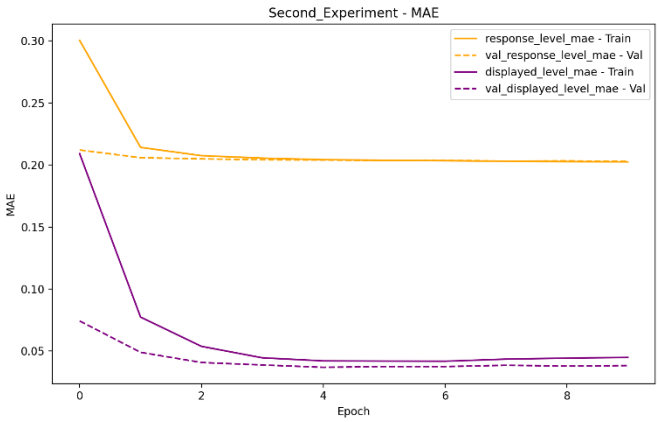

Figure 11 - MAE evolution in the Second Experiment

As we can observe, there was a significant increase in performance, across all metrics, achieving a displayed_valence accuracy over 92%. The only challenge remaining is related to the perceived, in this case, response_output, that is in the 53%. As such, it was decided to re-do this test, with a higher number of epochs to try to find the stabilising point of the model. However, the previous model trained in roughly a minute per round, while this model, being more robust, increased the training time to approximately 30 minutes per epoch, making harder to justify the utilization of this model considering the increase of performance doesn't truly justify such a heavy increase in runtime.

5.3 Experiment Report: Trial 3 — Centralized Multimodal Emotion Recognition on CREMA-D with fusion techniques and improved visual branch

Based on the perceived results of the previous experiment, a new trial was executed, applying changes to the visual branch to attempt to have increased performance in this modality, attempting to bring it closer to the other values. It follows the same structure as the previous trials:

- **Displayed emotion** (actor's expressed emotion)
- **Perceived (response) emotion** (listener's interpretation)
- **Displayed and response emotion intensity levels** (regression)
- **Displayed valence** (positive / neutral / negative categories)

The preprocessing of data is as follows:
Clips Used: Filtered and aligned using clipName and HDF5 indices

Modality Inputs:

- **Audio**: MFCCs, shape (100 × 13), normalized with training mean/std.
- **Text**: 50-dimensional embeddings, preprocessed and normalized.
- **Visual**: 10 RGB frames per clip, resized to 224 × 224, normalized.
- **Demographics**: Structured metadata features.

Labels:

- displayed_output: categorical (actor's displayed emotion)
- response_output: categorical (listener's perceived emotion)

- displayed_level: float in [0, 1] (displayed intensity)
- response_level: float in [0, 1] (perceived intensity)
- displayed_valence: categorical (positive/neutral/negative)

Data Split: 80% training / 20% testing (fixed seed).

The hardware utilised is the same desktop as in the previous experiment.

Hardware:

- **CPU:** AMD Ryzen 9 7950X 16-Core Processor
- **GPU:** NVIDIA GeForce RTX 4080 (16 GB VRAM)
- **RAM:** 32 GB (2 x 16 GB) DDR5 5200 MHz
- **Hard Drive:** SSD 500 GB (181 GB free out of 487 GB)
- **OS:** Windows 11

5.3.1 Model Architecture

As stated in the introduction, changes to the visual branch of the model were made to improve efficiency and performance. The stacked convolutional approach with the LSTM underperformed and was replaced by a pyramidical CNN-LSTM stack to compress spatial information hierarchically reducing overhead and redundancy. Fusion layers were also enhanced by applying gating mechanism to better evaluate the contribution the weights from the different modalities.

Table 9 - Model Components of the Third Experiment

Component	Description
Audio Branch	MFCC input (100×13) → Masking → 1D CNN (64 filters) → BatchNorm → Global Average Pooling → Dense(64, ReLU).
Text Branch	Precomputed text embeddings (50-dim) → Dense(64, ReLU) → Dropout(0.2).
Demographic Branch	Actor metadata (age/sex features) → Dense(32, ReLU) → Dropout(0.2).
Visual Branch	10 RGB frames (224×224) → TimeDistributed CNN (Conv2D ×3 with pooling & batchnorm, Global Avg Pool) → BiLSTM(64, bidirectional) → Dense(128, ReLU).

Component	Description
Gating Mechanism	Dense(1, sigmoid) applied to audio, text, and visual encodings, producing learnable importance weights for each modality.
Fusion 1: Audio-Text-Demo	Concatenation of gated audio, gated text, and demographics → Dense(128, ReLU) → Dropout(0.3).
Fusion 2: Visual-Demo	Concatenation of gated visual and demographics → Dense(128, ReLU) → Dropout(0.3).
Final Fusion	Concatenation of Fusion 1 and Fusion 2 → Dense(256, ReLU) → Dropout(0.4) → Dense(128, ReLU).
Output Heads	Five parallel heads: 3 × classification, 2 × regression

Loss Functions:

- Categorical Crossentropy (for classification tasks)
- Mean Squared Error (for regression tasks)

Loss Weights:

- response_output: 1.5
- displayed_output: 1.0
- response_level: 0.5
- displayed_level: 0.2
- displayed_valence: 1.0

Optimizer: Adam (learning rate = 1e-4)
Precision Policy: Mixed precision (float16) enabled on GPU
Training Duration: 15 epochs (roughly 20 minutes per epoch, for a total of 5 hours aproximately)
Batch Size: 16

Total parameters: 320,789

Trainable parameters: 320,213

Non-trainable parameters: 576´

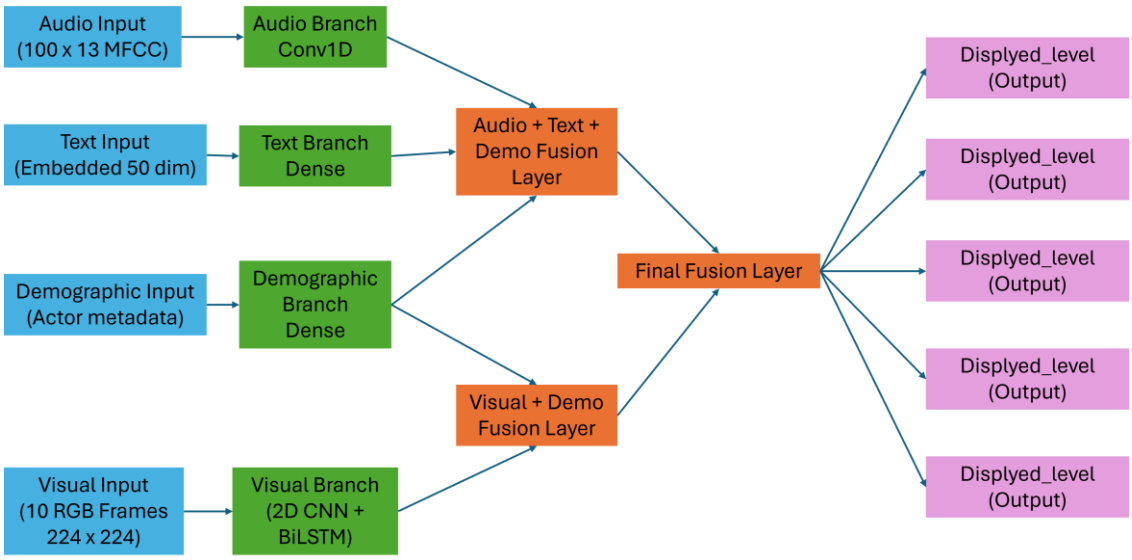

Figure 12 - Simplified flow of the Third Experiment mode

5.3.2 Training and Evaluation Metric

Table 10 - Classification results of the 3rd Experiment

Output	Training Accuracy	Validation Accuracy
Displayed Output	93.4%	95.9%
Response Output	57.1%	57.6%
Displayed Valence	93.6%	95.3%

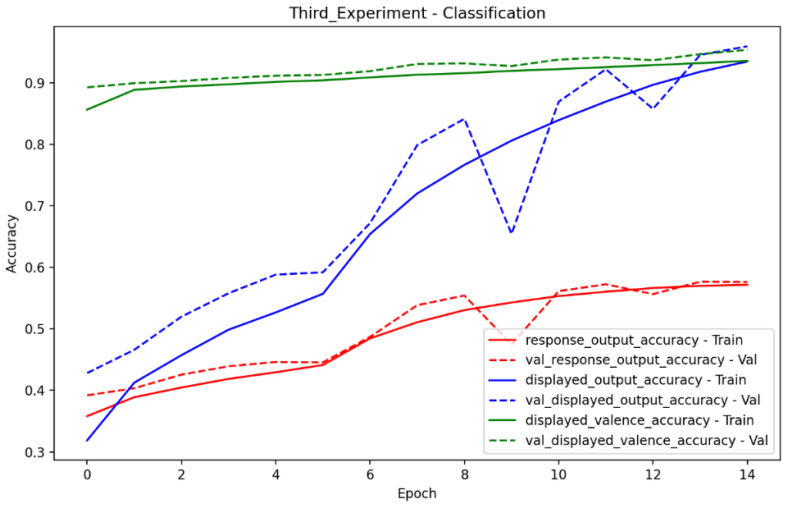

Figure 13 - Classification metrics evolution in the Third Experiment

Increasing the number of rounds and by adapting the model slightly to optimize, an increase in run time performance was noticed (a reduction of 10 minutes per round on the same setup as the previous experiment) as well as an increase in all metrics. However, the response output remains with a lower than desirable value, while showing some unstable fluctuations. During training it was noted that it stabilized around the 13th epoch, as the last epochs all had a similar result.

Table 11 - MAE results of the 3rd Experiment

Output	Training MAE	Validation MAE
Displayed Level	0.0431	0.0339
Response Level	0.2014	0.2015

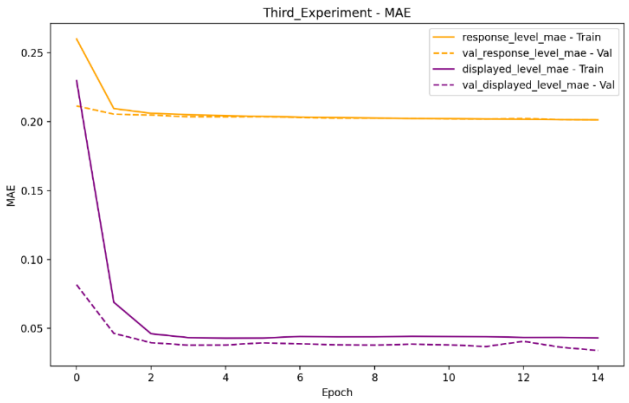

Figure 14 - MAE evolution in the Third Experiment

5.4 Experiment Report: Trial 4 — Federated Multimodal Emotion Recognition on CREMA-D

To conclude the development and trial of the system, the last developed prototype takes the model of the previous experimented and is integrated into the Federated framework. Data processing is slightly adjusted. To simulate the decentralised approach, clients access a dynamic portion of data, based on how many connect to the server. In the current example, since the experiment uses desktop, each client would access 50% of the total data, assigning via round-robin, so they can access similar data files. The remainder of features being evaluated and processed remains the same:

- **Displayed emotion** (actor's expressed emotion)
- **Perceived (response) emotion** (listener's interpretation)
- **Displayed and response emotion intensity levels** (regression)
- **Displayed valence** (positive / neutral / negative categories)

The preprocessing of data is as follows:
Clips Used: Filtered and aligned using clipName and HDF5 indices

Modality Inputs:

- **Audio**: MFCCs, shape (100 × 13), normalized with training mean/std.
- **Text**: 50-dimensional embeddings, preprocessed and normalized.
- **Visual**: 10 RGB frames per clip, resized to 224 × 224, normalized.

- **Demographics**: Structured metadata features.

Labels:

- displayed_output: categorical (actor's displayed emotion)
- response_output: categorical (listener's perceived emotion)
- displayed_level: float in [0, 1] (displayed intensity)
- response_level: float in [0, 1] (perceived intensity)
- displayed_valence: categorical (positive/neutral/negative)

Data Split: Linear equal split by clients where does a 80% training / 20% testing (fixed seed) internally for their own data.

The hardware utilised is the same desktop as in the previous experiment for client 1:

Hardware:

- **CPU:** AMD Ryzen 9 7950X 16-Core Processor
- **GPU:** NVIDIA GeForce RTX 4080 (16 GB VRAM)
- **RAM:** 32 GB (2 x 16 GB) DDR5 5200 MHz
- **Hard Drive:** SSD 500 GB (181 GB free out of 487 GB)
- **OS:** Windows 11

A second desktop is used as client 2. They share the same hardware however the disk space was more limited:

Hardware:

- **CPU:** AMD Ryzen 9 7950X 16-Core Processor
- **GPU:** NVIDIA GeForce RTX 4080 (16 GB VRAM)
- **RAM:** 32 GB (2 x 16 GB) DDR5 5200 MHz
- **Hard Drive:** SSD 500 GB (17,2 GB free out of 487 GB)
- **OS:** Windows 11

The server was deployed in a virtual machine, using a minimal setup due to the lack of available resources

Hardware:

- **CPU:** Intel® Xeon® Silver 4210 CPU @ 2.20 GHz
- **RAM:** 32 GB

- **Hard Drive:** SSD 100 GB (38,7 GB free out of 99,3 GB)
- **OS:** Windows Server

5.4.1 Model Architecture

As mentioned on the experiment description, the federated experiment was the adaptation of the centralised model of the third experiment into the federated framework, as such, the components described in **Erro! A origem da referência não foi encontrada.** and plot of Figure 12 remain the same in both clients during this experiment. The main notable difference compared to the centralised experiment, is that the training was faster. Client 1 averaged on 15 minutes per epoch, while client 2 averaged on 17 minutes per epoch. This could be explained by disk accesses to during the training and client 2 having less accessible disk space, could have caused some I/O access bottleneck

5.4.2 Training and Evaluation Metrics

As expected due to the split of data, the clients displayed different metric progression, however, the results were similar between both clients, keeping in pair with the centralised results. The results being similar to the centralised approach (with a slight drop of performance) which is expected in the federated framework, due to the splitting of data and weights averaging, the reduction of global training time shows that there's potential on the framework

Table 12 - Classification results of Client 1 during Federated Experiment

Output	Training Accuracy	Validation Accuracy
Displayed Output	82.2%	81.4%
Response Output	54.7%	54.2%
Displayed Valence	91.2%	91.9%

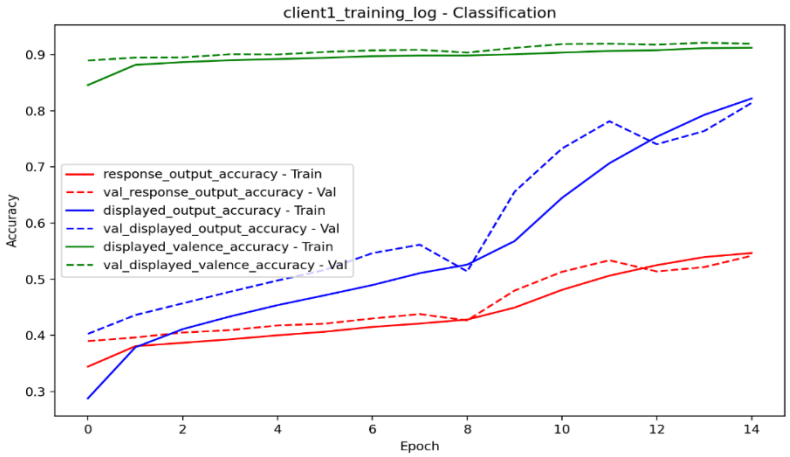

Figure 15 - Classification metrics evolution of Client 1 during Federated Experiment

Table 13 - MAE results of Client 1 during Federated Experiment

Output	Training MAE	Validation MAE
Displayed Level	0.0455	0.0414
Response Level	0.2033	0.2036

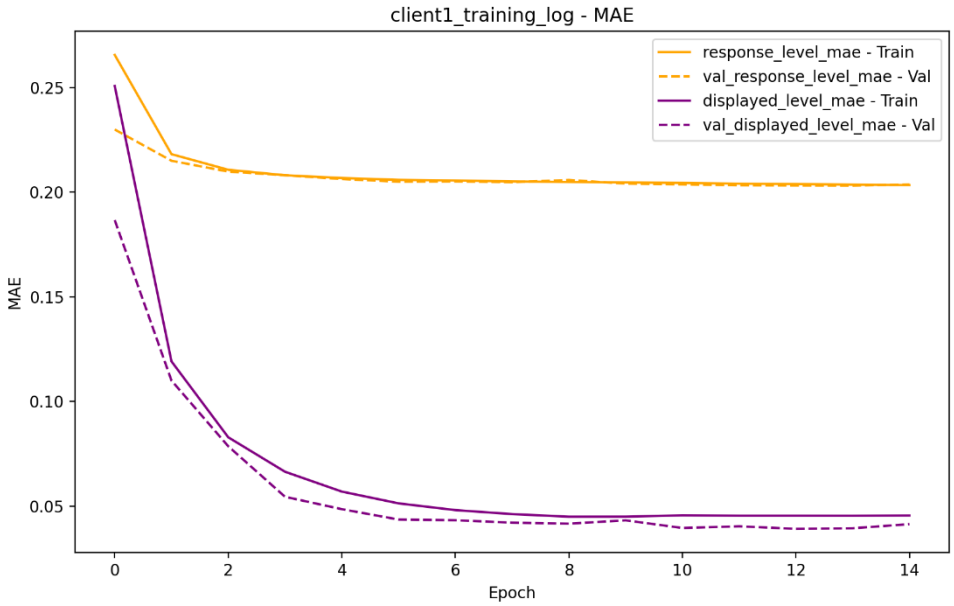

Figure 16 - MAE metrics evolution of Client 1 during the Federated Experiment

Table 14 - Classification results of Client 2 during Federated Experiment

Output	Training Accuracy	Validation Accuracy
Displayed Output	82.3%	81.8%
Response Output	54.8%	55.3%
Displayed Valence	91.4%	91.9%

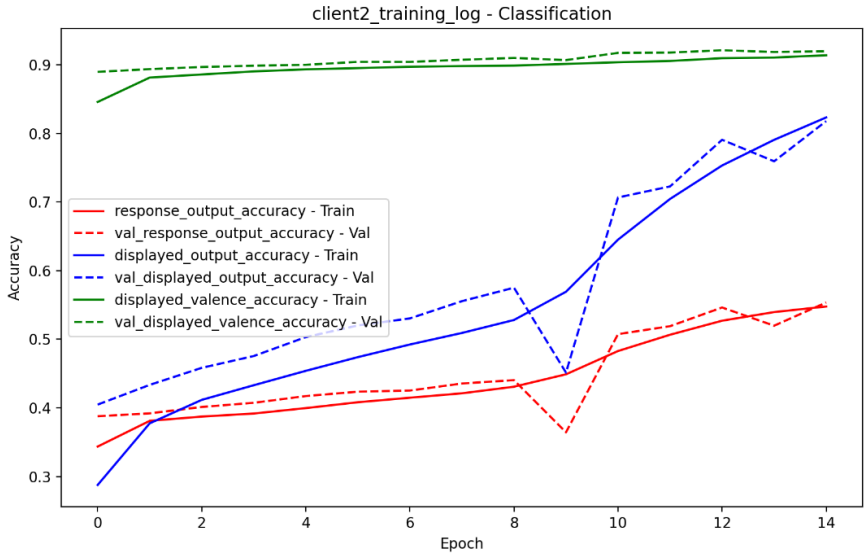

Figure 17 - Classification metrics evolution of Client 2 during Federated Experiment

Table 15 - MAE results of Client 2 during Federated Experiment

Output	Training MAE	Validation MAE
Displayed Level	0.0450	0.0412
Response Level	0.2031	0.2026

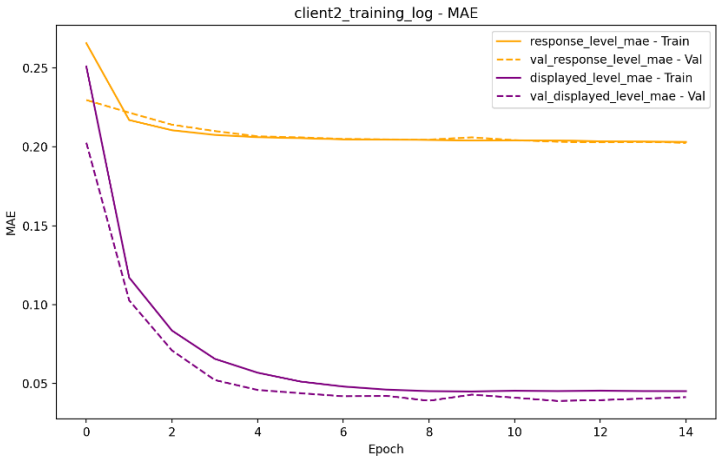

Figure 18 - MAE metrics evolution of Client 2 during Federated Experiment

5.5 Comparison with State-of-the-Art works

As mentioned in the previous chapters, the CREMA-D dataset that was used for the testing of the development of the system, is widely used standard benchmark in the field, having been used in several works to report the classification accuracy on displayed emotions. As such to display the results of this dissertation, a comparison between the achieved results with the literature will further display the capabilities of the system.

Most works in the literally only focus on the evaluation of displayed emotions, which in the case of the works is labelled as the "displayed_output" as mentioned in the experiments, and as such, this will be the metric used to compare with the literature.

Table 16 - Results Comparison with State-of-the-Art works

Method / Work	**Modalities Used**	**Task Type**	**Reported Performance (CREMA-D)**	**Notes**
This work – FL Clients 1 & 2	Audio + Video + Text + Demographics	Multi-task classification + regression	Displayed Output Accuracy: Client 1 = 81.8%, Client 2 = 81.7%	Federated, client-specific
(Salas-Cáceres *et al.*, 2025)	Audio + Video	Classification	Accuracy = 81.4%	Fusion + LSTM
(Mocanu, Tapu and Zaharia, 2023)	Audio + Video	Classification	Accuracy = 76.3%	Cross-modal audio-visual fusion

Method / Work	Modalities Used	Task Type	Reported Performance (CREMA-D)	Notes
(Islam, Karray and Muhammad, 2025)	Audio + Video	Classification	Accuracy = 84.6%	Multi-stage fusion network
(Radoi and Cioroiu, 2024)	Audio + Video	Classification	Accuracy = 74.2%	Lightweight model with uncertainty-based learning

As shown in Table 16, it is showcased that the models developed in this work achieve competitive performance on the CREMA-D dataset. For instance, FL Clients 1 and 2 reach displayed output accuracies of 81.8% and 81.7%, respectively, surpassing several recent audio/video-based approaches such as (Salas-Cáceres *et al.*, 2025), (Mocanu, Tapu and Zaharia, 2023) and (Radoi and Cioroiu, 2024), which report accuracies of 81.4%, 76.3% and 74.2%, respectively.

While most prior studies focus exclusively on displayed emotion classification, our approach extends the evaluation to multi-task predictions, including displayed valence, perceived response, and intensity regression. Notably, the ability to capture both categorical and continuous emotion representations demonstrates the flexibility of the proposed system and highlights the benefits of incorporating multimodal inputs (audio, video, text, and demographic data).

Additionally, the client-specific evaluation in a federated learning setting indicates that each client can maintain high performance locally while contributing to the broader model without sharing sensitive data. This approach contrasts with traditional centralized SoTA models, providing not only competitive accuracy but also enhanced privacy and adaptability across heterogeneous client data.

Overall, these results confirm that the proposed system is not only competitive with modern state-of-the-art works on CREMA-D but also introduces unique capabilities in multi-task and privacy-aware emotion recognition.

5.6 Chapter conclusions

During the development of the multiple prototypes, we could verify some progress on the development of the framework establishing a good basis for the project architecture. The gradual increase in the performance of the model with the performed changes was tangible but still leaving room for improvements.

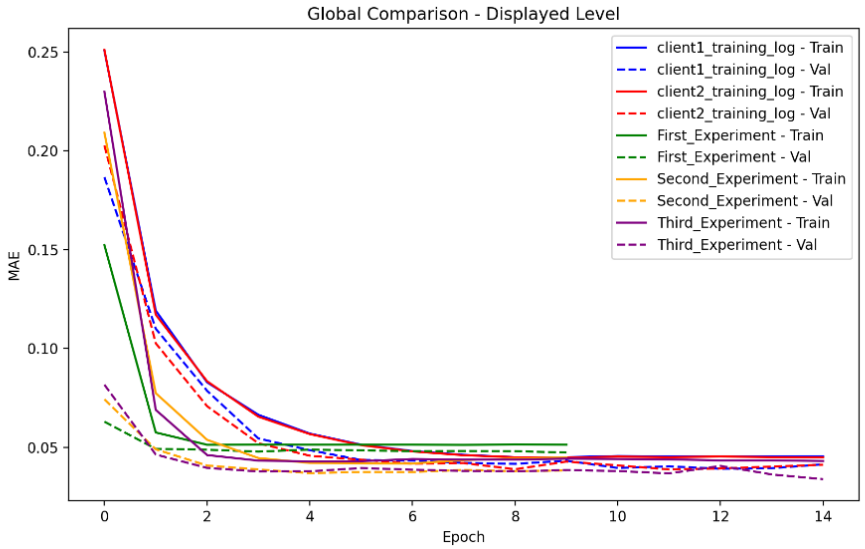

Figure 19 - Global metrics of displayed level MAE

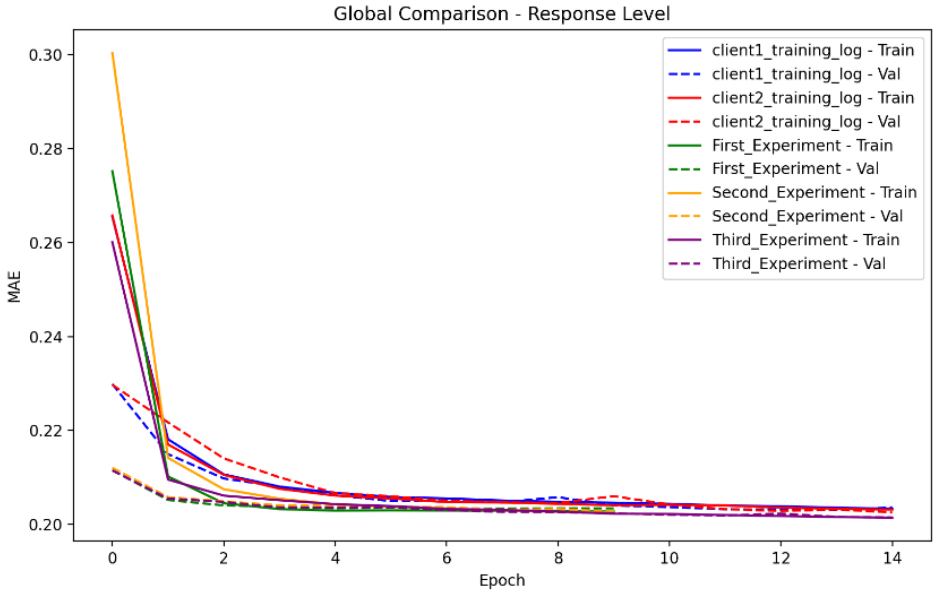

Figure 20 - Global metrics of response level MAE

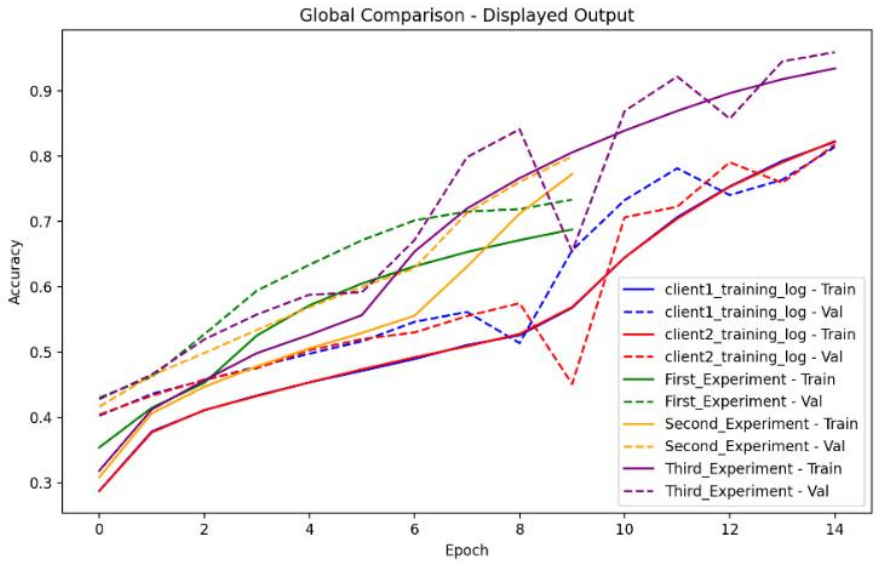

Figure 21 - Global metrics of displayed output accuracy

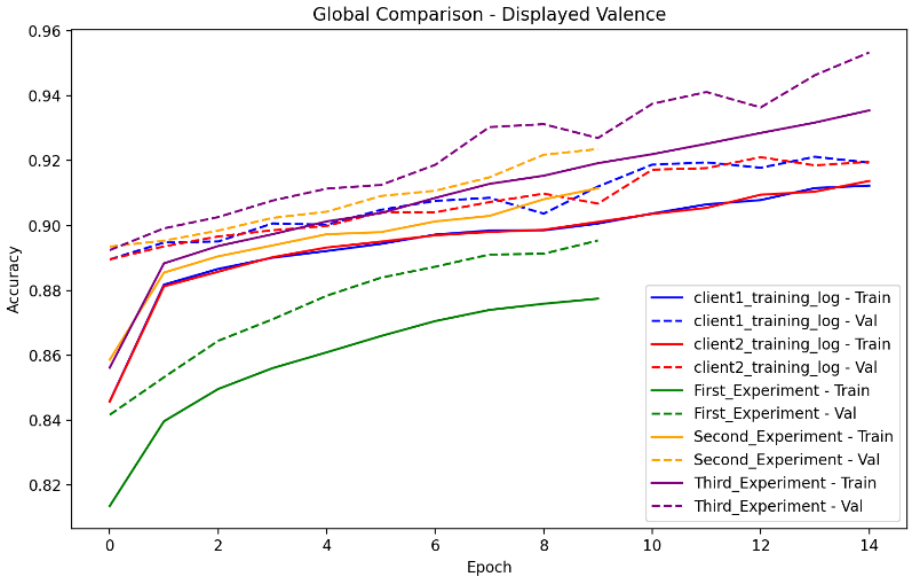

Figure 22 - Global metrics of displayed valence accuracy

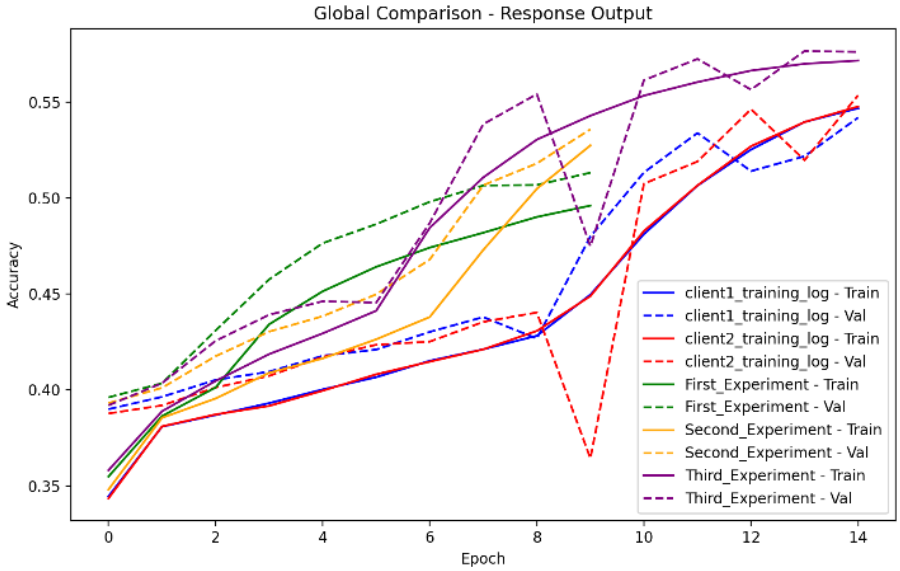

Figure 23 - Global metrics of response output accuracy

Observing the progress of the multiple evaluated metrics across the experiments, the evolution in the centralised experiments is noticeable, especially visible across the accuracy metrics, while the regression metrics tend to stabilise early enough across all experiments, fluctuating in very small changes, almost unnoticeable if we reduced the scale of accuracy. The increase of 10 epochs to 15 epochs also shows the potential growth of the model. Given the current evolution, displayed valence accuracy and displayed output seem to be stabilising but response output still has some growth room, a possible experiment with a higher number of epochs could get closer to the actual metrics ceiling. The regression metrics seem to achieve their ceiling fairly early on, leaving room low room for improvement with the current model.

6 Conclusions

6.1 Main conclusions

The presented CAPE framework demonstrates a promising approach for multimodal emotion recognition in a federated learning setting. While certain evaluation metrics, such as the response_output, still show lower-than-desired performance, the system achieves competitive results in other key metrics, particularly displayed_output and displayed_valence. For instance, both FL Client 1 and Client 2 reached displayed output accuracies above 81%, which surpasses several recent state-of-the-art multimodal approaches evaluated on the CREMA-D dataset.

From a scientific perspective, this work provides a concrete proof-of-concept that a federated deployment does not significantly compromise model performance compared to centralized environments. The system successfully preserves the privacy of sensitive user data while maintaining high classification accuracy, demonstrating the feasibility and advantages of a privacy-aware federated approach.

The integration of multiple modalities—audio, video, text, and demographic information—was shown to contribute positively to performance, highlighting the importance of multimodal fusion in emotion recognition tasks. This finding supports the notion that combining diverse sources of information can improve the robustness and generalization of emotion prediction models.

Although some research questions related to RQ1, specifically the integration with NLP models for chatbot interaction or with recommendation systems, remain open, the work thoroughly addresses RQ2. It confirms that federated learning can be effectively applied to multimodal emotion recognition without a significant trade-off in performance. Moreover, the architecture, supported by a scientific publication, positions the CAPE framework as a solid foundation for further research and development.

6.2 Future Work

Several directions remain for the continued improvement and application of the CAPE framework:

1. **Optimization of Emotion Recognition Models:** While the multi-task approach provides valuable insights into both classification and regression of emotional states, metrics such as response_output accuracy remain below the desired threshold. Future work should focus on refining model architectures and exploring advanced fusion strategies to enhance performance further.

2. **Computational Efficiency:** The current framework presents challenges in terms of runtime and resource requirements. Future development will aim to optimize models for deployment on lower-cost hardware, enabling broader accessibility and scalability.

3. **Extended Application and Integration:** The current work has not fully explored integration with NLP-based systems or recommendation engines. Future studies could incorporate CAPE into interactive systems, such as chatbots or adaptive interfaces, to evaluate real-world utility.

4. **Scientific Dissemination and Collaboration:** The results achieved provide a strong foundation for publication in scientific journals and conferences. Future work will aim to document and present the findings to the research community, fostering collaboration and contributing to advancements in federated, multimodal emotion recognition.

5. **Exploration of Novel Metrics:** Further research should also investigate additional evaluation metrics that capture nuanced aspects of perceived emotion, extending the framework beyond standard classification accuracy and regression measures.

The last step being left open is the adaptation of the work presented in this dissertation to be organised and presented on a scientific magazine, displaying the progress of the development of the project.

7 References

Abbruzzese, L. *et al.* (2019) 'Age and gender differences in emotion recognition', *Frontiers in Psychology*, 10(OCT), p. 479529. Available at: https://doi.org/10.3389/FPSYG.2019.02371/BIBTEX.

Acoustic Theory of Speech Production: With Calculations Based on X-Ray ... - Gunnar Fant - Google Livros (no date). Available at: https://books.google.pt/books?hl=pt-PT&lr=&id=qa-AUPdWg6sC&oi=fnd&pg=PA5&dq=Fant,+G.+(1960).+Acoustic+theory+of+speech+production&ots=repT2MD7u0&sig=-uq-0IV4zWs01nAXTJJZWInsMfw&redir_esc=y#v=onepage&q=Fant%2C%20G.%20(1960).%20Acoustic%20theory%20of%20speech%20production&f=false (Accessed: 1 September 2025).

Adamopoulou, E. and Moussiades, L. (2020) 'Chatbots: History, technology, and applications', *Machine Learning with Applications*, 2, p. 100006. Available at: https://doi.org/10.1016/J.MLWA.2020.100006.

Aguilera, A., Mellado, D. and Rojas, F. (2023a) 'An Assessment of In-the-Wild Datasets for Multimodal Emotion Recognition', *Sensors 2023, Vol. 23, Page 5184*, 23(11), p. 5184. Available at: https://doi.org/10.3390/S23115184.

Aguilera, A., Mellado, D. and Rojas, F. (2023b) 'An Assessment of In-the-Wild Datasets for Multimodal Emotion Recognition', *Sensors 2023, Vol. 23, Page 5184*, 23(11), p. 5184. Available at: https://doi.org/10.3390/S23115184.

Alarcão, S.M. and Fonseca, M.J. (2019) 'Emotions recognition using EEG signals: A survey', *IEEE Transactions on Affective Computing*, 10(3), pp. 374–393. Available at: https://doi.org/10.1109/TAFFC.2017.2714671.

Andreina, S. *et al.* (2021) 'BaFFLe: Backdoor detection via feedback-based federated learning', *Proceedings - International Conference on Distributed Computing Systems*, 2021-July, pp. 852–863. Available at: https://doi.org/10.1109/ICDCS51616.2021.00086.

Arya, R., Singh, J. and Kumar, A. (2021) 'A survey of multidisciplinary domains contributing to affective computing', *Computer Science Review*, 40, p. 100399. Available at: https://doi.org/10.1016/J.COSREV.2021.100399.

Atal, B.S. and Hanauer, S.L. (1971) 'Speech Analysis and Synthesis by Linear Prediction of the Speech Wave', *The Journal of the Acoustical Society of America*, 50(2B), pp. 637–655. Available at: https://doi.org/10.1121/1.1912679.

Atrey, P.K. *et al.* (2010) 'Multimodal fusion for multimedia analysis: A survey', *Multimedia Systems*, 16(6), pp. 345–379. Available at: https://doi.org/10.1007/S00530-010-0182-0/TABLES/9.

Baltrusaitis, T., Ahuja, C. and Morency, L.P. (2019) 'Multimodal Machine Learning: A Survey and Taxonomy', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), pp. 423–443. Available at: https://doi.org/10.1109/TPAMI.2018.2798607.

Bartlett, M.S. *et al.* (2005) 'Recognizing facial expression: Machine learning and application to spontaneous behavior', *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, II, pp. 568–573. Available at: https://doi.org/10.1109/CVPR.2005.297.

Berntson, G.G. *et al.* (1997) 'Heart rate variability: Origins, methods, and interpretive caveats', *Psychophysiology*, 34(6), pp. 623–648. Available at: https://doi.org/10.1111/J.1469-8986.1997.TB02140.X.

Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5–32. Available at: https://doi.org/10.1023/A:1010933404324/METRICS.

Brendan McMahan, H. *et al.* (2016) 'Communication-Efficient Learning of Deep Networks from Decentralized Data', *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017* [Preprint]. Available at: https://doi.org/10.48550/arxiv.1602.05629.

Burke, R. (2002) 'Hybrid recommender systems: Survey and experiments', *User Modelling and User-Adapted Interaction*, 12(4), pp. 331–370. Available at: https://doi.org/10.1023/A:1021240730564/METRICS.

Cacioppo, J.T. *et al.* (1986) 'Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions.', *Journal of personality and social psychology*, 50(2), p. 260.

Cao, H. *et al.* (2014) 'CREMA-D: Crowd-sourced emotional multimodal actors dataset', *IEEE Transactions on Affective Computing*, 5(4), pp. 377–390. Available at: https://doi.org/10.1109/TAFFC.2014.2336244.

Cheng, Z. *et al.* (2024) 'Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning', *Advances in Neural Information Processing Systems*, 37. Available at: https://arxiv.org/abs/2406.11161v2 (Accessed: 12 September 2025).

Cho, Y.J., Wang, J. and Joshi, G. (2022) 'Towards Understanding Biased Client Selection in Federated Learning'. PMLR, pp. 10351–10375. Available at: https://proceedings.mlr.press/v151/jee-cho22a.html (Accessed: 10 January 2023).

Cootes, T.F. *et al.* (1995) 'Active Shape Models-Their Training and Application', *Computer Vision and Image Understanding*, 61(1), pp. 38–59. Available at: https://doi.org/10.1006/CVIU.1995.1004.

Cortes, C., Vapnik, V. and Saitta, L. (1995) 'Support-vector networks', *Machine Learning 1995 20:3*, 20(3), pp. 273–297. Available at: https://doi.org/10.1007/BF00994018.

Cowen, A.S. and Keltner, D. (2021) 'Semantic Space Theory: A Computational Approach to Emotion', *Trends in Cognitive Sciences*, 25(2), pp. 124–136. Available at: https://doi.org/10.1016/J.TICS.2020.11.004/ATTACHMENT/CBABCF23-C7CF-4AF1-9BA1-5C1E581FA934/MMC1.MP4.

Davidson, R.J. (1992) 'Anterior cerebral asymmetry and the nature of emotion', *Brain and Cognition*, 20(1), pp. 125–151. Available at: https://doi.org/10.1016/0278-2626(92)90065-T.

Davis, S.B. and Mermelstein, P. (1980) 'Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), pp. 357–366. Available at: https://doi.org/10.1109/TASSP.1980.1163420.

Dawson, M.E., Schell, A.M. and Filion, D.L. (2007) 'The electrodermal system', *Handbook of psychophysiology*, 2, pp. 200–223.

D'Mello, S.K. and Graesser, A.C. (2014) '31 Feeling, Thinking, and Computing with Affect-Aware Learning', *The Oxford handbook of affective computing*, p. 419.

Dominguez-Catena, I., Paternain, D. and Galar, M. (2024) 'Metrics for Dataset Demographic Bias: A Case Study on Facial Expression Recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8), pp. 5209–5226. Available at: https://doi.org/10.1109/TPAMI.2024.3361979.

Ekman, P. (2004) 'Emotions revealed', *BMJ*, 328(Suppl S5), p. 0405184. Available at: https://doi.org/10.1136/SBMJ.0405184.

Ekman, P. and Friesen, W. V. (2019) 'Facial Action Coding System', *PsycTESTS Dataset* [Preprint]. Available at: https://doi.org/10.1037/T27734-000.

Elman, J.L. (1990) 'Finding Structure in Time', *Cognitive Science*, 14(2), pp. 179–211. Available at: https://doi.org/10.1207/S15516709COG1402_1.

EUR-Lex - 32016R0679 - EN - EUR-Lex (no date). Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj (Accessed: 16 December 2022).

Ezzeldin, Y.H. *et al.* (2021) 'FairFed: Enabling Group Fairness in Federated Learning'. Available at: https://doi.org/10.48550/arxiv.2110.00857.

Feng, S. (2022) 'Vertical federated learning-based feature selection with non-overlapping sample utilization', *Expert Systems with Applications*, 208, p. 118097. Available at: https://doi.org/10.1016/J.ESWA.2022.118097.

Fred Agarap, A.M. (2018) 'Deep Learning using Rectified Linear Units (ReLU)'. Available at: https://doi.org/10.48550/arxiv.1803.08375.

Gahlan, N. and Sethia, D. (2024a) 'AFLEMP: Attention-based Federated Learning for Emotion recognition using Multi-modal Physiological data', *Biomedical Signal Processing and Control*, 94, p. 106353. Available at: https://doi.org/10.1016/J.BSPC.2024.106353.

Gahlan, N. and Sethia, D. (2024b) 'Federated learning in Emotion Recognition Systems based on physiological signals for privacy preservation: a review', *Multimedia Tools and Applications 2024 84:13*, 84(13), pp. 12417–12485. Available at: https://doi.org/10.1007/S11042-024-19467-3.

Ganin, Y. and Lempitsky, V. (2014) 'Unsupervised Domain Adaptation by Backpropagation', *32nd International Conference on Machine Learning, ICML 2015*, 2, pp. 1180–1189. Available at: https://arxiv.org/abs/1409.7495v2 (Accessed: 7 September 2025).

Goodfellow, I.J. *et al.* (2013) 'Challenges in Representation Learning: A Report on Three Machine Learning Contests', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8228 LNCS(PART 3), pp. 117–124. Available at: https://doi.org/10.1007/978-3-642-42051-1_16.

Herlocker, J.L. *et al.* (2004) 'Evaluating collaborative filtering recommender systems', *ACM Transactions on Information Systems (TOIS)*, 22(1), pp. 5–53. Available at: https://doi.org/10.1145/963770.963772.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long Short-Term Memory', *Neural Computation*, 9(8), pp. 1735–1780. Available at: https://doi.org/10.1162/NECO.1997.9.8.1735.

Huang, M. *et al.* (2020) 'A Federated Multi-View Deep Learning Framework for Privacy-Preserving Recommendations'. Available at: https://arxiv.org/abs/2008.10808v1 (Accessed: 16 September 2025).

Islam, M.M., Karray, F. and Muhammad, G. (2025) 'MSF-Net: Multi-stage fusion network for emotion recognition from multimodal signals in scalable healthcare', *Information Fusion*, 119, p. 103028. Available at: https://doi.org/10.1016/J.INFFUS.2025.103028.

Kansizoglou, I. *et al.* (2022) 'Continuous Emotion Recognition for Long-Term Behavior Modeling through Recurrent Neural Networks', *Technologies 2022, Vol. 10, Page 59*, 10(3), p. 59. Available at: https://doi.org/10.3390/TECHNOLOGIES10030059.

Kim, D.Y. and Wallraven, C. (2021) 'Label quality in AffectNet: results of crowd-based re-annotation', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13189 LNCS, pp. 518–531. Available at: https://doi.org/10.1007/978-3-031-02444-3_39.

Kumari, K. *et al.* (2022) 'BayBFed: Bayesian Backdoor Defense for Federated Learning', *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1747–1764. Available at: https://doi.org/10.1109/SP46215.2023.00100.

LeCun, Y. *et al.* (1998) 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE*, 86(11), pp. 2278–2323. Available at: https://doi.org/10.1109/5.726791.

Lee, J.P. *et al.* (2024) 'Encoding of multi-modal emotional information via personalized skin-integrated wireless facial interface', *Nature Communications 2024 15:1*, 15(1), pp. 1–13. Available at: https://doi.org/10.1038/s41467-023-44673-2.

Li, L. *et al.* (2020) 'A review of applications in federated learning', *Computers & Industrial Engineering*, 149, p. 106854. Available at: https://doi.org/10.1016/J.CIE.2020.106854.

Li, S., Deng, W. and Du, J. (2017) 'Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild', pp. 2852–2861. Available at: http://whdeng.cn/RAF/model1.html (Accessed: 2 September 2025).

Li, S. and Tang, H. (2024) 'Multimodal Alignment and Fusion: A Survey'. Available at: https://arxiv.org/abs/2411.17040v1 (Accessed: 10 September 2025).

Liang, J. *et al.* (2019) 'Cross-culture Multimodal Emotion Recognition with Adversarial Learning', *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2019-May, pp. 4000–4004. Available at: https://doi.org/10.1109/ICASSP.2019.8683725.

Lim, W.Y.B. *et al.* (2022) 'Decentralized Edge Intelligence: A Dynamic Resource Allocation Framework for Hierarchical Federated Learning', *IEEE Transactions on Parallel and Distributed Systems*, 33(3), pp. 536–550. Available at: https://doi.org/10.1109/TPDS.2021.3096076.

Lin, W. *et al.* (2023) 'A federated collaborative recommendation model for privacy-preserving distributed recommender applications based on microservice framework', *Journal of Parallel and Distributed Computing*, 174, pp. 70–80. Available at: https://doi.org/10.1016/J.JPDC.2022.12.002.

Liu, R. *et al.* (2021) 'FLAME: Differentially Private Federated Learning in the Shuffle Model', *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10), pp. 8688–8696. Available at: https://doi.org/10.1609/AAAI.V35I10.17053.

Liu, Y. *et al.* (2018) 'Secure Federated Transfer Learning', *IEEE Intelligent Systems*, 35(4), pp. 70–82. Available at: https://doi.org/10.1109/MIS.2020.2988525.

Lucey, P. *et al.* (2010) 'The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression', *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, pp. 94–101. Available at: https://doi.org/10.1109/CVPRW.2010.5543262.

McMahan, B. *et al.* (2017) 'Communication-Efficient Learning of Deep Networks from Decentralized Data'. PMLR, pp. 1273–1282. Available at: https://proceedings.mlr.press/v54/mcmahan17a.html (Accessed: 17 September 2025).

McStay, A. (2020) ‘Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy’, *Big Data & Society*, 7(1). Available at: https://doi.org/10.1177/2053951720904386.

McTear, M.F. (2002) ‘Spoken dialogue technology’, *ACM Computing Surveys (CSUR)*, 34(1), pp. 90–169. Available at: https://doi.org/10.1145/505282.505285.

Mehrabian, A. (1996) ‘Pleasure-Arousal-Dominance: A general framework for describing and measuring individual differences in temperament’, *Current Psychology*, 14(4), pp. 261–292. Available at: https://doi.org/10.1007/BF02686918/METRICS.

Mendieta, M. *et al.* (2022) ‘Local Learning Matters: Rethinking Data Heterogeneity in Federated Learning’, pp. 8397–8406. Available at: https://github.com/mmendiet/FedAlign. (Accessed: 19 December 2022).

Mitra, V. *et al.* (2017) ‘Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition’, *Speech Communication*, 89, pp. 103–112. Available at: https://doi.org/10.1016/J.SPECOM.2017.03.003.

Mocanu, B., Tapu, R. and Zaharia, T. (2023) ‘Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning’, *Image and Vision Computing*, 133, p. 104676. Available at: https://doi.org/10.1016/J.IMAVIS.2023.104676.

Mohammad, S.M. and Turney, P.D. (2013) ‘CROWDSOURCING A WORD–EMOTION ASSOCIATION LEXICON’, *Computational Intelligence*, 29(3), pp. 436–465. Available at: https://doi.org/10.1111/J.1467-8640.2012.00460.X.

Mollahosseini, A., Hasani, B. and Mahoor, M.H. (2019) ‘AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild’, *IEEE Transactions on Affective Computing*, 10(1), pp. 18–31. Available at: https://doi.org/10.1109/TAFFC.2017.2740923.

Mothukuri, V. *et al.* (2021) ‘A survey on security and privacy of federated learning’, *Future Generation Computer Systems*, 115, pp. 619–640. Available at: https://doi.org/10.1016/J.FUTURE.2020.10.007.

Nandi, A. and Xhafa, F. (2022) ‘A federated learning method for real-time emotion state classification from multi-modal streaming’, *Methods*, 204, pp. 340–347. Available at: https://doi.org/10.1016/J.YMETH.2022.03.005.

Nguyen, T.D. *et al.* (2021) ‘FLAME: Taming Backdoors in Federated Learning’, *Proceedings of the 31st USENIX Security Symposium, Security 2022*, pp. 1415–1432. Available at: https://doi.org/10.48550/arxiv.2101.02281.

Pan, J. *et al.* (2023) ‘Multimodal Emotion Recognition Based on Facial Expressions, Speech, and EEG’, *IEEE Open Journal of Engineering in Medicine and Biology*, 5, p. 396. Available at: https://doi.org/10.1109/OJEMB.2023.3240280.

Pan, S.J. *et al.* (2010) 'Cross-domain sentiment classification via spectral feature alignment', *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pp. 751–760. Available at: https://doi.org/10.1145/1772690.1772767.

Pan, S.J. and Yang, Q. (2010) 'A survey on transfer learning', *IEEE Transactions on Knowledge and Data Engineering*, 22(10), pp. 1345–1359. Available at: https://doi.org/10.1109/TKDE.2009.191.

Pang, B. and Lee, L. (2008) 'Opinion Mining and Sentiment Analysis', *Foundations and Trends® in Information Retrieval*, 2(1–2), pp. 1–135. Available at: https://doi.org/10.1561/1500000011.

Pazzani, M.J. and Billsus, D. (2007) 'Content-based recommendation systems', in *The adaptive web: methods and strategies of web personalization*. Springer, pp. 325–341.

PLUTCHIK, R. (1980) 'A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION', *Theories of Emotion*, pp. 3–33. Available at: https://doi.org/10.1016/B978-0-12-558701-3.50007-7.

Qu, L. *et al.* (2022) 'Rethinking Architecture Design for Tackling Data Heterogeneity in Federated Learning', pp. 10061–10071.

Rabiner, L.R. (1989) 'A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition', *Proceedings of the IEEE*, 77(2), pp. 257–286. Available at: https://doi.org/10.1109/5.18626.

Radoi, A. and Cioroiu, G. (2024) 'Uncertainty-Based Learning of a Lightweight Model for Multimodal Emotion Recognition', *IEEE Access*, PP, p. 1. Available at: https://doi.org/10.1109/ACCESS.2024.3450674.

Retail and Wholesale Transformation May Require up to 600 Billion in Investments to Future-Proof the Sector - EuroCommerce (no date). Available at: https://www.eurocommerce.eu/2022/10/retail-and-wholesale-transformation-may-require-up-to-600-billion-in-investments-to-future-proof-the-sector/ (Accessed: 6 March 2025).

Ricci, F., Rokach, L. and Shapira, B. (2022) 'Recommender Systems: Techniques, Applications, and Challenges', *Recommender Systems Handbook: Third Edition*, pp. 1–35. Available at: https://doi.org/10.1007/978-1-0716-2197-4_1.

Roschewitz, D. *et al.* (2021) 'IFedAvg: Interpretable Data-Interoperability for Federated Learning'. Available at: https://doi.org/10.48550/arxiv.2107.06580.

Russell, J.A. (1980) 'A circumplex model of affect', *Journal of Personality and Social Psychology*, 39(6), pp. 1161–1178. Available at: https://doi.org/10.1037/H0077714.

Salas-Cáceres, J. *et al.* (2025) 'Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics', *Multimedia Tools and Applications*, 84(23), pp. 27327–27343. Available at: https://doi.org/10.1007/S11042-024-20227-6/TABLES/4.

Scherer, K.R. (1995) ‘Expression of emotion in voice and music’, *Journal of Voice*, 9(3), pp. 235–248. Available at: https://doi.org/10.1016/S0892-1997(05)80231-0.

Sharma, R. *et al.* (2022) ‘Unifying the Discrete and Continuous Emotion labels for Speech Emotion Recognition’. Available at: https://arxiv.org/abs/2210.16642v1 (Accessed: 1 September 2025).

Shoumy, N.J. *et al.* (2020) ‘Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals’, *Journal of Network and Computer Applications*, 149, p. 102447. Available at: https://doi.org/10.1016/J.JNCA.2019.102447.

Shum, H. yeung, He, X. dong and Li, D. (2018) ‘From Eliza to XiaoIce: challenges and opportunities with social chatbots’, *Frontiers of Information Technology and Electronic Engineering*, 19(1), pp. 10–26. Available at: https://doi.org/10.1631/FITEE.1700826/METRICS.

Simić, N. *et al.* (2024) ‘Enhancing Emotion Recognition through Federated Learning: A Multimodal Approach with Convolutional Neural Networks’, *Applied Sciences 2024, Vol. 14, Page 1325*, 14(4), p. 1325. Available at: https://doi.org/10.3390/APP14041325.

Spezialetti, M., Placidi, G. and Rossi, S. (2020) ‘Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives’, *Frontiers in Robotics and AI*, 7, p. 532279. Available at: https://doi.org/10.3389/FROBT.2020.532279/XML.

Su, H. *et al.* (2023) ‘Recent advancements in multimodal human–robot interaction’, *Frontiers in Neurorobotics*, 17, p. 1084000. Available at: https://doi.org/10.3389/FNBOT.2023.1084000/XML.

Suresh, V., Yeo, G. and Ong, D.C. (2021) ‘Critically examining the Domain Generalizability of Facial Expression Recognition models’. Available at: https://arxiv.org/abs/2106.15453v2 (Accessed: 2 September 2025).

Di Tecco, A., Foglia, P. and Prete, C.A. (2024) ‘Video Quality Prediction: An Exploratory Study with Valence and Arousal Signals’, *IEEE Access*, 12, pp. 36558–36576. Available at: https://doi.org/10.1109/ACCESS.2024.3374056.

Tomkins, S.S. (2008) *Affect imagery consciousness: the complete edition: two volumes*. Springer publishing company.

Vaswani, A. *et al.* (2017) ‘Attention Is All You Need’, p. 1. Available at: https://arxiv.org/abs/1706.03762v7 (Accessed: 7 September 2025).

Wang, Y. *et al.* (2022) ‘A systematic review on affective computing: emotion models, databases, and recent advances’, *Information Fusion*, 83–84, pp. 19–52. Available at: https://doi.org/10.1016/J.INFFUS.2022.03.009.

Wen, J. *et al.* (2022) ‘A survey on federated learning: challenges and applications’, *International Journal of Machine Learning and Cybernetics 2022 14:2*, 14(2), pp. 513–535. Available at: https://doi.org/10.1007/S13042-022-01647-Y.

Wu, Y., Daoudi, M. and Amad, A. (2023) ‘Transformer-based Self-supervised Multimodal Representation Learning for Wearable Emotion Recognition’, *IEEE Transactions on Affective Computing*, 15(1), pp. 157–172. Available at: https://doi.org/10.1109/TAFFC.2023.3263907.

Yang, Q. *et al.* (2019) ‘Federated Machine Learning’, *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2). Available at: https://doi.org/10.1145/3298981.

Ye, H., Zhou, Y. and Tao, X. (2023) ‘A Method of Multimodal Emotion Recognition in Video Learning Based on Knowledge Enhancement’, *Computer Systems Science and Engineering*, 47(2), pp. 1709–1732. Available at: https://doi.org/10.32604/CSSE.2023.039186.

Yu, E. *et al.* (2024) ‘A federated recommendation algorithm based on user clustering and meta-learning’, *Applied Soft Computing*, 158, p. 111483. Available at: https://doi.org/10.1016/J.ASOC.2024.111483.

Yu, X. *et al.* (2023) ‘Real-Time EEG-Based Emotion Recognition’, *Sensors (Basel, Switzerland)*, 23(18), p. 7853. Available at: https://doi.org/10.3390/S23187853.

Zeng, Z. *et al.* (2009) ‘A survey of affect recognition methods: Audio, visual, and spontaneous expressions’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1), pp. 39–58. Available at: https://doi.org/10.1109/TPAMI.2008.52.