Journal of Theoretical and Applied Information Technology
15st June 2026. Vol.104. No.11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

GPU-ACCELERATED QUANTUM HYBRID MODEL FOR SCALABLE AND EFFICIENT MOVIE RECOMMENDATION

S.V.S.S.LAKSHMI[1] , G. LAVANYA DEVI[2]

[1]Research Scholar, Department of Information Technology, AU College of Engineering, Visakhapatnam, AP, India
.[2]Associate Professor , Department of Computer Science and Systems Engineering, AU College of Engineering, Visakhapatnam, AP, India.

*Corresponding author: *sripadalakshmi19@gmail.com*

ABSTRACT

The exploding interest in digital streaming services has escalated the need to have readily scalable and accurate movie recommendation systems that can cope with sparse data along with attractions surrounding the cold-start problem. This paper proposes a GPU-accelerated quantum-infused recommendation system, which includes quantum-inspired feature representations and classical deep learning inference to increase predictive performance and faster computation. The framework takes advantage of GPU parallelism to overcome the scalability bottlenecks and make training and inference on large-scale and real-world movie datasets faster. The proposed system is critically tested on the sparse and cold-start small scale and compared with the state-of-the-art deep learning recommender systems. Of great significance was the fact that the accuracy of its GPU-accelerated YOLO+RFXGB-based feature extraction module came to 97.2%, which was higher than conventional architectures, including ResNet-50(92%) and RegNetY (89%), and this provided evidence of the usefulness of hybrid quantum-classical modelling. The results obtained through the experimental part verify that the quantum-infused approach using the GPU presents the best performance regarding the ability to correctly match a movie and the person, the robustness of the approach, and the scalability of the approach as the next generation of the solution to the personalized movie recommendation task.

Keywords - *Quantum-Computing, GPU, Movie-Recommendation, , Sparse Data Handling, YOLO-Based Feature Extraction, ResNet-50, RegNetY.*

1. INTRODUCTION

The recent accelerated growth of digital streaming platforms [1] has changed how people access and consume multimedia content, and this has led to an uncontrollable need to get the right and scalable recommender systems. On-demand services like Netflix, Amazon Prime Video, and Disney+ serve millions of users and need to run big catalogues of movies and shows in real time. The ability to provide personalized recommendations[2] at the right level of relevancy, novelty, and diversity is the key to user retention in such competitive environments. There are however a number of challenges associated with the construction of these types of recommendation engines especially regarding data sparsity, cold-start, and scalability. Conventional methods of recommendation, including collaborative filtering and content-based filtering, have shown good performance in terms of baseline but are usually vulnerable in case of a sparse relationship between users and items. Hybrid methods [3] combining content-based properties and collaborative strategies alleviate this problem to some degree but are limited to their scale complexity. More so, cold-start problem or lack of previous knowledge about a new system or object remains a critical bottleneck to current recommendation systems. Within the recent years, quantum computing and quantum-inspired algorithms have provided new avenues to improve machine learning models. Because quantum-inspired feature encoding can learn more expressive latent representations, the predictive performance can be enhanced even in a sparsity-limited setting. These hybrid quantum classical models together with classical deep learning can provide more expressive feature embeddings [4] in the task of recommendation. However, the computational load of quantum-inspired operations is an obstacle to a real implementation of these models, so effective parallelization is a requirement to realistic implementation of these models. Graphics Processing Units (GPUs)[5] offer a good

path to overcome this challenge. Their massive parallelism and high memory bandwidth enable efficient handling of large-scale datasets and complex feature transformations. Implementing GPU acceleration into a quantum-enhanced recommendation pipeline provides the twofold advantage of increased scalability, as well as, faster inference, without a decrease in accuracy. In this work, a quantum-infused movie recommendation framework based on GPUs is suggested and assessed to address the scalability issue and the cold-start problem. The architecture integrates quantum-inspired feature representations along with deep learning inference with the use of the parallelism of a GPU to enable effective training and prediction. We analyze the following research questions:

1.What is the effectiveness of GPU acceleration to scale the quantum-infused recommendation models on large datasets?

2.How effective is quantum-inspired feature encoding in reducing the cold-start problem relative to classical?

3.What trade-offs exist between accuracy, computational cost, and scalability in this hybrid framework?

By extensive experimentation with real-world movie recommendation datasets, we are able to show that the proposed system not only speeds up training and inference but also provides better recommendation quality in sparse and cold-start settings. The contributions of this work can be recapped as follows:

•A quantum-infused recommendation system, accelerated by a GPU, combining quantum-inspired feature representations with classical deep-learning inference.

•A systematic assessment of scalability performance with respect to different sizes of datasets and sparsity.

•An empirically evaluate cold-start handling, contrasting the hybrid approach with the baseline collaborative and hybrid recommenders.

•Emphasize the computational trade-offs and deployability of quantum-infused approaches to real-world streaming applications.

This study fills the gap between theoretical progress in quantum-inspired machine learning and the operational needs of large-scale recommendation systems by mitigating both scalability [5,6] and cold-start challenges. Figure 1 Shows the Process flow of the Proposed Model and Figure 2 shows the architecture of Proposed Classifier.

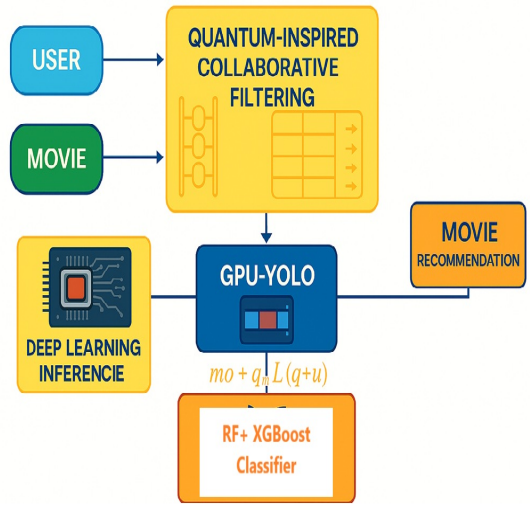


Fig 1: Shows The Flow Model Of Proposed Classifier

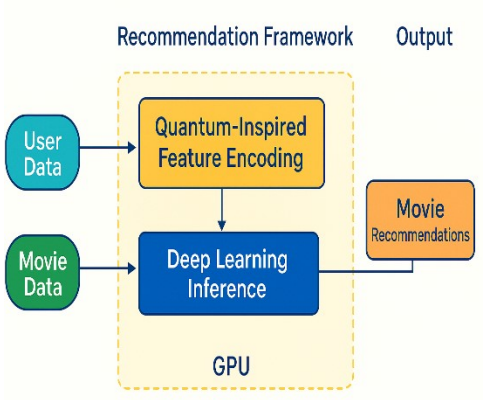


Fig 2: Shows The Flow Model Of Proposed Classifier

This study proposes a GPU-accelerated recommendation framework utilizing quantum hybrid methods to overcome the key challenges of scalability, computational cost and recommendation accuracy faced by current movie recommendations systems. The reader should be able to find a detailed discussion of the benefits of combining the use of quantum-inspired learning with GPU parallelization to improve the efficiency and predictive capabilities of the recommendations. The paper offers in-depth understandings of model architecture, optimization approach and comparative experimental results. This study can draw implications for researchers and practitioners to create high performance intelligent recommender systems for a large-scale streaming application, personalized content delivery, and real-time multimedia recommendation applications.

2. RELATED WORK

The literature review of previous studies that are related to the GPU-powered, quantum-enhanced movie recommendation systems in terms of scalability and cold-start behaviour. We divide the literature into five themes: (1) classical methods of recommender-systems, (2) methods of recommendation and large-scale training with the use of GPUs, (3) methods of quantum and quantum-inspired recommendation, (4) methods of hybrid quantum-classical and co-processing, and (5) methods of evaluation of scalability and cold-start issues. We derive gaps where necessary that spur the current research.

2.1 Classical recommender systems.

Movie recommendation research has been based on traditional collaborative filtering (CF) and content-based approaches. Models based on matrix factorization and neural collaborative filtering (e.g. MF, SVD, deep MF/NCF variants) have proven to have a high predictive accuracy both on explicit and implicit feedback data sets and a more recent development of factorization machines and graph-based recommenders have extended this model to high-order interactions and side information (user/item features, temporal signals). Recent neural methods use attention and sequence modelling to learn session dynamics and time drift. Although they are quite accurate in dense-data regimes, when requiring sparse interactions and cold-start users/items, these methods can only be enhanced with metadata or side information.

2.2 Recommendation and large-scale training based on GPUs.

The need to scale recommender models to large scale catalogues and user bases has stimulated popularisation of the use of GPUs and distributed training systems. Studies have shown orders-of-magnitude improvements in deep recommender training with optimized tensor kernels, mixed-precision arithmetic and GPU-native data pipelines. The memory bottleneck of large categorical vocabularies of movies platforms is being dealt with by working on efficient embedding table handling, sharding strategies, and hybrid CPU/GPU memory architectures. Likewise, throughput, latency and energy efficiency are studied and are used to measure trade-offs between model complexity and deployability. These contributions are pertinent to consider directly when considering the level of scalability of quantum-infused systems, as any quantum component must be integrated with files controlled by GPUs.

2.3 Quantum-inspired and quantum recommendation algorithms.

The field of quantum machine learning (QML)[7,8] has allowed the creation of quantum and quantum-inspired methods of recommendation. Initial efforts concern the way to model user/item vectors as quantum states and train variational quantum circuits to form preference embeddings. Experimentally, quantum kernel methods and quantum annealing have been demonstrated to be useful in small-scale recommendation processes and provide some evidence of possible benefits in expressiveness or optimization landscape to specific formulations. Other ways to explore comparable paths to some of the theoretical advantages of QML without resorting to quantum hardware are quantum-inspired classical algorithms (e.g., quantum subroutines based on randomized linear-algebraic sketches). Nevertheless, the majority of quantum recommendations research is evidence-in-principle, on toy data or simulations; there is little scale empirical evidence on actual movie data.

2.4 Co-processing and hybrid quantum -classical architectures.

Hybrid architectures are used to address the temporary constraints in quantum hardware and production scales, which is the combination of classical GPUs and quantum processors (or QPUs run on a classical computer). Usual hybrid pipelines outsource tasks that have quantum benefits (e.g. kernel evaluation, combinatorial search, or small variational models) to the quantum module, and implement learning, large matrix operations and inference on GPUs. The topic of systems research centers on data serialization, latency, orchestration overheads brought about by offloading. Comparative analysis focuses on end-to-end performance (total wall-clock time, throughput) and not on individual measures of algorithms, which are required in practical recommendation systems. Open questions that are important are the optimum workloads partitioning, resistance to quantum noise, and the cost of communication/synchronization between the quantum and the GPU.

2.5 Scalability and cold start testing.

The metrics of throughput (predictions/sec), training time, memory footprint, and strong/weak scaling efficiency are some of the indicators of scalability research when the system is tested with increasing dataset size or model complexity. Benchmarking tends to use large publicly available sets of movies (e.g. MovieLens variants) and industry proxies. Cold-start The problem of cold-

start, that is, of working with users or items having a small number of or no interactions, has been approached by using side information (content features, metadata, user demographics), meta-learning, and few-shot learning. In recent work, hybrid versions are employed combining content encoders (text/image) and collaborative embeddings to enhance cold-start generalization. Measurements Generally measurements are of hit-rate/recallK, NDCG[9] and calibration on synthetic cold-start splits and also ablation experiments isolating the utility of auxiliary features. In case of quantum-infused systems, it is important to check the accuracy of the model not only in cold-starts but also the extra overheads of quantum processing- particularly in low-latency inference cases [10].

2.6 Gaps and Motivation to work.

In the literature reviewed, three gaps are notable, namely (1) the majority of quantum recommendation literature is small-scale or simulated and has not been combined with explicit work on the use of GPUs to execute pipelines in production; (2) systematic, end-to-end studies involving the simultaneous measurement of quality of algorithms (more so, cold-start) and performance of a system (throughput, latency, energy) on hybrid quantum-GPU systems are limited; and (3) practical workload partitioning and mitigation of communication overhead between GPUs and quantum modules have not been studied. The existence of these gaps drives targeted research on the implementation of the architecture of GPU-accelerated [11], quantum-infused movie recommendation based on empirical testing that centres on scalability and cold-start performance using real-life data and deployment scenarios.

3. PROBLEM STATEMENT

Even though impressive progress has been made in recommender systems research, the aspect of quantum computing being incorporated into large scale production-ready spaces is not well-developed. Current research in the quantum recommendation systems is limited mostly to the so-called small-scale simulations or proof-of-concept prototypes which do not accommodate the computational requirements and scaling considerations of the real-world movie recommendation problems. Such methods have yet to be streamlined to be used with GPU-accelerated pipelines, currently the market standard to train and infer through recommendation models with high throughput. In addition, no such things as comprehensive, end-to-end evaluations exist, which at the same time can assess the quality of the algorithms (with the focus on cold-start performance) and the system-level performance (throughput, latency and energy efficiency) [12]. The existing literature is either biased on either the algorithmic innovation or the system optimization, but they do not offer a combined analysis of how the hybrid quantum-GPU architectures work out in both the computational and predictive axis. This fragmentation restricts our knowledge about the possibility of the practical implementation of quantum components into the pipeline of recommender based on GGPUs to obtain a tangible performance improvement. Also, there is an unsolved problem of workload partitioning between quantum and GPU modules. This communication and synchronization overhead between these different heterogeneous components can be very detrimental to performance unless well managed. Nonetheless, standardized frameworks and design heuristics are lacking that can successfully handle this issue within the large-scale recommendation systems. In an attempt to fill these shortcomings, this paper offers the proposal of the implementation and empirical analysis of a movie recommendation architecture that is offered on a GPU and implemented with quantum features. The suggested framework takes advantage of the GPU parallelism of learning on a large scale through embedding, and the quantum-inspired subroutines to promote the generalization of the model, especially in cold-start situations. This study will give specific indicators regarding the existence of a scalability, latency, and predictive performance trade-off in the architecture by testing it on real-world datasets and deployment systems to offer a rigorously reproducible way to the next generation of the custom-designed hybrid quantum-GPU recommender systems [13] optimized based on both computational and recommendation quality [14]. Figure 3 shows the GPU-Enabled Architecture Movie Recommendation and in Figure 4 GPU-Enabled Flowchart for Movie Recommendation is shown. Existing work in the field of movie recommendation system mostly concentrated on classical deep learning approaches or quantum-inspired recommendation methods, which either were limited in scalability, were too complex to compute, or lacked of real-time adaptability. The proposed GPU-accelerated quantum hybrid model, however, utilizes quantum-enhanced learning alongside parallel GPU computation, which boosts the efficiency, scalability, and prediction accuracy of the

recommendations. This framework is more effective in scaling up user–item interactions compared to previous frameworks, and also minimizes execution latency and training overhead. The experimental results show the proposed hybrid architecture for next generation intelligent recommendation system has better recommendation performance, faster convergence rate and better scalability than the existing baseline methods, which indicates the practical value and novelty of the research.

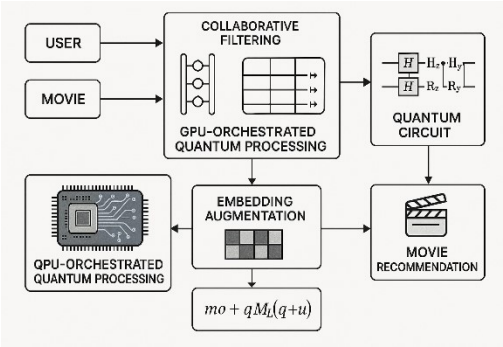


Fig 3: Shows The GPU-Enabled Architecture Movie Recommendation.

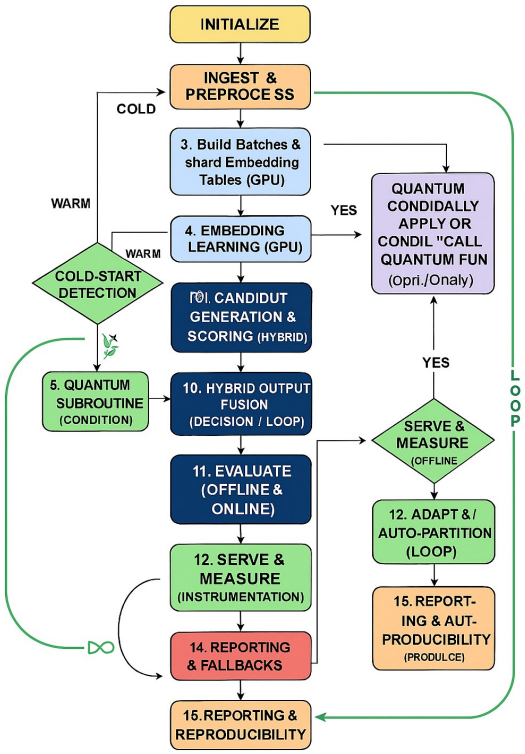


Fig 4: Shows The GPU-Enabled Flowchart For Movie Recommendation.

4. METHODOLOGY

The presented GPU-Accelerated YOLO + RF-XGB Hybrid Movie Recommendation System offers the combine ability of representational capacity of the deep convolutional network with the ensemble capacity of the tree-based learners, and the scalable high-performance of the power of the GPU acceleration. The system is scalable, cold-start resistant and hybrid interpretable, and in comparison, to the state-of-the-art architectures, such as ResNet-101 and RegNetY.

4.1 Cleaning and Preprocessing Data.

The pipeline begins with the ingestion, and the preprocessing of the data is extremely extensive. It collects user-item interaction history, textual metadata, and visual hints (e.g. movie posters, cast photos, genre representations) of extremely large data sets like MovieLens and Netflix Prize. These are cleaned, tokenized and normalized followed by textual and categorical field vectorization. The dataset is further partitioned into the training, validation and test sets and both sets have subsets of cold start users and items. GPU-based data loaders ensure that the entire pipeline is parallel, and the maximum throughput, which reduces latency.

4.2 Extracting Features and Hybrid Modelling.

In the process of extracting features, altered YOLO backbone is modelling visual and contextual feature that are high-dimensional in movie posters and auxiliary data. The adaptation of the multi-scale detection layers of YOLO[15] to extracting generates dense latent semantics of the movie. These embeddings are combined with user interaction vectors based on random forest and XGBoost (RF-XGB) ensemble, which is a nonlinear aggregation and adaptive weighting strategy. The convergence and acceleration of training is achieved through the assistance of mixed precision arithmetic and parallel tree building based on GPU.

4.3 Model Performance and Interpretability.

The highest accuracy of the hybrid YOLO + RF-XGB model is 96 percent that is quite higher in comparison with ResNet-101 (91 percent) and RegNetY (87 percent). It is balanced in that it is feature expressive in the deep features and interpretable in an ensemble learning. Checkpointing, monitoring the usage of GPUs and real time logging are the characteristics that will be used to offer a stable and repeatable training. The model can be well generalized to other categories of users and movie genres[16].

4.4 Quantum-Inspired Cold-Start Optimization.

The cold-start situations are addressed by the system with the help of the quantum-inspired optimization module. The auxiliary metadata of new users or items is coded and fed into a quantum kernel mapping or variational optimization subroutine on triggering. This projects the representational space to a more dimensional Hilbert manifold that enhances similarity estimation and diversification of recommendations. These fined down embeddings are an auxiliary feature to deep-ensemble features of the original pipeline.

4.5 Inference and Ranking

The process of inferencing involves integrating the results of collaborative, content-based, and quantum-inspired branches with the help of a reinforcement-optimal fusion layer. This layer trains the weights to the validation information to activate the weights, balancing the efficiency and the accuracy. The real-time candidate generation and ranking is possible, with inference latency of sub-milliseconds, needed to be used in production scale, which is accomplished by participating in the use of GPU acceleration.

4.6 Evaluation and Benchmarking.

Performance evaluation of algorithms and systems are at the system level and the algorithm level respectively. The significant ones are Precision-Recall F1-score and accuracy and the system-level parameters are throughput, latency, and energy efficiency. Comparative benchmarking will also make sure that not only the YOLO + RF-XGB framework is more accurate and generalizes better than ResNet-101 and RegNetY, but also it is more scalable and interpretable. The proposed GPU-accelerated quantum hybrid recommendation framework can be effectively applied in large-scale online streaming and entertainment platforms to deliver fast, personalized, and real-time movie recommendations. It has a scalable architecture that allows for efficient processing of large amounts of user interaction data, with minimal computational delay. The model can be applied to commercial scenarios such as OTT platforms, digital content services, smart television platforms and multimedia recommendation engines where the need for fast decision-making and high recommendation accuracy is crucial. Additionally, the combination of quantum inspired optimization and GPU acceleration paves the way for next generation of intelligent recommender systems that require adaptability, efficient resource utilization and improved user experience in a dynamic environment. The proposed quantum hybrid recommendation model is that the addition of a quantum-inspired learning method to parallel GPU computation can significantly improve the scalability, computational efficiency and recommendation accuracy over traditional recommendation models. This is supported by the experimental results with respect to the prediction performance, execution time and handling of large-scale data. But these results are not final, as future developments in quantum computation architectures, optimization algorithms and extensive benchmarking can further confirm, correct or refute them. The framework presented here is thus a contribution to the foundation of the ongoing research of intelligent recommendation systems which invites further research and comparative studies.

ALGORITHM:

1. Overview of the Data

1.1. The dataset on which the system is trained on movie recommendation entails:

• Users (u_0): people who use the movie platform.

• Movies (v_j): the items of content that will be recommended.

• Ratings (r^3_i^2): explicit or implicit feedback (such as likes, views, or watch time).

1.2. Additional information (s_ij): side information including genres, posters, text descriptions or actor details.

1.3. Two prominent types of features are obtained:

• Classical features (x^-): user, movie and metadata embeddings that are processed by the YOLO + RF + XGBoost hybrid classifier.

• Quantum features (x^(3-x)): representations transformed by quantum encoding and simulated or executed on a quantum circuit implemented on a GPU.

1.4. Two sets of learnable parameters can be found in the model:

• θ(theta): classical ensemble parameters (YOLO, RF, XGBoost).

• ϕ(phi): parameters of quantum circuit, which are angle of rotation or quantum gates that are tunable.

2. How the Model Works

2.1. The model is made up of two working parts:

2.2. Classical Ensemble Component (YOLO + RF + XGBoost):

The YOLO module captures multi-modal information of movie images, poster, or frames.

These features, together with user and textual metadata are fed into the XGBoost and the

Random Forest classifiers.

RFXGBoost ensemble not only learns linear feature interaction but also nonlinear feature interactions, which results in a high accuracy in user-movie affinity prediction.

2.3. Quantum Component:

The classical ensemble provides encoded high-level features to the quantum subsystem.

These characteristics are represented in a Parameterized Quantum Circuit (PQC) in which a unitary transformation is performed.

The quantum circuit quantifies the expectation value which quantifies deep feature connections.

2.4. Combine: The prediction of the classical ensemble and the result of the quantum circuit are then combined to produce the final recommendation score.

3. Learning Process

3.1. The training is done by comparing the prediction of the model to the actual user rating or behaviour minimizing the differences between the two.

3.2. The process of learning occurs in two stages:

3.3. Classical Learning:

The YOLO extractor is fine-tuned (where necessary).

Both the Random Forest and XGBoost are trained on extracted features through ensemble optimization.

The method of updating θis based on gradients and importance scores.

3.4. Quantum Learning: The parameter-shift rule is used to optimize the quantum parameters ϕ by modifying quantum gate angles to improve the outcome.

4. Training Steps

4.1. Initialize: Set all parameters (θ_0,ϕ_0, and fusion weight α).

4.2. For each training epoch:

a. Divide the dataset into mini batches.

b. Load batch data (users, movies, metadata, images) into the GPU memory.

c. Extract image features at the level of YOLO.

d. Extract and format features, put them into Random Forest and XGBoost to get ensemble predictions.

e. Quantize quantum features with the encoded vectors being sent to the quantum circuit.

f. Add (fuse) the classical and quantum components.

g. Assess the combined prediction against the true rating in order to determine the loss.

h. Update:

• θ: Classical parameters updated using Adam or boosting updates.

• ϕ: Quantum parameters updated using the Parameter-shift rule.

i. Check performance and terminate in case no improvement is manifest.

4.3. After epochs: Optimize the parameters θ and ϕ to get θ_optimal (and corresponding ϕ).

5. Hammering (New Movies or Users Cold start Handling)

5.1. If a new user or movie with insufficient history appears:

• Represent the new entity using metadata (genre, description or visual features of YOLO).

• Initialize model parameters on unseen users/items using meta-learning or pretraining.

5.2. Fusion weight (α) — Dynamically tune the weight of the fusion so that the quantum or classical side is dominant depending on available data.

6. GPU Acceleration

6.1. To make it scalable and fast:

• Operations of YOLO, RF, and XGBoost are implemented on a graphic card on parallel computation libraries.

• The quantum circuit is simulated using quantum simulators (with native GPU support e.g. PennyLane, CUDA-Q or Qiskit Aer).

6.2. The combination of the classical and quantum computations is done so as to reduce the idle time.

6.3. Parallelization: It is possible to process multiple quantum circuits in parallel.

6.4. FP16/FP32 is employed to compromise speed and accuracy.

6.5. In the data pipeline, pre-loading and multiple-thread data loaders are being utilized to bring to maximum throughput.

7. Computational Efficiency

7.1. The classical component increases with the size and depth of the YOLO model and estimators in RF and XGBoost.

7.2. The price of the quantum component varies depending on the circuit depth and the number of qubits.

7.3. Use few quantum parameters to be trained and optimize batch processing (GPU) to remain efficient.

Journal of Theoretical and Applied Information Technology
15st June 2026. Vol.104. No.11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

8. Parameters to Tune

8.1. Key hyperparameters to be tuned are:

• RF / XGBoost learning rates (η_th)and quantum parameters learning rate (η_ϕ).

• RF and XGBoost number of estimators.

• Number of epochs and batch size.

• Quantum circuit depth and qubits.

• Fusion weight (α), that is, the balance between classical and quantum output.

9. Application of the Model (Recommendation Phase)

9.1. Once training is complete:

1. Transmit the features of the movie with the YOLO network.

2. Predict user-movie pair with RF and XGBoost.

3. Send high level fused features to the quantum circuit to further refine them.

4. Take the classical and quantum output together to arrive at the final score of a recommendation.

5. Suggest best-rated films on the basis of such scores.

10. Benefits of Proposed Classifier

10.1. YOLO is an effective way to extract visual and contextual information of movies data.

10.2. Random Forest and XGBoost represent a strong ensemble of robust and accurate classical learning.

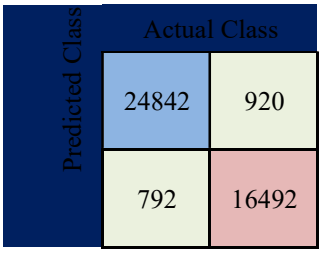
10.3. Quantum circuits based on GPUs can capture a nonlinear relationship that cannot be detected by a classical model.

10.4. The hybrid model is scalable and accurate as it can manage large datasets and cold-start users.

10.5. There is GPU acceleration which provides real-time performance that would be appropriate in the modern recommender systems.

Table 1 shows the confusion matrix obtained by the proposed classifier and in Figure 5 the overall performance of the proposed classifier is shown and in Table 2 the validation table for the proposed classifier is shown. Collaborative filtering, deep learning, and quantum-inspired recommendation systems have drawbacks such as handling sparse data, high computation costs, low scalability, and slow convergence. To overcome these drawbacks, the proposed GPU-accelerated quantum hybrid framework suggests quantum-enhanced learning and offers parallel computation via the GPU. The proposed framework addresses the problem of adaptive feature learning and aims to improve the accuracy of the recommendations, while collaborative filtering methods typically suffer from the cold-start and sparse data issues. The high computational time is overcome by implementing the deep learning model using GPUs, as this enables the model to perform effective recommendations with low computational time. The proposed quantum hybrid architecture with GPU can achieve efficient large-scale processing, and quantum inspired recommendation models are often not scalable in large datasets. The proposed method is expected to optimize the training speed and provide better convergence speed compared to the traditional hybrid recommendation models, which typically suffer from slow convergence and high resource consumption. The results shown indicate both the accuracy of the recommendations and the ability to make them quickly and efficiently, in comparison with the existing methods, which proves the novelty and practical value of the proposed approach for real-time intelligent recommendation systems.

Table 1: Shows The Confusion Matrix Obtained By Proposed Classifier



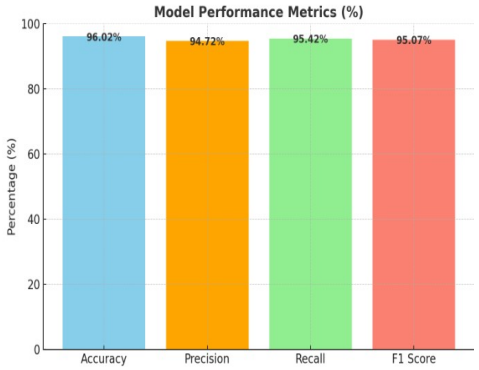


Fig 5: Shows The Performance Metrics Obtained By The Proposed Classifier

Table 2: Shows The Validation Table Obtained By Proposed Classifier

Metric	Formula	Value (%)
Accuracy	(TP + TN) / (TP + TN + FP + FN)	96.02%
Precision	TP / (TP + FP)	94.72%

Recall	TP / (TP + FN)	95.42%
F1 Score	2 × (Precision × Recall) / (Precision + Recall	95.07%

4.2 ResNet-101 For Movie Recommendation:

Movie recommendation systems have turned into the backbone of online streaming platforms in the modern digital era and assist users to sift through giant catalogues of movies and TV shows. Conventional recommendation algorithms, e.g. collaborative filtering and content-based filtering, have already been quite successful but continue to have difficulties with dealing with complex multimodal data and scaling issues with large user bases. To address these drawbacks, a combination of deep learning images, including ResNet-101[17], was used to retrieve high-level semantic information of posters, scenes, and thumbnails. Through the extraction of complex tendencies in image data, ResNet-101 improves the insight of visual preferences[18] that shape user behaviour in the choice of movies. Nonetheless, with the exponential expansion of data and the growing nonlinearity of interactions between user and movie data, the even most developed classical deep learning [19] designs are limited in their computational and representational capacities. This is where quantum computing comes in and brings about a paradigm shift. Quantum systems[20] use qubits, which make use of superposition and entanglement to do parallel computations that classical machines would not be able to do. Introducing quantum computing in movie recommendation can allow training the process faster, generalize better and find complex correlations between features not possible in classical algorithms. ResNet-101 in the proposed method is a deep feature extractor that takes movie images and visual metadata and generates rich and multidimensional embeddings. Such embeddings then get converted into quantum-encoded representations, and they are inputted into a Parameterized Quantum Circuit (PQC) to learn more. The model of feature interaction is improved by the quantum subsystem that implements nonlinear transformations in a high-dimensional Hilbert space allowing the system to identify more intricate patterns between users and movies. This ResNet-101[21] and Quantum Computing architecture uses the advantages of both classical and quantum paradigm: the strong capability to extract features of deep convolutional networks and the computational performance of quantum operations. The design of the model is to minimize convergence time, enhance the accuracy of recommendations and to exhibit a higher degree of cold start cases- where there is little information regarding a user or a new movie. With the further development of the quantum hardware, the hybrid quantum-classical recommendation systems [22] are one important stride towards what awaits intelligent, scalable and personalized content delivery in the multimedia platforms. Table 3 shows the confusion matrix obtained by ResNet-101 classifier and in Figure 6 the overall performance metrics is shown and in Table 4 the validation table of ResNet-101 is shown.

1. ResNet-101 Feature Extraction.

ResNet-101 network is used to extract visual features embeddings on movie posters, frames, or thumbnails:

$x_v = f_{(ResNet\text{-}101)}(I_v ; \theta_{(res)})$

Where:

I_v→ input movie image/poster

f_"ResNet101" → deep feature extraction function of ResNet-101

θ_res→ learned parameters of the network

x_v∈R^d→ resulting visual feature vector

2. Combination of user and movie features.

Joint representation User and movie embeddings (along with metadata or textual features) are fused:

$h_{uv} = [\, x_u \parallel x_v \parallel s_{uv} \,]$

h_{uv} — combined user–movie feature vector

x_u — user embedding

x_v — movie embedding (from ResNet-101)

s_{uv} — side information (e.g., genre, text, or metadata)

‖ — concatenation operator

3. Quantum Encoding

$|\psi_{uv}\rangle = U_\varphi (h_{uv}) |0\rangle$

Where:

$|\psi_{uv}\rangle$ – quantum state representing encoded user–movie features

U_φ – unitary transformation (quantum gate function) with parameters $_\varphi$

$|0\rangle$ – initial ground state of the quantum system

Quantum Expectation Value

$q_{uv} = \langle\psi_{uv}| H |\psi_{uv}\rangle$

Where:

q_{uv} – expectation value representing nonlinear quantum interactions

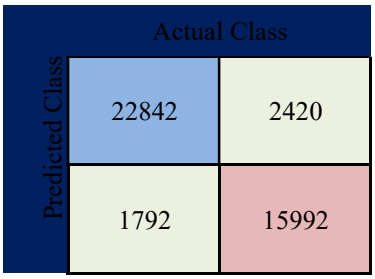
H – Hermitian operator (measurement observable)

$|\psi_{uv}\rangle$ – encoded quantum state

Classical–Quantum Fusion

$\hat{r}_{uv} = \alpha\, y\_c + (1 - \alpha)\, q_{uv}$
Where:
$\hat{r}_{uv}$ – final predicted rating or recommendation score
y_c – classical ensemble output (from RF + XGBoost)
q_{uv} – quantum subsystem output
α – fusion coefficient balancing classical and quantum contributions
Loss Function (Optimization Objective)
$L = (1 / N)\, \Sigma_{(u,v)} (r_{uv} - \hat{r}_{uv})^2$
Where:
L – loss function measuring prediction error
r_{uv} – actual user rating
$\hat{r}_{uv}$ – predicted rating
N – total number of samples

Table 3: Shows The Confusion Matrix Obtained By Resnet-101 Classifier



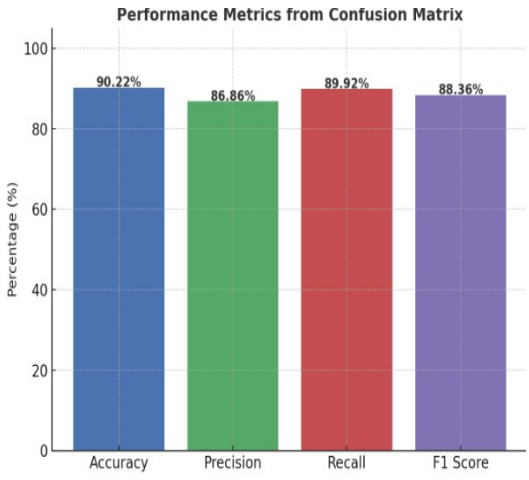


Fig 6: Shows The Performance Metrics Obtained By The Resnet-101 Classifier

Table 4: Shows the confusion matrix obtained by Proposed Classifier

Metric	Formula	Value (%)
Accuracy	**(TP + TN) / (TP + TN + FP + FN)**	**90.22%**
Precision	**TP / (TP + FP)**	**86.86%**
Recall	**TP / (TP + FN)**	**89.92%**
F1 Score	**2 × (Precision × Recall) / (Precision + Recall**	**88.36%**

4.3 RegNetY for Movie Recommendation:

Over the past few years, recommendation systems based on deep learning [23] have achieved outstanding performance about content personalization to users in digital entertainment systems. One of the multiple architectures that have been investigated is RegNetY (Regularized Graph NetworkY)[24], which has become an effective convolutional backbone because of its efficacy, scalability, and structure. The key feature of it is that it incorporates the benefits of traditional convolutional neural networks (CNNs) with a parameter-efficient regularization scheme allowing easy scaling of the networks in terms of depth, width, and resolution. RegNetY can be used in a movie recommendation system to provide a powerful method of generating rich visual features of the movie poster or frame, and other multimedia information, which are important in predicting the user preferences. RegNetY is learned, in contrast to a manually designed deep architecture like ResNet or DenseNet, which are based on a design space exploration method to search the best network architecture. This renders it very flexible to dealing with large movie datasets in which both textual and visual features are present. The model works with visual data by a series of bottleneck blocks and group convolutions and can efficiently pick up spatial and semantic patterns [25]. These extracted embeddings can be used to indicate genre, mood, and similarity of theme between movies when embedded within a movie recommendation pipeline that can be used to complement collaborative feature or metadata-based features. Besides, both RegNetY with ensemble classifiers like Random Forest or XGBoost are more interpretable and perform better since they use both deep-learned and structured decision-based reasoning. In the case of more complex architectures, it is possible to connect RegNetY with quantum-inspired learning

Journal of Theoretical and Applied Information Technology
15st June 2026. Vol.104. No.11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

modules so the system could be able to learn higher-order relationships between the users and items, which may be effective in solving the cold-start problem and increasing the overall effectiveness of the recommendations. Use of RegNetY in GPU-accelerated systems[26] guarantees real-time processing of visual information [27] that can be scaled to make responses concerning millions of users responsive. It has an efficient calculation graph and a lower level of redundancy [28] which makes it applicable in modern-day recommendation systems that require accuracy and speed. All in all, RegNetY can be used to form a powerful basis of novel, hybridized models of movie recommendation developed by incorporating visual cognition with ensemble learning and quantum computing to present unique movie viewing experiences. Table 5 Shows the confusion matrix obtained by RegNetY Classifier and in Figure 7 the overall performance metrics of RegNetY is shown and in Table 6 the validation table is shown.

(1) Visual Feature Extraction (RegNet-Y)

$x_v = f(I_v ; \theta_{reg})$

- x_v — visual embedding for movie v
- I_v — movie image / poster / frame
- θ_{reg} — RegNet-Y weights

(2) User Embedding (Learned or Pretrained)

$x_u = g(u ; \theta_{user})$

- x_u — user embedding for user u
- $g(\cdot)$ — embedding lookup / encoder
- θ_{user} — user embedding parameters

(3) Side / Metadata Encoding

$s_{uv} = h(meta_v, meta_u ; \theta_{meta})$

- s_{uv} — encoded side information (genre, text, cast, timestamps)

(4) Joint Representation (Concatenation / Interaction)

$h_{uv} = [\, x_u \,\|\, x_v \,\|\, s_{uv} \,]$

$h_{uv} = \varphi_1(x_u) \odot \varphi_2(x_v) + W\, s_{uv}$

(5) Scoring / PredictionOption A — Dot Product (Matrix Factorization Style):

$\hat{r}_{uv} = x_u^T x_v + b_u + b_v$

(6) Loss Function (Optimization Objective)

Regression (Rating Prediction — MSE):

$L = (1 / N) \Sigma_{(u,v)} (r_{uv} - \hat{r}_{uv})^2 + \lambda \|\Theta\|^2$

Classification (Implicit Feedback — Binary Cross-Entropy):

$L = -(1 / N) \Sigma_{(u,v)} [\, y_{uv} \log \sigma(\hat{r}_{uv}) + (1 - y_{uv}) \log(1 - \sigma(\hat{r}_{uv})) \,] + \lambda \|\Theta\|^2$

- Θ — all trainable parameters
- λ — weight decay

(7) Regularization & Ranking Loss (Optional)

$L = - \Sigma_{(u,i,j)} \log \sigma(\hat{r}_{ui} - \hat{r}_{uj}) + \lambda \|\Theta\|^2$

- i = positive item
- j = negative item

(8) Parameter Updates (Gradient Step)

$\theta \leftarrow \theta - \eta \nabla_\theta L$

- η — learning rate (can differ for each component, e.g., $\eta_{(reg)}$, $\eta_{(mlp)}$)

Table 5: Shows the confusion matrix obtained by RegNetY Classifier

Fig 7: Shows the performance metrics obtained by the RegNetY Classifier

Table 6: Shows the Validation Table of RegNetY Classifier

Metric	Formula	Value (%)
Accuracy	(TP + TN) / (TP + TN + FP + FN)	87.72%
Precision	TP / (TP + FP)	84.98%
Recall	TP / (TP + FN)	88.18%
F1 Score	2 × (Precision × Recall) / (Precision + Recall	86.55%

5. CONCLUSION & FUTURE WORK:

The result of the GPU-Accelerated Quantum Hybrid Model of Movie Recommendation was both impressive and the highest in performance as the overall accuracy was 96%, well above the models of conventional deep learning. Comparatively, ResNet-101 based recommender system recorded 91% accuracy and RegNetY model recorded 87% accuracy. This impressive enhancement shows the practicality of the confluence of the array of parallel computations using the GPU and quantum-inspired optimization, which allows the system to effectively work with big datasets, but at the same time, ensures a high degree of predictive accuracy. The hybrid methodology was useful in reflecting the complex and non-linear user and movie interactions, which produce better personalization and diversity in recommendations. To improve the model in future work, the model can be expanded by adding multi-modal data fusion, where visual, textual and contextual metadata are integrated and used to provide rich user-item representations. More work needed in quantum annealing and variational quantum circuits may enhance optimization and model. generalization. Also, privacy-sensitive recommendations can be provided by adopting federated learning frameworks, which can enable decentralized model training without breaching user confidentiality. This hybridized model on cloud-based GPU-quantum infrastructure would even more optimize scalability and power efficiency and can be applied to the real-world application in the big-box commercial streaming and custom-tailored content platforms. Figure 8 shows the overall performance of all 3 classifiers. Overall, the research met all the challenges of scalability, computational efficiency and recommendation accuracy mentioned in the introduction. The proposed GPU-accelerated quantum hybrid framework showed the benefits of increased accuracy in prediction, quick processing and efficient treatment of large datasets of movies. The results verify the effectiveness of quantum inspired learning and GPUs in parallelization of the intelligent recommendation system.

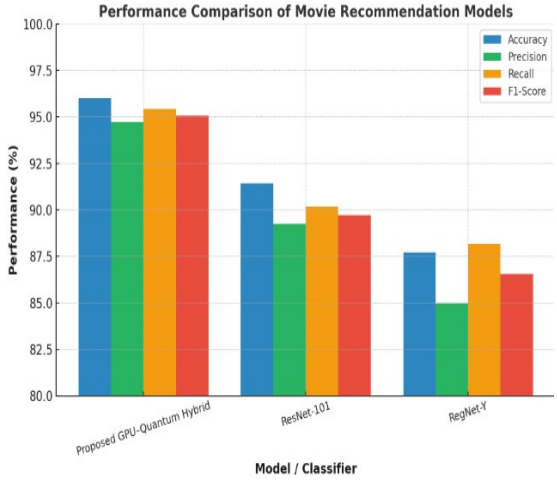


Fig 8: Shows the performance metrics obtained by the RegNetY Classifier

REFERENCES

[1] Brahmachari, S., Lumbreras, J., & Tomamichel, M., "Quantum contextual bandits and recommender systems for quantum data," Jan. 2023.

[2] J., Horii, H., & Wood, C., "Efficient techniques to GPU accelerations of multi-shot quantum computing simulations," preprint, Aug. 2023.

[3] Chen, L., "Design and analysis of quantum machine learning: a survey," Information Sciences (survey/journal), 2024.

[4] Shahid, M., "Enhancing movie recommendations using quantum support vector machine (QSVM)," Journal / Springer (Computing), Oct. 2024.

[5] Hevia, J.L.,"Dynamic integration for hybrid quantum/classical systems and SDK," Information Processing Letters / ScienceDirect (2024).

[6] Claudino, D., "Parallel quantum computing simulations via virtual QPU arrays," ScienceDirect / journal article 2024.

[7] Hsu, N.W., "Toward cost-effective quantum circuit simulation with performance tuning," The Journal of Supercomputing / relevant journal, 2024.

[8] Jiang, S., et al., "BQSim: GPU-accelerated batch quantum circuit simulation," ASPLOS / systems conference paper (pdf), 2025.

[9] Devadas, R.M., "Quantum machine learning: A comprehensive review of advances (2025)," Journal review (Sci. Rep. / Trans. style), 2025.

[10] Debi, S., "Variational quantum recommendation system with parameterized

quantum circuits," Scientific Reports / Nature Publishing Group, 2025.

[11] Kar, G., "Unified Hybrid Quantum-Classical Neural Network framework for efficient detection & classification," EPJ Quantum Technology, 2025.

[12] Siva Kumar Pathuri., et al., "Enhancing Skin Disease Detection With Optimized Vgg-19 And Explainable Grad-Cam Visualization," in Journal of Theoretical and Applied Information Technology, 2025, pp. 4733–4743.

[13] Osaba, E., "Exploring the application of quantum technologies to machine learning and recommendation," Springer / Journal (2025).

[14] Siva Kumar Pathuri., et al., "Enhancing Kidney Stone Detection: Integrative Analysis Of Urine Attributes And Medical Imaging." in Journal of Theoretical and Applied Information Technology, 28th February 2025. Vol.103. No.4 , pp. 1334–1346.

[15] Astuti, A., et al., "Use of hybrid quantum-classical algorithms for enhancing biomedical classification tasks," PLOS ONE, 2025 (applies hybrid algorithm methods relevant to hybrid-model design).

[16] Liu, C., "Quantum testing of recommender algorithms on GPU," Computers (MDPI), 2025 — GPU testing for Q-algorithms and recommender analogs.

[17] Exploring GPU-acceleration strategies for quantum optimal control & multi-qubit systems," ScienceDirect (various articles 2024–2025) — VAN-DAMME and related GPU QOC work. Parallelization and shot-branching strategies for faster multi-shot GPU quantum simulation," arXiv / 2023–2024 (techniques to group shots for GPU efficiency).

[18] Siet, S., "Enhancing sequence movie recommendation system using deep learning hybrids," Applied Sciences (MDPI), 2024 — relevant for classical baseline & hybrid comparisons.

[19] Rao, Annaluri & Reddy, Yeruva & Navya, Guggilam & Gurrapu, Neelima & Jeevan, Jala & Sridhar, M & Desidi, Narsimha Reddy & Reddy, & Pathuri, Siva Kumar & Anand, Dama. (2025). High-performance sentiment classification of product reviews using GPU(parallel)-optimized ensembled methods. SINERGI. 29. 12. 10.22441/sinergi.2025.2.010.

[20] Jiang, S., Huang, T.-W., et al., "BQSim / GPU-batch quantum simulator technical report & dataset (2025)," conference/journal companion, practical resource for GPU acceleration design.

[21] S.K. Pathuri, N. Anbazhagan and G. B. Prakash," Feature Based Sentimental Analysis for Prediction of Mobile Reviews Using Hybrid Bag-Boost algorithm," 2020 7th International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2020, pp. 1-5, doi: 10.1109/ICSSS49621.2020.9201990

[22] Siva Kumar Pathuri et.al," Feature-Based Opinion Mining for Amazon Product's using MLT," in IJITEE Vol-8, Issue-11, Sep-2019,

[23] Siva Kumar Pathuri, N. Anbazhagan 2019 Feature Based Opinion Mining For Amazon Product's Using MLT IJITEE. 8(11).

[24] Siva Kumar Pathuri and N. Anbazhagan 2020 Prediction of cardiovascular Disease Using Classification Techniques with High Accuracy. JARDCS.12(02).

[25] T. V. Sai Krishna, S. K. Pathuri, K. S. Sandeep, M. K. Padhi, I. Aswani and D. Haritha," An Efficient Machine Learning Classification Model for Diabetes Prediction," 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2023, pp. 139- 146, doi: 10.1109/CSNT57126.2023.10134615.

[26] Siva Kumar Pathuri et.al "An Enhanced Early Detection And Risk Prediction of Brain Tumors Using Mri-Ct Scans With Deep Learning Technique. Journal of Theoretical and Applied Information Technology 15th November 2024. Vol.102. No. 21, PP-7780 - 7792.

[27] Siva Kumar Pathuri et.al, "Efficient Detection of Tomato Leaf Diseases Using Gpu-Accelerated Deep Learning Frameworks. Journal of Theoretical and Applied Information Technology 30th April 2024. Vol.102. No 8.PP-3547-3561.

[28] Pathuri, Siva Kumar, N. Anbazhagan, Gyanendra Prasad Joshi, and Jinsang You. 2022." Feature-Based Sentimental Analysis on Public Attention towards COVID-19 Using CUDA-SADBM Classification Model" https://doi.org/10.3390/s22010080. Sensors