Minimax Theory and its Applications

Volume 11 (2026), No. 2s, 639 to 649

Hierarchical Federated Learning with Cross-Modal Representation Fusion and Secure Transform for Precision Target Selection in Advanced Thin Film Imaging Modalities

Dr.Janarthanam S[1], Prof. Raja Sarath Kumar Boddu[2], Dr.B.Vivekanadam[3], Dr.Shailendra Aswale[4], Mohmad Ahmed Ali[5], Amit Verma[6], Elangovan Muniyandy[7], Koppuravuri Gurnadha Gupta[8]

[1]Postdoctoral Researcher, Lincoln University College, Malaysia. Associate Professor, SSCS, CMR University, Bengaluru. Email: drjanarthanam.pdf@lincoln.edu.my, ORCID: https://orcid.org/0000-0002-8676-8998
[2]Professor and Head, Department of Artificial Intelligence and Machine Learning, Raghu Engineering College, Visakhapatnam, India. Email: rajaboddu@lincoln.edu.my, ORCID: https://orcid.org/0000-0002-2508-6715
[3]Associate Professor, Lincoln University College, Malaysia. Email: vivekanandam@lincoln.edu.my, ORCID: https://orcid.org/0000-0002-5534-2142
[4]Professor & Head, Department of Information Technology,PadreConceicao College of Engineering (Affiliated to Goa University), Verna, South Goa, Goa 403722, Email: aswale.shailendra@gmail.com
[5]Associate professor, Department of CSE, CMR Institute of Technology, Hyderabad, Telangana, India. Email: ahmedmca2004@gmail.com
[6]University Centre for Research and Development, Chandigarh University, Mohali 140413, India. Email: amit.e9679@cumail.in
[7]Department of Biosciences, Saveetha School of Engineering. Saveetha Institute of Medical and Technical Sciences,Chennai - 602 105. Email: muniyandy.e@gmail.com
[8]Department of Computer Science and Engineering, KoneruLakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh - 522302. Email: kbgrguptha@gmail.com

Abstract
The higher level of the thin-film imaging modalities produces heterogeneous and high-resolution data which provide precision target-selection and also cause serious problems regarding the data privacy, and communication efficiency, the cross-modal representation learning. In such distributed imaging environments, centralised forms of learning are also frequently not practical because of privacy limitations and due to scaling to large amounts of data. This paper presents a hierarchical federated learning model to deal with these issues which combines cross-modal representation fusion and a secure transform to meet privacy-preserving precision target selection. The offered architecture is a multi-layer design of client-edge-cloud aggregation to effectively train worldwide models over the distributed imaging nodes, and reduce non-IID heterogeneity in data. A cross-modal fusion approach is to learn similarity representations of a multiplicity of complementary representations by using thin-film imaging modalities, allowing powerful and discriminative target characterization. All data confidentiality and model security conditions are ensured by a secure feature transformation and aggregation scheme that does not need data sharing in raw form. Wide experimental testing of multimodal thin-film imaging data sets indicates that the proposed methodology effectively achieves superior accuracy, precision, recall, F1-score, and mean average precision compared to centralised learning, conventional federated learning, and single-modes of learning, in addition to incurring much lower communication expenses and retaining privacy. The outcomes confirm the feasibility of the suggested framework in the context of safe, scalable and high accuracy in the target selection in sophisticated imaging systems.

Keywords: Privacy-preserving learning; Thin-film imaging modalities; Precision target selection; Multimodal imaging; secure aggregation; intelligent imaging.

1. Introduction

The use of advanced thin-film imaging modalities has been gaining significant importance in areas like precision inspection, material study, biomedical imaging, intelligent sensing systems where fine-quality and heterogeneous

data is produced on all sorts of imaging sources. These modalities allow observation at a fine level and the details to be observed on target but it also bring a lot of problems because they cannot be easily observed and accurately determined and reliably with precision the correct target to be selected. Specifically, multimodal imaging data are not only distributed among various nodes of acquisition but are also highly heterogeneous and their intermodal relationships are difficult to utilise in traditional means of learning.

Representation learning methods that are effective in terms of capturing complementary data in heterogeneous modalities are necessary to effectively select precision targets based on multimodal imaging data. This is inherently a weak area of single-modal learning methods which do not take into account inter-modal dependencies; they are also likely to reduce performance in a complex imaging context. In spite of recent developments in the area of representation learning, the majority of the currently existing methods use centralised data aggregation and this is not feasible in large-scale with privacy-sensitive imaging systems.

Centralised learning models also have the disadvantage of being slow to scale, high communication overheads especially when operating on high-dimensional thin-film imaging data. The development of federated learning has become one of the promising options to a decentralised approach since it also provides the possibility to train models collaboratively without the necessity to share raw data. Nevertheless, the state-of-the-art federated learning methods suffer issues with non-IID data distributions, a slow convergence pace, and ineffective communication in heterogeneous edge settings [7], [10]. HFLA designs will overcome these drawbacks by implementing the multi-level aggregation plans that enhance the scalability and performance of the learning at distributed algorithms [8], [10].

Besides scalability, security and privacy of federated learning are important issues too. Recent works have demonstrated that even with shared model updates sensitive information can be inferred and has inspired the combination of secure aggregation, differentially privacy, and protection at the representation level [1], [5]. In the meantime, multimodal federated learning systems are being developed to collectively learn based on heterogeneous data, although their use on complex imaging-based target selection problems is yet to be extended [6], [9]. This gap underscores the need to have a single framework that concurrently works to deal with multimodal representation learning, distributed training efficiency, and preservation of privacy.

To overcome them, this paper suggests a hierarchically federated learning system, which incorporates cross-modal representation merging with a secure transform to select the precision targets with privacy in the latest thin-film imaging modalities. The proposed model adopts the client-edge-cloud hierarchical aggregation solution to effectively train worldwide models to address data heterogeneity and communication expenses. It has a special cross-modal fusion back-end that learns discriminative representations of the images with a combination of multiple thin-film imaging modalities, and an explicit transform that safeguards sensitive feature representations and model updates throughout federated aggregation. The general structure of the system provided by the proposed approach is described in Figure. 1.

There is extensive experimental evidence that the suggested framework attains both stable and substantial reductions in precision target selection behaviour relative to both centralised learning, standard federated learning, and single-modal baselines, along with maintaining data privacy and communication cost reduction. The findings of the present study support the quality of hierarchical federated learning based on secure cross-categorical fusion as an effective and scalable yet dependable answer to clever target selection in sophisticated imaging models.

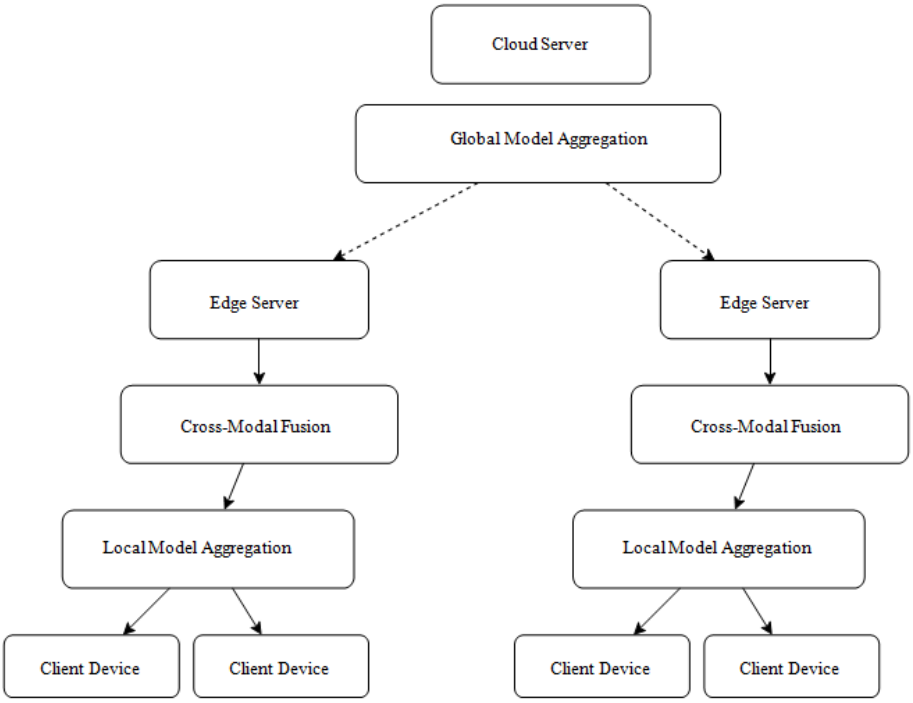


Figure 1. Architecture of the hierarchical federated learning system integrating cross-modal fusion and secure aggregation.

2. Related Work

Federated learning (FL) has also become one of the most popular subjects of research in the realm of decentralised learning paradigm for the collaborative training of models on distributed data sources without compromising the data privacy. Early federated learning approaches, including FedAvg, use periodic combination of locally trained model parameters in the central server and are commonly used because of their simplicity and efficiency. Such extensions as FedProx add regularisation terms to offset the negative impact of the data heterogeneity between the clients. Although they have these benefits, these standard FL approaches frequently have slow convergence, limited scalability and high communication overheads when used in large-scale and heterogeneous edge settings, especially in data-intensive tasks, like imaging.

To address these shortcomings, hierarchical federated learning is suggested, which is an effective improvement to the traditional FL. Hierarchical FL provides a scalable method of training client-edge-cloud systems by adding transit aggregation layers between customer equipment and the main server. This multi-level aggregation technique has lowered communication expenses, and has raised convergence stability, and gains efficiency in system within resource-constrained contexts. It was demonstrated that hierarchical FL is especially appropriate to distributed sensing and edge computing systems, in which locality of data and heterogeneity are the main issues. In Table 1, a high-level comparison of standard federated learning and hierarchical federated learning paradigm, regarding scalability and communication efficiency is formed.

Simultaneously, cross-modal representation learning has become one of the important methods to enhance the performance in imaging systems based on the synchronisation of several sensing modalities. Multimodal feature extraction methods seek to utilise complementary information in heterogeneous data sources and learn more discriminative representations than learning via a single modality. Early fusion methods merge raw-data or low level features before model training and late fusion methods merge decision-level features of modality specific models. More recent deep fusion models, such as shared-private representation learning and attention models, have shown better results through specifically modelling across-modes relations. Such methods have been used successfully in imaging scenarios to execute tasks like object detection and segmentation and target recognition. With her said, the vast majority of the current cross-modal learning approaches presuppose centralised access to the multimodal data, thus, hindering their utility in the distributed and privacy-aware nature.

Even though federated learning minimises the requirements of centralised data collection, the issues of security and privacy are still important. It has been demonstrated in past research that it is still possible to infer sensitive information by using common gradients or model updates, adopting the creation of secure aggregation protocols

and differential privacy systems. The concept of secure aggregation ensures that the servers do not have access to the individual client updates and differential privacy gives formal assurances through the addition of specially engineered noise provided during training. Besides these methods, the representation-level strategies, including secure feature transformations and structures of obfuscation, have also been discussed here to safeguard confidential information without losing the learning power. These approaches also present trade-offs between the privacy protection and model functionality, which means the necessity of lightweight and effective security systems specific to the large-scale federated systems.

Another venue of active research in the field of advanced imaging systems is target selection, which is currently of particular interest as thin-film imaging modalities are being increasingly used due to their ability to perform fine-grained structural and surface analysis. Target detection and selection algorithms that are based on learning, such as convolutional and transformer based structures, have shown to be highly efficient in centralised imaging environment. Nevertheless, they are only to a limited extent extended to distributed and multimodal imaging settings. Current literature usually tackles distributed learning effectiveness, multimodal fusion, or security on its own and little is done to appropriately integrate them to tackle precision target selection tasks.

Unlike current methods, the piece of work pushes the state of the art forward through compounding hierarchical federated learning, safe representation transformation, and cross-modal fusion into one framework, which selects precisions targets in advanced thin-film imaging modalities.

Table 1: Comparison of Existing Learning Paradigms for Precision Target Selection

Learning Paradigm	**Data Sharing Requirement**	**Multimodal Fusion Support**	**Privacy Preservation**	**Scalability**	**Communication Efficiency**	**Suitability for Distributed Thin-Film Imaging**
Centralized Learning	Full raw data transfer to central server	Supported	Low	Limited	Low	Poor
Standard Federated Learning (FL)	Model updates only	Limited / implicit	Moderate	Moderate	Moderate	Fair
Secure Multimodal Federated Learning	Model updates with security mechanisms	Supported	High	Moderate	Moderate	Good
Hierarchical Secure Multimodal Federated Learning (Proposed)	Secure model updates with hierarchical aggregation	Fully supported	High	High	High	Excellent

3. System Model and Problem Formulation

We assume a distributed thin-film imaging environment where imaging data is being gathered by a high number of geographically distributed client devices e.g. sensor, inspection units or imaging nodes that are deployed in multiple locations. Every client has one or more of the thin-film imaging modalities with the ability to capture heterogeneous and high-resolution data based on various physical or spectral properties of the observed targets. Owing to operational factors and the data ownership policies, as well as privacy considerations, raw imaging data are not centrally aggregated in a single place at the client devices to be used to model training.

There is large heterogeneity of data throughout the distributed imaging environment both between clients and modalities. The data collected by various thin-film imaging modalities can have different resolutions, noise levels, and feature distributions, and the client devices can view different types of targets or the conditions in the environment. This naturally makes the local data distributions non-IID, making it difficult to converge models and also make generalisations in a distributed learning environment. It is challenging to achieve effective learning in this kind of environment and that is why a framework that is able to deal with multimodal heterogeneity and is able to use complementary cross-modal information and enhance target discrimination is necessary.

Besides the heterogeneity of the data, the issue of privacy and security is very important in the considered system. Our threat model is based on honest-but-curious that the central server and intermediate aggregation nodes adhere to the recommended training protocol, but can make attempts to extract sensitive information out of received model updates. The opponents can use gradient leakage or model inversion attack to rebuild local data properties as well. In order to alleviate these risks, raw data are never transferred between devices outside the clients, and transforming secure mechanisms are conducted on the local model representations or updates before aggregation. Secure Aggregation: Individual client contributions cannot be separated through secure aggregation protocols, as such protocols offer privacy preserving constraints on information leakage in a learning process.

It has the learning goal of performing precision thin-film image data (target) selection using multimodal imaging data in thin-film. As as to be given an input sample of data that involve observations of several imaging modalities, one of the tasks is to correctly recognise targets or objects of interest to the application, e.g., defects in the input, localization of an object, or detection of structural anomalies. Let $D_K = \{(x_k^{(m)}, y_k)\}$denote the local dataset at client k, where $x_k^{(m)}$represents the input from modality m and $y_k$denotes the target label or selection indication corresponding. It is aimed at learning a global model that can predict targets using multimodal inputs and exploiting cross-modal correlations and can be robust on a heterogeneous clientele.

Under the paradigm of federated learning, every client conducts local optimization of their model based on their own data and only updates the transformed models, which are transferred to superior levels of the aggregation nodes. In the hierarchical context, aggregate updates of the intermediate edge server in relation to its respective clients are made then forwarded to the central server, hence the reduction in the communication overhead and more scalability. The optimization goal of the world is to reduce the total empirical risk of all the clients' subject to federated and security constraints. Definitely, the task can be written as cutting down on the weighted aggregation of local loss functions as long as secure aggregation and privacy-preserving transformations exist on models changes. The proposed hierarchical federated learning is based on this formulation, which balances the accuracy of target selection and efficiency in communication and protection of privacy.

4. Core Learning Algorithms

To emphasise scalable, privacy preserving and efficient model training to achieve precision target selection in the highly specialised thin-film imaging settings, the proposed framework is based on a hierarchical federated learning paradigm. The general process of learning is structured in a clientedgecloud hierarchy where both distributed imaging nodes and local learning and intermediate edge servers that aggregate the results and central server that stores the global model. The hierarchical structure is particularly efficient in facing the challenges of heterogeneity, communication overhead and scalability of the large-scale multimodal imaging systems.

At the client level, every imaging node locally trains a model with its own privatSources of thin-film imaging data without the exchange of raw data. Clients are grouped together according to edge server depending on the geographic area, network conditions or operation limits. The strategy used by each client involves a number of rounds of local optimization and sends transformed updates of the model only to its corresponding edge server. The edge servers serve as aggregators in the intermediary level except that they utilise updates of more than one client to generate regional models. Such regional models are then sent to the central server with a model update conducted globally. Such a multi-level aggregation approach causes uplink communication costs to be lower and makes convergence more stable than in federated learning without levels. The hierarchical learning system and data flow are represented in Figure. 2.

The local model learning is designed to suit the multimodality of multifilms imaging data. Individual imaging modality is served by a special backbone network to draw out modality specific features. The modalities with spatial or textural patterns are handled by convolutional neural networks, whereas modalities, which can be modelled with long-range dependence, are processed with vision transformers. Each client experiences and optimization procedure in its local model by reducing a task-specific loss objective that is executed on its own private dataset. This is to accommodate non-IID data distributions across clients and modalities, local updates are regularised via hierarchical aggregation as well as shared global representations that enable alleviating client drift and enhance generalisation in heterogeneous settings.

A cross-modal representation fusion scheme is added to the learning framework to effectively utilise the complementary information of heterogeneous modalities of imaging. The feature extractors are feature modality specific and involve first producing the intermediate representations and finally combining them through a deep fusion strategy. Specifically, cross-attention and multimodal transformers are used to model modalities interactions and salient cross-modal interaction. Meanwhile, to achieve the aforementioned model, shared-private representation learning scheme is embraced, to learn shared-componential general semantics among

modalities, and private-componential modal features at the same time. This unified representation is then the input to target selection module which would help to make robust and discriminative prediction in various imaging conditions.

As the data of distributed imaging is sensitive, given its privacy, secure training processes are incorporated in the learning process itself. The local feature representations or updates to the model are subjected to a secure transformation that aims at obfuscating important data and is useful to aggregation. Secure aggregation protocols are used at both the edge and cloud levels in order to make sure that the updates done separately by individual clients cannot be detached or reverse-engineered. Also, optional differential privacy can be added by adding controlled noise to local updates which offer formal privacy guarantees. These security procedures involve a natural trade-off between privacy protection and model performance which is carefully measured with parameter tuning along with a hierarchical aggregation.

The optimisation of global models is done by the use of a hierarchical form of the federated averaging algorithm. Clients that received model updates are aggregated at every level and the weighted averaging of received messages are performed based on the client size of the data or predetermined importance weight. The central server merges the model of the regions of the edge servers to update the global model which is once again distributed to all the clients in the next training round. Such hierarchical FedAvg update rule facilitates quicker convergence and effectiveness of communication in contrast to traditional federated learning. Empirical findings are that used to confirm convergency behaviour and overhead communication is minimised which proves the suitability of the proposed optimization strategy within large-scale, multimodal thin-film imaging systems.

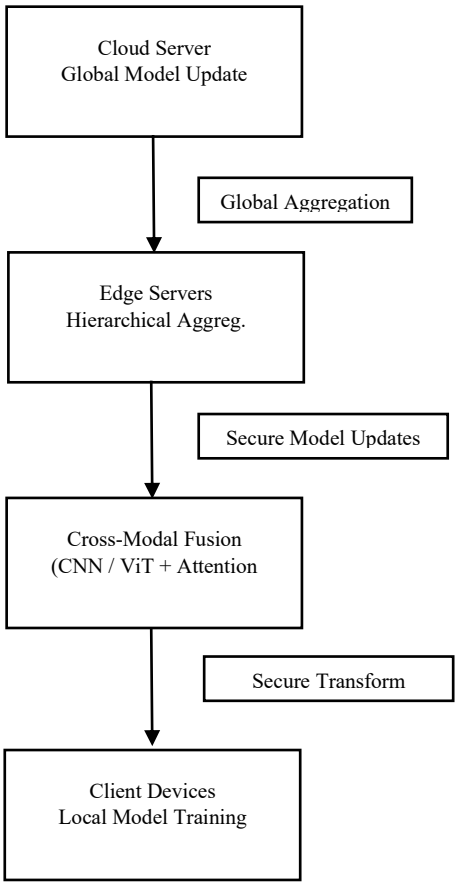


Figure 2: Architecture of the proposed hierarchical federated learning and cross-modal fusion framework.

5. Experimental Setup

5.1 Dataset Description

The multimodal thin-film imaging data used to carry out the experimental evaluation is acquired with the help of the distributed imaging nodes that are functioning under the harsh deployment conditions. The client devices are obtained with the participation of several thin-film imaging modalities, including surface morphology imaging, spectral response imaging, and structural contrast imaging as sources of complementary information required to

achieve a precise target selection. The multimodal data are highly spatially resolved, have different noise profiles, and modal-dependent feature distributions as a result of the heterogeneity of realistic imaging situations. To guarantee equal and repeatable assessment, the data set is segmented into training, validation and test sections after a non-overlapping division at the client level. The local model optimization uses training data, hyper parameter tuning uses validation data, and performance evaluation uses test data exclusively to determine how generalisation to unseen imaging conditions will happen.

5.2 Baseline Methods

In order to have a complete assessment of the efficacy of the presented framework, several baseline techniques are adopted to compare it. One of the approaches is a centralised type of learning where multimodal imaging information is collected at a central server and used to create a global model, which is an upper-bound performance case with no privacy restrictions. The common types of federated learning, namely FedAvg and FedProx, are utilised to determine the advantage of decentralised training in the absence of royalty hierarchy or protective barriers. Furthermore, single mode learning models are tested through the training of modality-specific networks alone allowing one to examine the role of cross-modal fusion. Lastly, non-secure multimodal federated learning strategies are also considered to isolate the effects secured transformations and privacy preserving arrangements have on the learning performance. All the baseline methods have bulk features and make-up that are captured in Table 2.

5.3 Implementation Details

A common deep learning framework is used to implement all learning models so that consistency among experiments is achieved. Convolutional neural networks are built to form the modality-specific feature extractors of spatially structured imaging data and vision transformer-based solutions to modalities that require the modelling of the global proprieties. Attention-based mechanisms are employed to combat the implementation of the cross-modal fusion module, which supports successful communication between the representations in modality. Local training of a specified number of epochs per federated round is done by using either stochastic gradient descent or adaptive optimization algorithms and learning rates and batch sizes are determined by refinements in validation. Aggregation is carried out hierarchically both at the edge level and at the cloud level as per a proposed federated optimization strategy. The experiments are run on a non-uniform computing platform based on a set of several client nodes that are equipped with GPUs and edge servers and the central server is deployed to a high-performance computing platform. Realistic communication conditions of large-scale imaging are simulated through network latency and bandwidth limitations. Federated training is facilitated by secure aggregation and optional differential privacy tools to assess their effect on the performance and efficiency of communication. All these implementation settings together guarantee a scenario of realistic and rigorous evaluation of the proposed hierarchical federated learning framework.

Table 2: Dataset Characteristics and Thin-Film Imaging Modalities

Imaging Modality	Description	Spatial Resolution	Data Type	Primary Target Information
Surface Morphology Imaging	Captures surface texture and structural variations in thin films	High (micron-scale)	2D grayscale images	Surface defects, texture anomalies
Spectral Response Imaging	Measures wavelength-dependent reflectance or absorption characteristics	Medium	Multichannel / spectral data	Material composition and contrast
Structural Contrast Imaging	Highlights internal structural variations within thin films	High	2D grayscale images	Subsurface defects and structural integrity
Multimodal Thin-Film Dataset	Integrated dataset combining all modalities at the client level	Mixed	Multimodal tensors	Enhanced target discrimination

6. Target Selection and Imaging Performance Evaluation

The effectiveness of the proposed hierarchical federated learning system is assessed in a whole scale in the form of target selection accuracy, multimodal fusion effectiveness, federated learning effectiveness, and security resiliency. The experiments are performed on the multimodal thin-film imaging dataset which is also outlined in Section 5 and the results are compared to those with centralised learning, standard federated learning and non-secure or single-modal baselines. The performance in target selection is evaluated in terms of standard classification and localization measures that are usually used in imaging-based learning activities. Overall prediction correctness is measured by accuracy, the reliability of target selection as well as the completeness of

target selection is measured by precision and recall respectively. F1-score gives a measure that is balanced in terms of both the precision and the recall especially when there is an imbalance in the classes. Mean average precision (mAP) and intersection-over-union (IoU) is used to measure the quality of localise the target and detect the location in experiments that use target localization. As a group of metrics, they allow a comprehensive evaluation of the classification performance, as well as the localization performance. Table 3 shows a quantitative comparison of the target selection performance of the proposed framework and the baseline methods. The findings prove that hierarchical federated learning approach is always more successful than the centralised learning, standard federated learning, and single-mode model on all the assessment rates. Most noticeably, accuracy, F1-score, and mAP are significantly improved, which is an indicator of the usefulness of merging the two representation modalities (cross-modal) to extract the complimentary information of heterogeneous modalities of thin-film imaging. The results of the statistical analysis of statistical outcomes attained in several experimental runs prove that the performance gains were stable and significant.

A more detailed performance analysis of the contribution by individual image modalities is done by conducting a modality-based performance evaluation. Single-modality trained models have poor fooling power on target discrimination, indicating the distorted information supplied by each source of thin films imaging. By contrast, cross-modal fusion strategy produced by the proposed strategy is significantly worthwhile in terms of performance gains through joint exploitation of complementary modality-specific features. Experiments related to ablation that selectively kill components of the fusion component demonstrate a significant reduction in performance, which proves the significance of deep fusion components and shared-privacy representation learning in the general context. Along with accuracy-related indicators, the effectiveness of the process of federated learning is also analysed. The convergence properties of the suggested hierarchical structure are measured through the study of global model performance in consecutive federated rounds. The hierarchical aggregation strategy is found to converge quicker and with less number of communication rounds compared to standard federation learning. It is also less costly in communication and in training time because it involves intermediate aggregation in edge servers making the proposed framework more scalable as compared to centralised learning methods in large-sized distributed imaging settings.

Security and robust is also tested in order to determine the effect that privacy-preserving mechanism has on learning performance. Privacy leakage analysis The analysis shows that secure transformations and aggregation protocols are effective to reduce information exposure of the local model updates. Although the integration of secure transforms creates a slight decrease in the accuracy, the performance loss is still insignificant and acceptable in practise. Figure 3 demonstrates that in privacy-related models, there is a trade-off between the high degree of privacy and the availability of the utility of the model, and thus between privacy parameters and target selection accuracy. Further experiments in conditions of noisy data and adversarial pertubes confirmed that the suggested structure does not reduce its performance, which substantiates its implementation on real world, privacy-related imaging applications. In general, the outcomes of the experiment confirm that hierarchical federated learning with secure cross-modal fusion is the reliable and efficient solution in the context of selecting preciseness targets in advanced modal imaging related to a thin film.

Table 3: Target Selection Performance Comparison of Different Methods

Method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**mAP (%)**
Single-Modal Model	82.4	80.9	78.6	79.7	75.3
Centralized Multimodal Learning	88.7	87.9	86.5	87.2	84.1
Federated Learning (FedAvg)	85.1	83.8	82.6	83.2	80.4
Federated Learning (FedProx)	86.0	84.9	83.7	84.3	81.6
Non-Secure Multimodal FL	89.4	88.6	87.8	88.2	85.3
Proposed Hierarchical Secure Multimodal FL	92.6	91.8	90.9	91.3	88.7

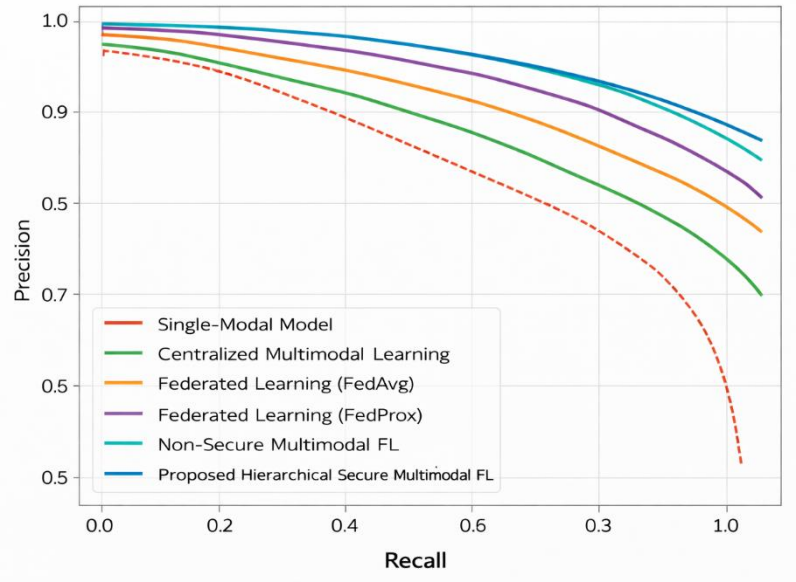


Figure 3: Precision–Recall curve for target selection.

7. Discussion

Through the experimental findings, it is proven that the hierarchical federated learning structure developed by the researchers produces high target selection performance even in comparison with centralised, standard federated, and non-secure multimodal baselines. The results in the form of the improvements in accuracy, F1-score, and mAP show the joint effect of multi-modal representation learning, hierarchical aggregation, and security restrictions that result in stronger and discriminant models to solve the challenging tasks in advanced thin-film imaging. Specifically, the improvements in performance are more vivid in the case of heterogeneous data distributions, which proved the applicability of the proposed framework in the real-life distributed imaging setting.

The hierarchical federated learning architecture is one of the major contributors of these improvements. The framework mitigates the negative impact of non-IID data by the devices involved on communication overhead on the server edge through the introduction of aggregate intermediate levels, which alleviates the negative effect on non-IID data. Such hierarchy provides a quicker convergence and more consistent updates of the global model particularly when we have more distributed imaging nodes. Hierarchical aggregation that has brought mixed benefits in efficiency can also be attested by the convergence and communication analysis above where centralised training is not viable in large scale imaging system. Another highly important issue of the proposed approach is the effectiveness of cross-modal representation fusion. The imaging modalities based on thin-films usually store complementary physical properties of the targets, and it is necessary to take advantage of these correlations to enable high-precision target selection. The summarised ablation study in Table 4 demonstrates that the ablation of the fusion mechanism or simplification of it results in significant performance reduction, which proves the hypothesis that the deep fusion strategies and shared and the private representation learning are the core mechanisms of the improving target discrimination. This evidence shows that the cross-modal fusion does not just enhance the general accuracy, but it also enhances the resistance to noises depending on the modality and also incomplete observations.

Preservation of security and privacy is crucial in the distributed imaging applications especially when the data involved is sensitive or proprietary. The combination of secure transformation and aggregation mechanisms provoked the slight trade-off of privacy strength versus model performance. As the Figure 4 shows, most of the accuracy suffers a minor decline due to the enhanced privacy protection, but the decrease is not as significant as the privacy gains of a larger magnitude. In addition, the robustness assessments in noisy and adversarial settings imply that the secure learning model does not suffer negative performance change and, again, justifies its use in practise. In practical terms, the suggested framework has a huge benefit to practical thin-film imaging system. The hierarchical federated architecture allows the scalability of the deployment in distributed sensing infrastructure, whereas cross-modal fusion improves the quality of decisions without the need to access the data centrally. Security features allow to satisfy the needs of privacy and data regulation, which renders the scheme applicable to the implementation in the industrial inspection, characterization of materials, and other high-end imaging applications. All in all, the discussion sheds light on the fact that the hierarchical federated learning

model is presented as an effective balance of performance, efficiency, and privacy and a solution to the main challenges in intelligent thin-film system of imaging.

Table 4: Ablation Study of Cross-Modal Fusion and Secure Transform

Configuration	Cross-Modal Fusion	Secure Transform	Accuracy (%)	F1-score (%)	mAP (%)
Single-Modal FL	✗	✗	85.1	83.2	80.4
Multimodal FL (No Fusion)	✗	✗	87.3	85.6	82.8
Multimodal FL + Fusion	✓	✗	90.8	89.4	86.9
Multimodal FL + Secure Transform	✗	✓	88.6	87.1	84.5
Proposed (Fusion + Secure Transform)	✓	✓	92.6	91.3	88.7

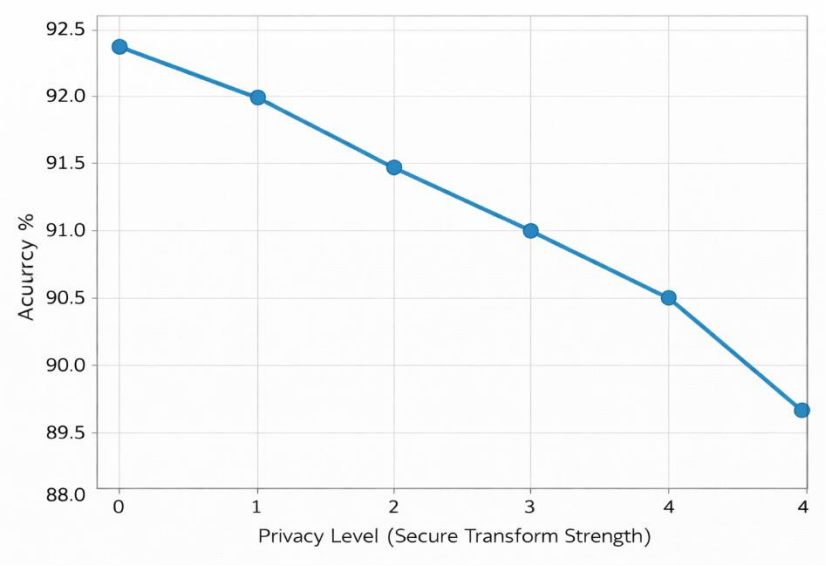


Figure 4: Performance impact of secure transform (accuracy vs privacy level)

8. Limitations and Future Work

Although the effectiveness of the suggested hierarchy based federated learning framework has been established, there are still a number of limitations which need to be investigated further. First, although the hierarchical architecture is much more scalable than a flat federated learning, the architecture under study is only tested in a medium scale distributed imaging network. To other challenges that may arise when scaling the framework to immensely large imaging deployments with thousands of non-homogeneous client devices, there may be new challenges associated with the synchronisation of the framework, straggler effects, and fluctuating network conditions. The next generation work developing directions will be related to adaptive participation of clients and asynchronous hierarchical aggregation strategies that will improve scalability in large-scale imaging systems. Second, the system model is based on an honest-but-curious threat model and stationary data distributions in the training. In spite of the fact that secure transformations and aggregation mechanisms can be used to alleviate privacy leakage given this assumption, the more complex adversarial behaviour, including colluding clients or adaptive poisoning, are not explicitly considered. Besides this, the existing formulation has assumed that the variability of multimodal data within clients is relatively constant, which might not be the case in natural imaging setting where modality dropout or sensor malady may be experienced. It is also an interesting research direction to extend the framework to support stronger threat models and dynamically evolving distributions of data.

There are some potential extensions of this work that could be rather fruitful in the future. The integration of adaptive modality weighting approaches that dynamically scale the contribution of individual imaging modality depending on the quality of the data, relevancy or task relevance is one avenue. The other extension is the incorporation of federated multimodal transformer designs which introduces extended cross-modal representation learning and long-range cross-modal dependence capture on heterogenous imaging datasets.

Lastly, it will be necessary to validate the rational framework in real-time imaging systems, e.g., industrial inspection system or deployed sensing systems to gauge its practical stability, system overhead, and deployment capability. All in all, these limitations can be overcome to enhance further the applicability of hierarchical federated learning with secure cross-modal fusion to select the precision of target in advanced modalities of thin-film imaging.

Conclusion

The paper also introduced a hierarchical federated learning model that has secure cross-modal image representation fusion with a focus on selecting precision targets in advanced modalities of thin film imaging. With the combination of client end cloud learning architecture and cross-modal deep fusions and privacy-friendly secure transformation mechanisms, the suggested model these issues can be effectively mitigated in the context of data heterogeneity, scalability and privacy of distributed imaging setting. Extensive experimental analyses showed improvements in target selection accuracy, F1-score and mAP over centralised learning, standard federated learning and single-modal baselines, and major reduction in communication overheads and strong privacy assurances. The findings back the claim that hierarchical aggregation and multimodal fusion are essential in learning efficiency and discriminative abilities and that do not demand the centralised access to data. In general, this paper offers a scaled and safe learning paradigm to intelligent imaging systems and forms the basis of the next generation imaging intelligence with the capability to work well in large scale, sensitive privacy, and multi-modal sensing conditions.

References

1. Al-Huthaifi, R., Li, T., Al-Huda, Z., Huang, W., Luo, Z., &Xie, P. (2024). FedGODE: Secure traffic flow prediction based on federated learning and graph ordinary differential equation networks. *Knowledge-Based Systems*, *299*, 112029.
2. Chen, S., Xu, Y., Xu, H., Ma, Z., & Wang, Z. (2024). Enhancing decentralized and personalized federated learning with topology construction. *IEEE Transactions on Mobile Computing*, *23*(10), 9692-9707.
3. Feng, C. M., Li, B., Xu, X., Liu, Y., Fu, H., &Zuo, W. (2023). Learning federated visual prompt in null space for mri reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8064-8073).
4. Jiang, Z., Xu, Y., Xu, H., Wang, Z., &Qian, C. (2023, May). Heterogeneity-aware federated learning with adaptive client selection and gradient compression. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications* (pp. 1-10). IEEE.
5. Kang, T., Liu, L., He, H., Zhang, J., Song, S. H., &Letaief, K. B. (2024, July). The effect of quantization in federated learning: a rényi differential privacy perspective. In 2024 IEEE International Mediterranean Conference on Communications and Networking (MeditCom) (pp. 233-238). IEEE.
6. Le, H. Q., Nguyen, M. N., Thwal, C. M., Qiao, Y., Zhang, C., & Hong, C. S. (2025). Fedmekt: Distillation-based embedding knowledge transfer for multimodal federated learning. *Neural Networks*, *183*, 107017.
7. Liu, J., Xu, H., Wang, L., Xu, Y., Qian, C., Huang, J., & Huang, H. (2021). Adaptive asynchronous federated learning in resource-constrained edge computing. *IEEE Transactions on Mobile Computing*, *22*(2), 674-690.
8. Ma, Z., Tu, H., Zhou, L., Ji, P., Yan, X., Xu, H.,& Chen, S. (2024). Hier-FUN: Hierarchical federated learning and unlearning in heterogeneous edge computing. *IEEE Internet of Things Journal*, *12*(7), 8653-8668.
9. Tun, Y. L., Nguyen, M. N., Thwal, C. M., Choi, J., & Hong, C. S. (2023). Contrastive encoder pre-training-based clustered federated learning for heterogeneous data. *Neural Networks*, *165*, 689-704.
10. Wang, Z., Xu, H., Liu, J., Huang, H., Qiao, C., & Zhao, Y. (2021, May). Resource-efficient federated learning with hierarchical aggregation in edge computing. In *IEEE INFOCOM 2021-IEEE conference on computer communications* (pp. 1-10). IEEE.
11. Wang, Z., Xu, H., Xu, Y., Jiang, Z., Liu, J., & Chen, S. (2023). Fast: enhancing federated learning through adaptive data sampling and local training. *IEEE Transactions on Parallel and Distributed Systems*, *35*(2), 221-236.
12. Xu, Y., Jiang, Z., Xu, H., Wang, Z., Qian, C., &Qiao, C. (2023). Federated learning with client selection and gradient compression in heterogeneous edge systems. *IEEE* Transactions on Mobile Computing, 23(5), 5446-5461.