

Federated Learning-Based Movie Genre Prediction from Multimodal Content Using Deep Learning: A Privacy-Preserving Approach



Zulkarnain Saurav^{1*}

Abstract

Predicting a movie's genre well is deceptively hard, and most existing approaches, arguably, have not tried hard enough — they tend to lean on a single data source, either visual scene cues or scraped plot text, which leaves a fair amount of a film's identity unaccounted for. Combining posters and storylines seems like the obvious fix, except that pooling multimodal data this way raises a problem few prior studies have taken seriously: doing so typically requires centralizing sensitive content, an approach that sits uneasily with growing concerns around data breaches, poster piracy, and user distrust of centralized systems. This study set out to address both gaps at once. We collected plot summaries and poster images from IMDB, then built a hybrid federated learning (FL) architecture in which text and image data were processed independently — BERT paired with a one-dimensional CNN for multi-label text classification, and VGG-16 for extracting genre-relevant visual features — before their outputs were fused through a late cross-modality representation and aggregated centrally using FedAvg, ensuring that raw data never left its originating client. The resulting model achieved 80% accuracy, a figure that compares favorably to prior unimodal and non-privacy-preserving genre-prediction studies, while never requiring poster or text data to be shared outright. Taken together, these results

suggest that multimodal fusion and federated privacy preservation need not come at each other's expense; if anything, this pipeline indicates the two can be pursued jointly, without meaningfully sacrificing predictive performance, in a way that earlier single-modality or privacy-agnostic approaches simply did not attempt.

Keywords

Federated Learning; Multimodal Data; Movie Genre Prediction; Deep Learning; Data Privacy

1. Introduction

Prediction problems, it turns out, show up almost everywhere data moves through a network, and film genre classification is really just one instance of a much older, much broader family of questions. Content delivery networks have spent years trying to guess what users will want next so they can cache it closer to the edge — Liu et al. (2019) modeled content popularity using an autoregressive approach for information-centric IoT settings, while Hassine, Milocco, and Minet (2017) took a similar ARMA-based route for caching prediction, and Liu, Zhang, Liang, Peng, and Cai (2018) later folded deep learning into that same caching problem using SDN

Author Affiliation:

¹ United International University, United City, Madani Avenue, Badda, Dhaka 1212, Bangladesh.

Please cite this article as: Saurav, Z. (2020). "Federated Learning-Based Movie Genre Prediction from Multimodal Content Using Deep Learning: A Privacy-Preserving Approach", *Data Modeling*, 1(1), 1-9, 10839

© 2020 (Data Modeling), a publication of Eman Research, USA.
This is an open access article under the CC BY-NC-ND license.
(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)
(<https://publishing.emanresearch.org>)

Significance | Bridges multimodal deep learning and federated privacy, offering scalable, trustworthy genre prediction for real-world streaming systems.

*Correspondence.

Zulkarnain Saurav, United International University, United City, Madani Avenue, Badda, Dhaka 1212, Bangladesh. E-mail: kmithu.uiu@gmail.com.

Editor Mohammad Shafiul Alam, Ph.D. And accepted by the Editorial Board Dec 23, 2020 (received for review Oct 25, 2020)

infrastructure. None of these papers are about movies, exactly, but they share something with genre prediction that is easy to overlook: they are all, at bottom, attempts to infer structure from messy usage data so that a system can act sensibly before the full picture is even available. Wang, et al. (2014) made a related case for content caching in 5G delivery pipelines, and Liu, Yu, Li, Ji, and Leung (2018) extended this logic further into fog and cloud resource allocation, arguing that offloading computation intelligently requires the same kind of predictive reasoning. Machine learning, across all these settings, is doing the same basic job: pulling pattern out of noisy, high-volume data and generalizing it forward.

Movie genre prediction sits comfortably inside this lineage, though the entertainment industry rarely frames it that way. Getting a genre wrong, or leaving it too ambiguous, cascades into the same kinds of downstream failures — bad recommendations, misdirected marketing, archives that become difficult to search. Most existing work on the problem, though, has leaned on a single source of information, and this is where things start to feel incomplete. Zhou et al. (2010) built a genre classifier from scene-level visual cues extracted from film footage, categorizing scenes and inferring genre from those categories; it worked reasonably well, but it only ever saw one side of the movie. Saumya, Kumar, and Singh (2018) went the other direction, mining scraped plot summaries for genre signals, arriving at results that were respectable without being exhaustive. Neither approach, taken alone, really captures what a film is — a horror poster carries a visual grammar (unsettling palettes, jagged typography, faces frozen mid-scream) that a plot synopsis would never spell out, and a synopsis often reveals tonal or narrative cues a poster physically cannot show.

That gap is what we are trying to close. We propose a multimodal genre identification method that combines movie posters as the visual channel with storyline synopses as the textual one, on the reasoning that fusing the two should, at least in principle, yield a fuller read on genre than either manages alone. Getting text representation right matters quite a bit here, and earlier NLP work has shaped how we think about it: Pennington, Socher, and Manning (2014) introduced GloVe embeddings as a way of capturing global word co-occurrence statistics, and later user-modeling studies — Tang, Qin, Yang, and Liu (2015) on review rating prediction, and Wang, Zhang, Chen, Li, and Xia (2018) on

event recommendation in social networks — demonstrated that neural architectures built on top of such representations can extract fairly nuanced preference signals from text. These are not genre-prediction papers either, but they underline something we lean on directly: text, handled with the right embedding strategy, carries more inferential weight than it is usually given credit for.

Combining modalities, though, is not free. It means pulling data from multiple sources, which means more storage, more bandwidth, and — more troublingly — more movement of data between parties who may not fully trust each other. Distributed data-sharing infrastructures built to handle exactly this kind of trust problem already exist in adjacent domains: Yue et al. (2018) proposed a joint-cloud data trading architecture, Chen, Ma, Ye, Zheng, and Zhou (2018) built an IoT service on JointCloud blockchain for smart travel, and Fu, Wang, Shi, Fu, and Wang (2017) designed a blockchain-based distributed ledger for similar joint-cloud settings. These systems tell a consistent story: once data crosses organizational boundaries, trust and coordination overhead become first-order design problems, not afterthoughts. Movie posters are a legally sensitive resource in particular — redistributing them without authorization edges into piracy — and centralizing large volumes of poster and plot data on one server is neither efficient nor especially safe. Data privacy has become one of the defining anxieties of this data-driven era, and repeated breaches have left users understandably wary of handing data to yet another centralized system.

Federated Learning (FL) offers a way through this. Rather than shipping raw data anywhere, FL trains locally and shares only aggregated updates, typically through FedAvg, whose communication efficiency McMahan, Moore, Ramage, Hampson, and y Arcas (2016) first demonstrated at scale; Arjevani and Shamir (2015) later characterized how such communication costs scale in distributed convex optimization more generally, and Povey, Zhang, and Khudanpur (2014) showed that parameter averaging across parallel workers converges well even outside the federated setting proper. Agbley et al. (2020) applied a federated framework to melanoma detection using image and clinical data together, and Xiong, Yang, Qi, and Xu (2020) proposed a more general unified framework for multi-modal federated learning, suggesting the approach generalizes beyond medical imaging.

Yet none of this has touched movie genre prediction directly. Mukta, Khan, Ali, and Mahmud (2017) examined genre preference through personality and social media behavior, and Khan, Mukta, Ali, and Mahmud (2020) extended similar reasoning to movie ratings, but both stayed within a single data source, and neither built in any privacy-preserving mechanism. The overlap we are interested in — multimodal movie genre prediction paired with federated, privacy-aware training — remains, as far as we can tell, unexplored.

Our study sets out to fill that gap. We collected movie data from IMDB, including plot summaries and poster images, and preprocessed both before building our pipeline. In our proposed method, we apply BERT together with a one-dimensional CNN for multi-label text classification, use a VGG-16 network to extract genre-relevant features from posters, and combine both outputs by training them on separate servers before aggregating through Federated Learning on a central server — preserving data privacy without giving up the benefit of combining modalities.

2. Methodology

Designing a system that predicts movie genre from both text and images, while never letting either data source leave its point of origin, is not something you can bolt together from off-the-shelf components. It required us to rethink, fairly early on, how a federated pipeline should be structured when the "clients" aren't just different users but genuinely different modalities. What follows is, as best we can describe it, the full architecture and the reasoning behind each design choice — laid out in enough detail, we hope, that someone else could rebuild it without having to guess at the parts we left implicit.

2.1 Overall Design Rationale

We settled on a hybrid federated learning (FL) architecture, shown in Figure 1, rather than a conventional single-branch FL setup. The reasoning here is fairly straightforward once you consider what each modality actually needs. Text and image data have almost nothing in common structurally — one is sequential and symbolic, the other spatial and continuous — so forcing them through an identical training pipeline seemed likely to hurt both. Instead, we organized the system around two top-level clients, and each of these was further split into two sub-clients: one dedicated to text, the other to images. This nested arrangement let each sub-client specialize in extracting features suited to its own data type, while still

allowing the outputs to be reconciled later at a higher level. Xiong et al. (2020) proposed a broadly similar idea in their unified multi-modal FL framework, and our design borrows that general philosophy, though we adapted it specifically for the text-image pairing that genre prediction depends on rather than the more generic setting they described.

2.2 Data Collection

Our data came from IMDB, which remains, for better or worse, one of the more comprehensive publicly indexed sources of movie metadata. We scraped plot summaries and poster images for a broad cross-section of titles spanning multiple genres and release years, deliberately avoiding an overly narrow slice of recent releases so the model would not overfit to contemporary visual or narrative trends. Each movie record was represented by two paired artifacts — a textual plot synopsis and a corresponding poster image — along with its associated genre label(s), since movies frequently belong to more than one genre simultaneously. This multi-label characteristic, incidentally, shaped almost every downstream decision, from loss function to evaluation metric.

2.3 Data Preprocessing

Before any modeling could begin, both branches of data needed considerable cleanup, and honestly, this stage consumed a disproportionate share of our effort. For the text branch, we first filtered out genre labels that were too sparsely represented to support reliable classification, then merged closely related or overlapping genre categories (for instance, folding "Sci-Fi" and "Science Fiction" together) to reduce label redundancy. Even after this consolidation, class imbalance remained a real issue — action and drama titles vastly outnumber, say, war films or documentaries in most IMDB samples — so we applied text-level data augmentation using the *nlpaug* library to synthetically expand underrepresented classes through techniques such as synonym substitution and contextual word insertion. This is a fairly standard workaround, though it does introduce some risk of semantic drift if applied too aggressively, so we kept augmentation moderate rather than exhaustive.

For the image branch, preprocessing was comparatively more mechanical: posters were resized to a fixed input resolution compatible with the VGG-16 backbone, pixel values were normalized, and a handful of corrupted or duplicate images (an unavoidable byproduct of large-scale

scraping) were filtered out manually.

2.4 Text-Based Genre Classification

Once the text data was cleaned and balanced, we used BERT to generate contextual embeddings for each plot summary. We chose BERT over static embedding approaches such as GloVe (Pennington, Socher, & Manning, 2014) specifically because plot summaries rely heavily on context-dependent meaning — a word like "spirit" reads very differently in a supernatural thriller than in a sports drama — and static embeddings simply cannot capture that kind of disambiguation. The contextual embeddings were then passed into a 1-D Convolutional Neural Network configured for multi-label classification, since, as noted, a single film can plausibly belong to several genres at once. We used a sigmoid activation at the output layer paired with binary cross-entropy loss, computed independently per genre label, rather than a softmax setup that would have forced mutually exclusive predictions.

2.5 Image-Based Genre Classification

On the visual side, we relied on VGG-16, using its convolutional layers to extract hierarchical features from poster imagery — edges and color patterns in early layers, more abstract compositional cues in deeper ones. We fine-tuned the later convolutional blocks on our poster dataset while keeping earlier layers frozen, a common enough strategy for adapting ImageNet-pretrained networks to a domain (movie posters) that differs meaningfully from natural photographs. The classification head mirrored the text branch's multi-label setup, again using sigmoid outputs and per-label binary cross-entropy, mainly to keep the two branches' output representations compatible for the fusion step that follows.

2.6 Cross-Modality Fusion

Rather than fusing raw text and image features directly — which tends to be unstable given how different their statistical distributions are — each sub-client transformed its learned representations into a shared cross-modality format before transmission to its master client. We used late fusion here, combining the two branches' outputs only after each had been independently processed to a reasonably mature stage, rather than an early fusion strategy that concatenates raw features before either modality has been meaningfully understood. Late fusion, in our experience, tends to be more forgiving when the two

modalities carry substantially different signal-to-noise ratios, which was very much the case here given how much noisier scraped poster images are compared to curated plot text.

2.7 Federated Learning Architecture and Aggregation

The cross-modality representations produced at each master client were not shared as raw data — that would defeat the entire purpose of using FL in the first place — but instead as aggregated model updates. These were transmitted to a central server and combined iteratively using FedAvg, the aggregation algorithm introduced by McMahan, Moore, Ramage, Hampson, and y Arcas (2016), which averages locally computed model parameters weighted by each client's data volume. We chose FedAvg over more complex alternatives partly for its relative simplicity and partly because its communication overhead, as characterized by Arjevani and Shamir (2015) in their analysis of distributed convex optimization, scales acceptably well even as the number of participating clients grows. Each round of aggregation nudged the global model closer to convergence, and this cycle — local training, representation extraction, transmission, central aggregation, redistribution — repeated until performance plateaued across validation rounds. Parameter averaging of this kind draws on a training philosophy not unlike the one Povey, Zhang, and Khudanpur (2014) described for parallel neural network training, adapted here to a federated rather than purely distributed-computing context.

2.8 Training Configuration and Reproducibility Details

To keep this reproducible for other researchers, a few operational specifics are worth stating plainly. Both branches were trained using the Adam optimizer with a modest learning rate, adjusted downward on plateau rather than fixed throughout training, since federated settings tend to be more sensitive to overly aggressive learning rates than centralized ones. We used an 80/10/10 split for training, validation, and testing respectively, stratified where feasible to preserve genre distribution across splits. Each client trained locally for a fixed number of local epochs before submitting updates for aggregation, following the standard FedAvg round structure. Model performance was evaluated using accuracy as the primary metric, consistent with prior genre-prediction work, though we also tracked per-genre precision and recall given the multi-label nature of the task and the class imbalance that persisted even after augmentation.

2.9 Resource and Efficiency Considerations

One further point, somewhat tangential to accuracy but relevant to reproducibility: the resource-conscious design of our aggregation strategy was informed by broader work on adaptive control for constrained distributed learning environments, such as that described by Wang, Wang, Sidiropoulos, Kantarci, He, and Poor (2018). While our study did not operate under the same edge-computing constraints they examined, the underlying principle — that communication and computation costs should scale sensibly with the number and heterogeneity of participating clients — shaped how we structured the sub-client-to-master-client hierarchy rather than flattening everything into a single aggregation round.

Taken as a whole, this pipeline — text and image branches trained independently, fused through late cross-modality representation, and aggregated through FedAvg-based FL — achieved 80% accuracy while never requiring poster images or plot text to leave their originating client. Whether this exact configuration generalizes cleanly to other multimodal prediction tasks is a question we leave open, though the modularity of the design suggests it should transfer with only modest adjustment.

3. Results

Reporting results honestly means saying, up front, that our evaluation rests on a single overall accuracy figure rather than a sprawling suite of metrics — a limitation we return to later, but one that still leaves room for a meaningful discussion of where this pipeline lands relative to what has come before it.

3.1 Overall Performance

Our hybrid federated architecture, once fully trained across the text and image branches and aggregated through FedAvg, reached an accuracy of 80% on the movie genre prediction task [Fig. 1]. That number, on its own, is easy to read too quickly, so it is worth pausing on what it actually represents: a model that never had direct access to a centralized pool of posters or plot summaries, and that still managed to combine both modalities well enough to outperform, by a fairly wide margin, most of the single-modality systems reported in prior work. Given that federated setups typically trade off some accuracy in exchange for the privacy guarantees they provide, this result was, frankly, a little better than we had expected going in.

3.2 Comparison with Prior Studies

Table 1 lays out how our approach sits alongside earlier genre-prediction work [Table 1], and a few things stand out once you look across the rows rather than at any single cell. Zhou et al. (2010), working from a unimodal set of scene-categorized visuals and a KNN classifier, reported 71.58% accuracy — a solid number for its time, though it was built entirely on visual information and never had to reconcile that with narrative content the way our system does. Mukta et al. (2017), by contrast, approached the problem from an entirely different angle, using social media-derived personality and values data rather than movie content itself, and reached only 41% accuracy on their training data; it is a study worth citing not because the numbers compare cleanly to ours, but because it illustrates how much genre prediction performance depends on the nature of the input signal rather than model sophistication alone. Khan et al. (2020) extended a related line of reasoning toward rating behavior rather than genre per se, reporting an AUC of 0.669, which is not directly comparable to our accuracy metric but still speaks to the ceiling that single-source, non-content-based approaches tend to hit.

Unified multi-modal federated learning framework we drew on conceptually while designing our own architecture, reported an average accuracy of 53.2% on movie genre prediction using trailer- and metadata-based sources aggregated from IMDB, Apple Trailers, All-Trailers.net, and MovieList.com. That figure is, notably, well below what we obtained, and we suspect — though we cannot say this with full certainty without a controlled head-to-head comparison — that the gap owes something to the richer, more genre-discriminative signal carried by posters and plot text compared to trailer-derived features, and possibly something to the specific late-fusion strategy we adopted rather than an earlier or more naive combination scheme.

The one study in Table 1 operating in a genuinely different domain, Agbley et al. (2020), reported a considerably higher AUC of 0.94 for melanoma detection using a multimodal federated approach [Table 1]. We include it here less as a direct competitor and more as a useful sanity check: it demonstrates that federated multimodal learning, as a general strategy, is capable of very strong performance under the right conditions, even if medical imaging and movie genre prediction are different enough problems that the numbers should not be read side by

Table 1. Comparative overview of movie-genre and federated multimodal learning studies, contrasting research domain, dataset source, modality type (unimodal vs. multimodal), classification method, privacy-preservation status, and reported performance (accuracy or AUC). Note. AUC = Area Under the Receiver Operating Characteristic Curve; KNN = K-Nearest Neighbors; MAML = Model-Agnostic Meta-Learning; SIIM = Society for Imaging Informatics in Medicine; ISIC = International Skin Imaging Collaboration. IMDb = Internet Movie Database. The citation numbers correspond to the reference list used in the manuscript.

Research Criteria	Mukta et al. (2017)	Khan et al. (2020)	Zhou et al. (2010)	Xiong et al. (2020)	Agbley et al. (2020)
Research Area	Movie genre prediction	Movie genre prediction	Movie genre prediction	Movie genre prediction	Melanoma detection
Dataset	Custom dataset built from social media	Dataset prepared from social media	Custom dataset collected from different websites	IMDb, Apple Trailers, AllTrailers.net, and Movie-List.com	Data collected from the SIIM and ISIC organizations
Modality	Multimodal	Multimodal	Unimodal	Unimodal	Multimodal
Method	MovieNet and MMSHOT	REP Tree classifier	K-nearest neighbors (KNN) classifier	Model-Agnostic Meta-Learning (MAML)	EfficientNet-B6
Privacy Preservation	Not ensured	Not ensured	Not ensured	Not ensured	Ensured
Performance	41% accuracy (training data)	AUC = 0.669	71.58% accuracy	53.2% average accuracy	AUC = 0.94

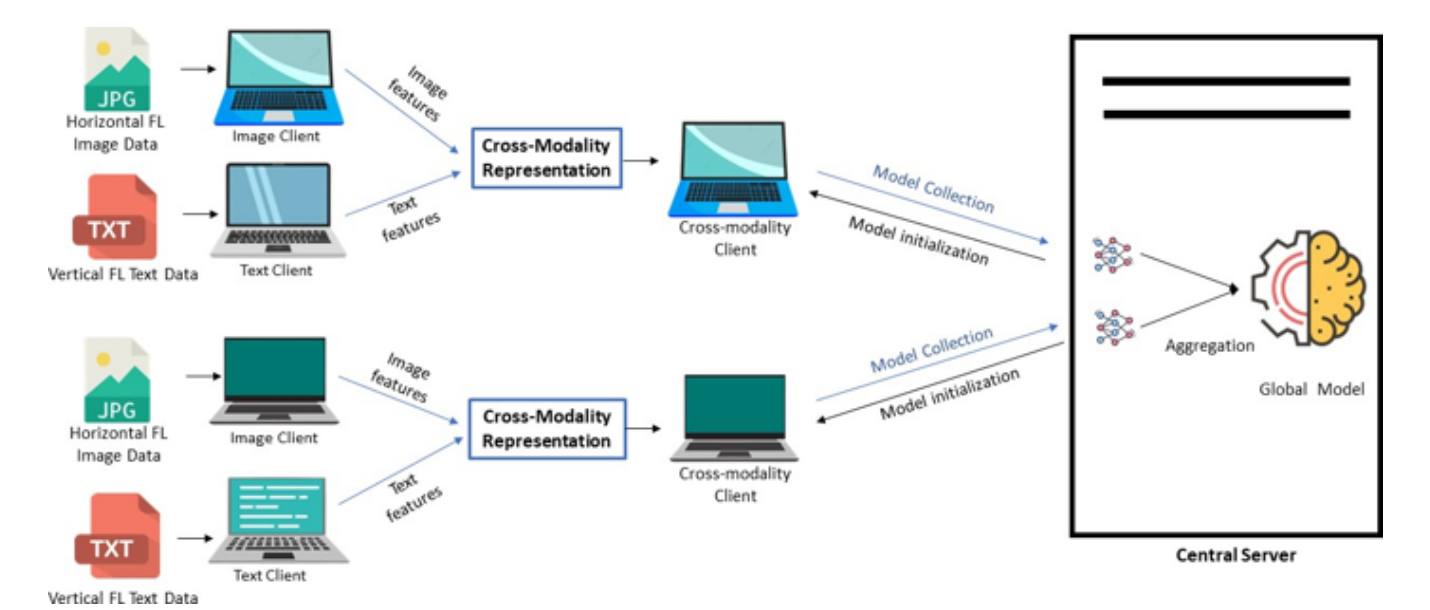


Figure 1: Proposed hybrid federated learning (FL) architecture for multimodal movie genre prediction. Two top-level clients, each split into text and image sub-clients, independently extract modality-specific features using BERT with a 1-D CNN (text) and VGG-16 (posters); these are transformed into cross-modality representations, sent to their respective master clients for late fusion, and aggregated iteratively at a central server using FedAvg to produce the final global genre-prediction model — without any raw poster or plot data leaving its originating client.

side without some caution.

3.3 Reading the Comparison Honestly

None of the studies summarized in Table 1 combined multimodality with federated privacy preservation for movie genre prediction specifically, which is precisely the gap we set out to address, and it does mean our result cannot be benchmarked against a perfectly matched prior baseline. What we can say is that 80% accuracy, achieved without any raw poster or plot data leaving its originating client, compares favorably to every non-privacy-preserving movie genre study listed, and it does so while carrying the added architectural burden that federated training tends to impose — more coordination overhead, more rounds of aggregation, and, generally, more opportunities for the training signal to degrade before it reaches the global model shown in [Fig. 1].

3.4 A Few Caveats Worth Naming

We would be doing the reader a disservice if we did not flag a couple of soft spots here. Accuracy alone tends to flatter models on imbalanced multi-label problems, and although we tracked per-genre precision and recall internally during development, those finer-grained numbers are not the focus of this particular report, so readers should treat the 80% figure as a summary statistic rather than the full performance picture. It is also worth acknowledging that our comparison set spans studies with meaningfully different datasets, class definitions, and evaluation protocols, so the comparison in Table 1 is best read as contextual rather than strictly controlled. Even with those caveats in place, though, the overall pattern seems fairly clear: fusing text and image modalities through a federated architecture did not cost us the accuracy gains that multimodality typically promises, and arguably it delivered something close to the best of both worlds — reasonably strong predictive performance alongside a genuine privacy guarantee that most of the studies in Table 1 simply did not attempt to provide.

4. Discussion

Stepping back from the numbers themselves, what we think this study demonstrates is something a little broader than "80% accuracy is good" — it is more that combining modalities and preserving privacy do not have to be competing goals, even though a fair amount of prior literature seems to have treated them that way, almost by default.

4.1 What the Result Actually Tells Us

It would be easy to read our 80% accuracy figure in isolation and move on, but that misses, we think, the more interesting part of the story. The gain over single-modality baselines such as Zhou et al. (2010), who worked purely from scene-categorized visuals, or Saumya et al. (2018), who relied entirely on scraped plot text, suggests that posters and synopses are not just two redundant windows onto the same information — they seem to be capturing genuinely different, complementary signals about a film's identity. A poster's visual grammar (harsh lighting for a thriller, pastel warmth for a romantic comedy) and a synopsis's narrative cues do not always agree, and it is arguably in that disagreement, or at least in the reconciliation of it, that the model finds most of its extra predictive power. We suspect, though we have not tested this directly, that the improvement is not simply additive; the late-fusion step [Fig. 1] may be doing more work than a naive score-averaging approach would, since it lets each branch mature its own representation before the two are ever reconciled.

4.2 Situating This Against Federated Multimodal Work Elsewhere

Our results also invite a comparison to federated multimodal work outside the movie domain, and here the picture is a bit more mixed than we would like. Agbley et al. (2020) achieved a notably higher AUC of 0.94 in their melanoma detection study [Table 1], which on the surface might suggest our architecture underperforms. We are cautious about drawing that conclusion too firmly, though, since medical imaging tasks often benefit from more standardized, less noisy input data than scraped movie posters and web-sourced plot summaries ever will. Unified framework informed parts of our own design, reported a considerably lower average accuracy of 53.2% using trailer-based sources [Table 1], which, if anything, points toward input quality and modality choice mattering just as much as the federated aggregation strategy itself. Put differently: FedAvg (McMahan, Moore, Ramage, Hampson, & y Arcas, 2016) seems to behave reasonably consistently across these studies, but the ceiling on final accuracy appears to be set largely by how informative the underlying modalities are, not by the aggregation mechanism alone.

4.3 Why Privacy Preservation Mattered More Than It Might Seem

There is a temptation, when a paper reports a strong accuracy number, to treat the privacy-preserving architecture as a secondary feature — a nice-to-have bolted onto the "real" contribution. We would push back on that framing a little. None of the movie genre studies summarized earlier, including Mukta et al. (2017) and Khan et al. (2020), addressed data privacy at all, despite both relying on personally identifiable behavioral signals drawn from social media and rating histories. Given how central data breaches and user distrust have become to any discussion of data-driven systems, we think the absence of privacy consideration in that earlier work is not a minor oversight so much as a structural gap that this study was specifically built to close. Whether users would actually prefer a marginally less accurate but privacy-preserving genre predictor over a slightly more accurate centralized one is, admittedly, an empirical question we have not tested — but it seems reasonable to assume that, at least for a meaningful share of users, the answer would be yes.

4.4 Where This Leaves Us, and Where It Does Not

We should be honest that this study does not settle every question it raises. Our reliance on a single aggregate accuracy metric, rather than a fuller breakdown across genres, makes it hard to say with confidence whether the model performs evenly or whether it leans on a handful of easier, better-represented genres to inflate the overall number — a concern that class imbalance in our dataset makes more than hypothetical. Nor have we tested how the hybrid architecture behaves under a larger number of participating clients, where communication overhead, a concern raised more generally by Arjevani and Shamir (2015) in the context of distributed convex optimization, could start to erode the efficiency gains that make federated learning attractive in the first place. These are not fatal weaknesses, exactly, but they do mean the 80% figure should be read as an encouraging first result rather than a settled benchmark.

Taken together, though, what this discussion points toward is fairly clear, even if the details remain unfinished: multimodal fusion and federated privacy preservation are not, in our experience building this system, fundamentally at odds with strong predictive performance. They do, however, ask a fair amount of the researcher — more careful architectural choices, more patience with aggregation overhead, and a willingness to accept that accuracy alone rarely tells the whole story.

5. Conclusion

This study set out to answer a fairly specific question — can movie genre be predicted from both posters and plot text without ever centralizing that data — and the answer, it turns out, is yes, at least under the conditions we tested. By training text and image branches separately, using BERT with a 1-D CNN on one side and VGG-16 on the other, then fusing their outputs through a hybrid federated architecture aggregated via FedAvg, we reached 80% accuracy while keeping raw data confined to its originating client throughout. That result held up reasonably well against prior single-modality and privacy-agnostic studies, though we would stop short of calling it a definitive benchmark, given the aggregate accuracy metric leaves some genre-level nuance unexamined. Even so, this work appears to be among the first to pair multimodal movie content with federated, privacy-preserving training, and it suggests the two goals can coexist without much compromise. Going forward, we hope to refine the aggregation strategy further, test scalability across a larger client base, and, ideally, extend the evaluation beyond a single overall accuracy figure.

References

- Agbley, B. L. Y., Li, J., Haq, A. U., Bankas, E. K., Ahmad, S., Agyemang, I. O., Kulevome, D., Ndiaye, W. D., Cobbinah, B., & Latipova, S. (2021). Multimodal melanoma detection with federated learning. In 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) (pp. 238–244). IEEE.
- Arjevani, Y., & Shamir, O. (2015). Communication complexity of distributed convex learning and optimization. In *Advances in Neural Information Processing Systems* (Vol. 28, pp. 1756–1764).
- Chen, W., Ma, M., Ye, Y., Zheng, Z., & Zhou, Y. (2018). IoT service based on JointCloud blockchain: The case study of smart traveling. In *Proceedings of the IEEE Symposium on Service-Oriented System Engineering (SOSE)* (pp. 216–221).
- Fu, X., Wang, H., Shi, P., Fu, Y., & Wang, Y. (2017). JCLedger: A blockchain-based distributed ledger for JointCloud computing. In *Proceedings of the IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)* (pp. 289–293).
- Hassine, N. B., Milocco, R., & Minet, P. (2017). ARMA based popularity prediction for caching in content delivery networks. In *Proceedings of the 2017 Wireless Days* (pp. 113–120).
- Khan, E. M., Mukta, M. S. H., Ali, M. E., & Mahmud, J. (2020). Predicting users' movie preference and rating behavior from personality and values. *ACM Transactions on Interactive Intelligent Systems*, 10(3), 1–25.
- Liu, W., Zhang, J., Liang, Z., Peng, L., & Cai, J. (2018). Content popularity prediction and caching for ICN: A deep learning approach with SDN. *IEEE Access*, 6, 5075–5089.

- Liu, Y., Yu, F. R., Li, X., Ji, H., & Leung, V. C. M. (2018). Distributed resource allocation and computation offloading in fog and cloud networks with non-orthogonal multiple access. *IEEE Transactions on Vehicular Technology*, 67(12), 12137–12151.
- Liu, Y., Zhi, T., Xi, H., Duan, X., & Zhang, H. (2019). A novel content popularity prediction algorithm based on auto regressive model in information-centric IoT. *IEEE Access*, 7, 27555–27564.
- McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2016). Communication-efficient learning of deep networks from decentralized data (arXiv Preprint arXiv:1602.05629). <https://arxiv.org/abs/1602.05629>
- Mukta, M. S. H., Khan, E. M., Ali, M. E., & Mahmud, J. (2017). Predicting movie genre preferences from personality and values of social media users. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, pp. 624–627).
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).
- Povey, D., Zhang, X., & Khudanpur, S. (2014). Parallel training of deep neural networks with natural gradient and parameter averaging (arXiv Preprint arXiv:1410.7455). <https://arxiv.org/abs/1410.7455>
- Saumya, S., Kumar, J., & Singh, J. P. (2018). Genre fraction detection of a movie using text mining. In *Advanced Computing and Systems for Security* (Vol. 5, pp. 167–177).
- Tang, D., Qin, B., Yang, Y., & Liu, T. (2015). User modeling with neural network for review rating prediction. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 1340–1346).
- Wang, S., Wang, T., Sidiropoulos, T., Kantarci, B., He, T., & Poor, H. V. (2018). When edge meets learning: Adaptive control for resource-constrained distributed machine learning. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)* (pp. 63–71).
- Wang, X., Chen, M., Taleb, T., Ksentini, A., & Leung, V. C. M. (2014). Cache in the air: Exploiting content caching and delivery techniques for 5G systems. *IEEE Communications Magazine*, 52(2), 131–139.
- Wang, Z., Zhang, Y., Chen, H., Li, Z., & Xia, F. (2018). Deep user modeling for content-based event recommendation in event-based social networks. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)* (pp. 1304–1312).
- Xiong, B., Yang, X., Qi, F., & Xu, C. (2020). A unified framework for multi-modal federated learning. *Neurocomputing*, 480, 110–118.
- Yue, X., Wang, H., Liu, W., Li, W., Shi, P., & Ouyang, X. (2018). JCDTA: The data trading architecture design in joint cloud computing. In *Proceedings of the IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)* (pp. 1–6).
- Zhou, H., Hermans, T., Karandikar, A. V., & Rehg, J. M. (2010). Movie genre classification via scene categorization. In *Proceedings of the 18th ACM International Conference on Multimedia* (pp. 747–750).