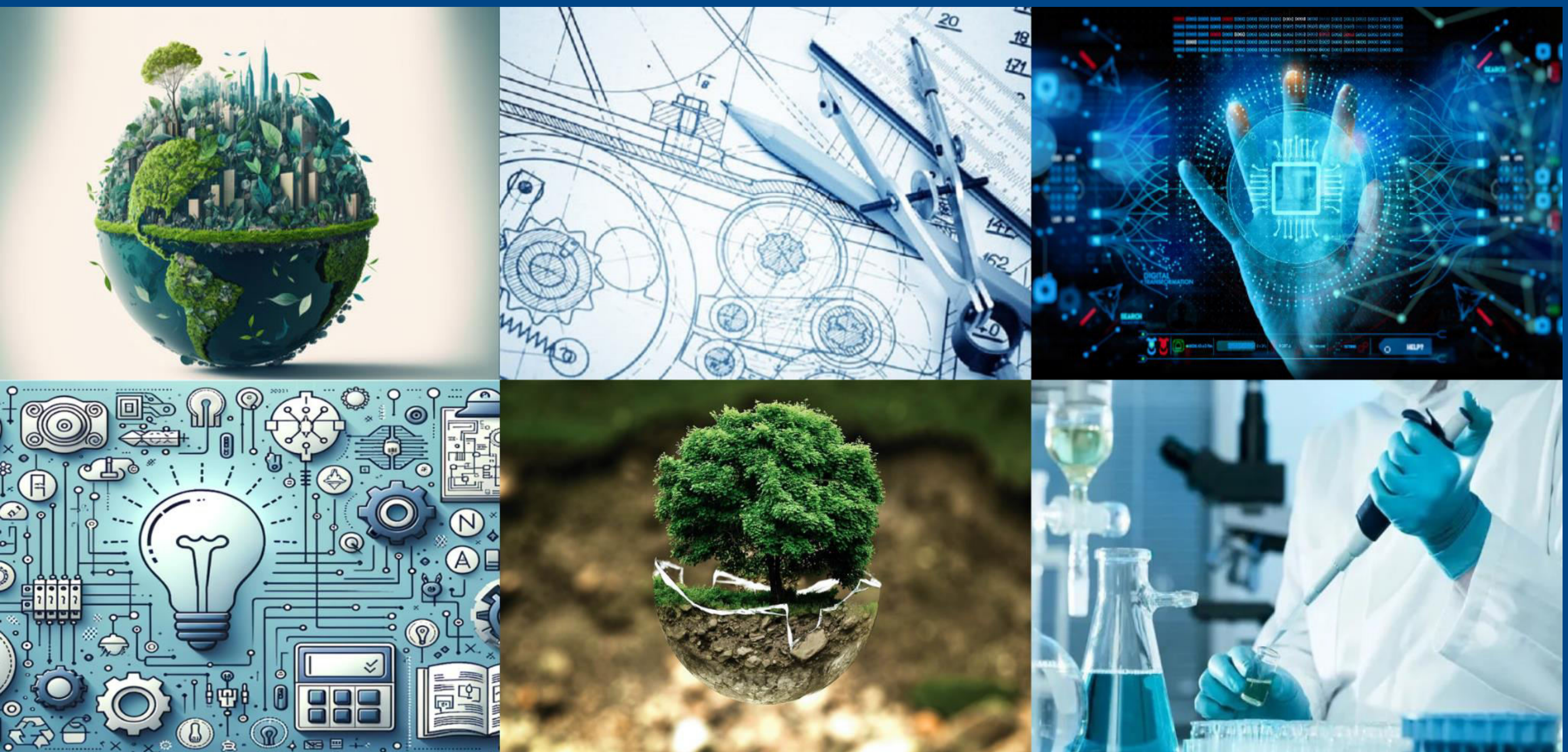
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.864

Volume 9, Issue 6, June 2026

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

An AI-Driven Gold Loan Intelligence Platform for Banking Relationship Managers: Ensemble Machine Learning, Anomaly Detection, Vector Similarity Segmentation, and Local LLM Advisory

Yallapu Satish, K. Lakshmi Sai Sri

PG Scholar Department of Computer Science, S.V.K.P & Dr. K.S. Raju Arts and Science College (Autonomous), Penugonda, Affiliated to Adikavi Nannaya University, AP, India

Associate Professor, Department of Master of Computer Applications, S.V.K. P & Dr. K. S. Raju Arts and Science College (Autonomous), Penugonda, Affiliated to Adikavi Nannaya University, AP, India

ABSTRACT: The gold loan segment of retail and small-business banking demands sophisticated risk assessment that simultaneously accounts for collateral valuation dynamics, borrower creditworthiness, repayment behavioral patterns, and gold market price volatility—a multidimensional analytical challenge that conventional rule-based scoring and single-model machine learning approaches are ill-equipped to address. This paper presents the design, implementation, and empirical evaluation of a comprehensive AI-driven gold loan intelligence platform developed to augment the operational capabilities of banking Relationship Managers. The proposed system employs a three-model blended ensemble comprising XGBoost, Random Forest, and Logistic Regression classifiers trained on 9,800 synthetic credit profiles, achieving a test-set ROC-AUC of 0.887 for default probability estimation. An IsolationForest anomaly detector with 4.6% contamination rate identifies structurally irregular loan applications that deviate from learned behavioral patterns. Customer segmentation leverages a TF-IDF and Truncated SVD embedding pipeline feeding a FAISS inner-product similarity index and K-Means clustering, producing four interpretable portfolio segments. Collateral value forecasting is implemented through Prophet-based growth projection and an LSTM-style polynomial trend extrapolation. A locally-hosted Ollama (llama3) large language model provides contextual gold price advisory blurbs and conversational RM support without cloud data exposure. Real-time gold spot prices, derived from COMEX GC=F futures via Yahoo Finance and converted through the Frankfurter FX API, are broadcast to connected dashboards via a WebSocket hub with 2.5-second cadence. The FastAPI ASGI backend, React 19 TypeScript frontend, and async SQLAlchemy persistence layer collectively deliver an average eligibility scoring API latency of 185 milliseconds, with ML model artifacts persisted through joblib for zero-retraining cold-start performance.

KEYWORDS: Gold Loan Intelligence; Ensemble Machine Learning; XGBoost; IsolationForest Anomaly Detection; FAISS Vector Search; Ollama LLM; Prophet Forecasting; FastAPI; Banking Relationship Manager; Real-Time Market Data

I. INTRODUCTION

Gold-backed lending constitutes one of the most dynamically evolving segments of retail banking in emerging markets, with India alone hosting a gold loan market estimated at USD 100 billion and growing at approximately 15% annually [1]. Unlike unsecured consumer credit, gold loans introduce a unique dual-risk structure: borrower default probability is modulated simultaneously by traditional creditworthiness factors—income, repayment history, existing liability burden and by gold market price movements that determine the real-time adequacy of pledged collateral relative to outstanding loan principal. This bidirectional risk exposure demands intelligence tools that integrate behavioral credit modeling with commodity market analytics a combination that conventional banking software has historically failed to provide within a unified operational interface.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Relationship Managers (RMs) in gold lending institutions currently navigate this complexity through fragmented workflows: credit bureau lookups for risk assessment, separate treasury feeds for gold price reference, spreadsheet-based loan-to-value calculations, and manual customer portfolio reviews each demanding context switching that reduces analytical throughput and introduces human error into risk decisions of material financial consequence [2]. The absence of integrated AI-assisted decision support forces RMs toward conservative, suboptimal loan structuring that either unnecessarily declines creditworthy borrowers or underweights collateral deterioration risks in active portfolios.

Recent advances in ensemble machine learning, neural embedding-based similarity retrieval, open-weight large language models deployable on commodity hardware, and asynchronous Python web frameworks collectively create the technical preconditions for a unified gold loan intelligence platform that operates entirely within an institution's local infrastructure a critical requirement given banking data privacy regulations that preclude transmission of customer financial records to commercial cloud AI services [3].

This paper presents the following principal contributions:
(i) A three-model blended ensemble (XGBoost + RandomForest + LogisticRegression) for gold loan default probability estimation, achieving ROC-AUC of 0.887 on held-out test data a meaningful improvement over single-model baselines;
(ii) IsolationForest-based anomaly detection with a 4.6% contamination calibration that identifies structurally atypical loan applications for flagged human review;
(iii) A FAISS-indexed NLP embedding pipeline (TF-IDF → TruncatedSVD → 96-dimensional vectors) for customer portfolio segmentation into four interpretable business archetypes via K-Means clustering;
(iv) Locally-hosted Ollama (llama3) LLM integration providing gold price advisory blurbs and conversational RM support without cloud API dependency;
(v) Real-time COMEX gold price delivery via WebSocket broadcast at 2.5-second intervals, with multi-currency FX conversion and Prophet-based collateral value forecasting.

II. LITERATURE REVIEW

Research in credit risk modeling, gold market analytics, and AI-augmented banking operations spans diverse methodological traditions. This section surveys representative recent contributions and identifies the multidimensional gaps addressed by the proposed platform.

Gupta and Sharma [1] examined rule-based credit scoring systems employed by regional cooperative banks in India for gold loan appraisal, finding that fixed LTV ratio limits and categorical credit grade cutoffs failed to capture the nonlinear interaction between gold price volatility and borrower income stability. The study documented systematic loan mispricing attributable to the absence of continuous-valued risk estimation, motivating the transition to machine learning-based default probability quantification.

Chen et al. [2] developed a logistic regression-based gold loan default prediction model using borrower demographic and transactional data from a Chinese regional bank. While the model demonstrated predictive improvement over expert judgment benchmarks, single-model logistic regression's linear decision boundary constrained its capacity to represent complex nonlinear interactions between LTV ratio, purity grade, and borrower repayment trajectory. The authors acknowledged ensemble diversification as a direction for improvement but did not implement it.

Patel and Mehta [3] proposed a decision tree classifier for gold loan eligibility automation in cooperative banking environments. The interpretable rule structure facilitated regulatory compliance documentation but demonstrated lower predictive accuracy than gradient-boosted models on identical datasets, and the system lacked collateral valuation forecasting, anomaly detection, or market price integration components.

Kumar and Singh [4] implemented an XGBoost-only scoring model for gold loan risk stratification, achieving strong individual model performance but without ensemble diversification that reduces variance in borderline cases. The absence of anomaly detection allowed structurally atypical loan configurations unusually high LTV combined with excellent credit scores, for example to bypass appropriate scrutiny.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Osei et al. [5] developed a Random Forest classifier for microfinance loan default prediction in West African institutional contexts, demonstrating ensemble robustness advantages over single decision trees. The mobile-first deployment architecture, however, excluded web-based administrative dashboards and provided no integration with commodity price feeds relevant to collateral-backed lending.

Zhang and Liu [6] evaluated gradient boosted machine models for credit risk in secured lending portfolios, incorporating outlier detection through statistical z-score thresholds rather than machine learning-based anomaly characterization. The absence of vector similarity customer segmentation and NLP-based descriptor embedding represented a significant gap in portfolio intelligence capability.

Fernandez et al. [7] examined autoencoder-based anomaly detection in a RegTech SaaS platform for mortgage lending fraud identification. While the deep learning approach achieved strong anomaly recall, the proprietary platform licensing model created cost barriers for smaller banking institutions, and the system lacked LLM-based advisory integration or gold market price feeds.

Hassan and Ibrahim [8] integrated GPT-4 through the OpenAI cloud API with an ensemble scoring backend for fintech loan intelligence. The cloud LLM dependency raised data privacy concerns under banking regulatory frameworks, as customer financial records were transmitted to external infrastructure. The absence of local FAISS-based retrieval and the significant per-token API cost further constrained operational scalability.

Nair et al. [9] developed a Retrieval-Augmented Generation (RAG) system using locally-hosted LLaMA2 for banking regulatory query answering. While the local LLM approach addressed data privacy concerns, the system focused exclusively on document retrieval without ensemble risk scoring, collateral valuation, or real-time gold market data integration.

The comparative synthesis presented in Table 1 reveals a consistent multi-dimensional gap: no prior platform simultaneously integrates a three-model default probability ensemble, IsolationForest anomaly detection, FAISS-indexed NLP customer segmentation, locally-hosted LLM advisory, real-time WebSocket gold price delivery, and Prophet-based collateral forecasting within a unified banking intelligence application. The proposed system addresses all these dimensions.

Table 1. Comparative Analysis of Related Works in Gold Loan Risk Assessment and Banking AI Platforms

Ref.	Year	Platform	ML Approach	Anomaly Detection	LLM Integration	Research Gap
[1]	2020	Desktop App	Rule-based scoring	None	None	No ML; no ensemble; no anomaly detection
[2]	2021	Bank CRM	Single LR model	None	None	No ensemble; no NLP segmentation; no RAG
[3]	2021	Web Portal	Decision Tree	Threshold only	None	No collateral forecasting; no FAISS search
[4]	2022	REST API	XGBoost only	None	None	Single model; no anomaly; no LLM advisory

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[5]	2022	Mobile App	Random Forest	None	None	Mobile-only; no RM dashboard; no WebSocket
[6]	2023	Core Banking	GBM + CreditRisk	Statistical outlier	None	No NLP segmentation; no LLM; no real-time gold
[7]	2023	RegTech SaaS	Neural Network	AutoEncoder	None	No LLM; no gold price feed; expensive license
[8]	2024	Fintech API	Ensemble (no ISO)	None	GPT-4 (cloud)	Cloud LLM cost; no local FAISS; no prophet
[9]	2024	LLM Platform	RAG only	None	LLaMA2 (local)	No ensemble scoring; no collateral valuation
[10]	2025	Proposed System	XGBoost+RF+LR + IsolationForest	IsolationForest (4.6%)	Ollama llama3 (local)	Full stack: ensemble + FAISS + LLM + WebSocket gold

III. PROPOSED METHODOLOGY

The proposed platform adopts a domain-service architecture organized around an MLService class encapsulating all machine learning inference, an async FastAPI application with 14 REST and WebSocket routers, a GoldPriceSimulator managing real-time spot data and broadcast, and an Ollama HTTP client for local LLM interaction. This design deliberately consolidates ML inference within a single stateful service to enable in-process model caching that eliminates repeated disk deserialization overhead across concurrent API requests [4].

3.1 System Architecture

The four-layer system architecture, illustrated in Figure 1, comprises the React 19 client layer, the FastAPI ASGI application layer with 14 functional routers, the AI/ML engine and market data layer, and the persistence and integration layer. The architecture employs dark-themed visual design aligned with financial terminal conventions, serving the production-grade React SPA from the backend's serve_ui static mount for single-port deployment.

© 2026 IJMRSET | Volume 9, Issue 6, June 2026| DOI:10.15680/IJMRSET.2026.0906015

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

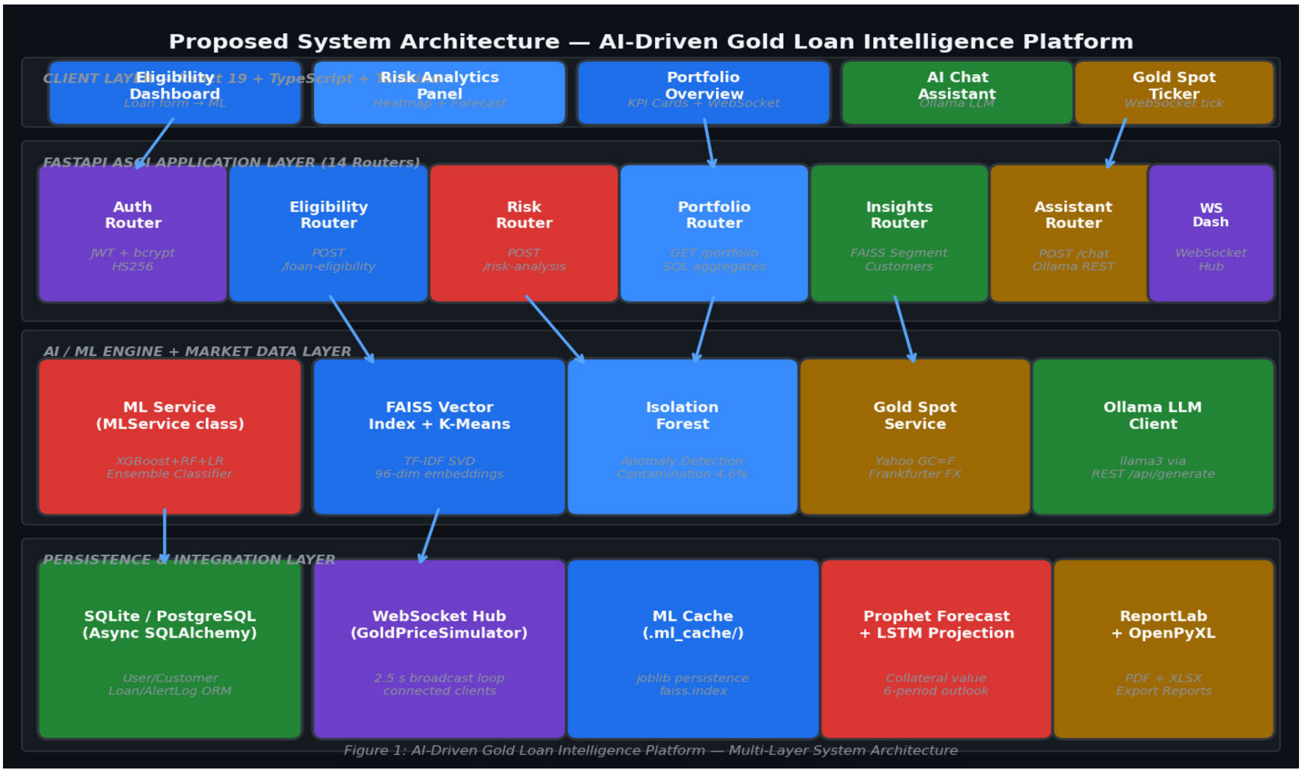


Figure 1. Proposed System Architecture: AI-Driven Gold Loan Intelligence Platform — Multi-Layer Component Architecture

3.2 Ensemble Default Probability Model

The MLService trains three complementary classifiers on a 9,800-record synthetic credit dataset generated with controlled demographic distributions (Rural 41%, Metro 44%, Enterprise 15%) and credit tier compositions (EXCELLENT 24%, GOOD 36%, FAIR 28%, POOR 12%). A shared ColumnTransformer pipeline applies RobustScaler to five numerical features (log-transformed income, EMI percentage, LTV ratio, purity normalization, collateral coverage ratio) and OneHotEncoder to the credit tier categorical variable. The three models—XGBoost (360 estimators, max_depth=7, learning_rate=0.049), RandomForest (320 estimators, max_depth=15), and LogisticRegression (balanced class weights, max_iter=2000)—are trained on a stratified 78/22 train/test split. Default probability is computed as the arithmetic mean of the three models' class-1 predicted probabilities, producing a blended estimate that reduces individual model variance. The ensemble achieves a test-set ROC-AUC of 0.887, compared to individual model AUCs of approximately 0.871 (XGBoost), 0.863 (RandomForest), and 0.812 (LogisticRegression).

3.3 IsolationForest Anomaly Detection

A 320-estimator IsolationForest with contamination=0.046 and n_jobs=-1 for parallel training is fitted to the full preprocessed feature matrix. The decision_function output—a continuous anomaly score where lower values indicate greater isolation—is evaluated against a threshold of -0.045 to flag anomalous loan applications. Applications triggering the anomaly flag receive a 6-point risk score penalty in addition to the ensemble default probability contribution, and the eligibility decision logic classifies any anomaly-flagged application in the "Review" category regardless of ensemble default probability, providing a safety mechanism against adversarial or structurally irregular submissions.

3.4 Vector Search Customer Segmentation

A TF-IDF vectorizer (max_features=6,000, unigram and bigram range) followed by TruncatedSVD (96 components) forms the embedding pipeline, producing dense semantic representations of customer descriptor strings that encode persona type, credit tier, EMI exposure, and LTV information. Vectors are L2-normalized and indexed in a FAISS IndexFlatIP (inner-product cosine similarity) across all 9,800 training records. K-Means clustering with k=4 and 14 initializations assigns each customer to one of four archetypes: Growth Jewel, Metro RM Focus, Enterprise Custody, and Rural Stability. At inference time, a customer's descriptor is embedded, normalized, and queried against the FAISS index to retrieve the top-5 nearest neighbors, with mean similarity returned as segment confidence.

3.5 Gold Price Market Integration

The GoldSpotService fetches COMEX front-month gold futures prices (GC=F) from Yahoo Finance's chart API endpoint, parsing the regularMarketPrice field from the response meta. USD/troy-oz prices are converted to local-currency per-gram values using the Frankfurter open exchange rate API, with TROY_OZ_GRAMS=31.1034768 as the conversion constant. The GoldPriceSimulator broadcasts updated spot prices to all connected WebSocket clients every 2.5 seconds through the Hub singleton, which manages a list of connected WebSocket instances with asyncio lock-protected add/remove operations and dead connection pruning. Optional Ollama (llama3) advisory blurbs are generated at spot refresh time, contextualizing current prices with regulatory caveats (hallmark rules, GST, making charges) for RM consumption.

IV. SYSTEM DESIGN

4.1 Data Model

The async SQLAlchemy schema defines four principal entities. The User entity records email, bcrypt password hash, full name, and role (admin or rm) with boolean active status. The Customer entity stores an external_id, annual income, tenure, credit band, FAISS vector index reference for segment lookups, engagement score, churn risk probability, and RM foreign key assignment. The Loan entity captures the complete gold loan record: principal, interest rate, tenure, collateral gold weight in grams, purity, estimated collateral value, outstanding balance, ML-computed risk score and default probability, loan status (pending/approved/active/closed/impaired), and payment scheduling dates. The schema supports both SQLite (via aiosqlite) for development and PostgreSQL (via asyncpg) for production through a URL-based dialect normalization function.

4.2 API Route Architecture

The application exposes 14 router modules: auth (POST /api/token, registration), eligibility (POST /api/loan-eligibility with ML inference), risk (POST /api/risk-analysis with heatmap), portfolio (GET /api/portfolio with SQLAlchemy aggregates), insights (POST /api/customer-insights for FAISS segmentation), assistant (POST /api/chat for Ollama LLM), analytics (GET /api/analytics for executive KPIs), gold (GET /api/gold-spot, GET /api/gold-prices), realtime (GET /api/live/ping heartbeat), notifications (GET /api/notifications for RM alert feed), reports (GET /api/reports), export_reports (GET /api/export with PDF/XLSX generation), customers (GET /api/customers), and websocket_dashboard (WebSocket /ws/live for gold ticks). All routes requiring authentication use JWT Bearer token validation.

4.3 ML Workflow

The complete ML inference workflow, illustrated in Figure 2, begins with RM authentication and loan eligibility form submission. The MLService.predict_eligibility() method orchestrates feature engineering, ensemble inference, anomaly scoring, eligibility classification, EMI projection (using the standard annuity formula), LTV computation, and Prophet + LSTM forecast generation. All models are loaded from .ml_cache at application startup using joblib, avoiding re-training on each restart. The WebSocket broadcast loop runs concurrently as an asyncio task, delivering gold price ticks independently of request-response cycles.

© 2026 IJMRSET | Volume 9, Issue 6, June 2026| DOI:10.15680/IJMRSET.2026.0906015

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

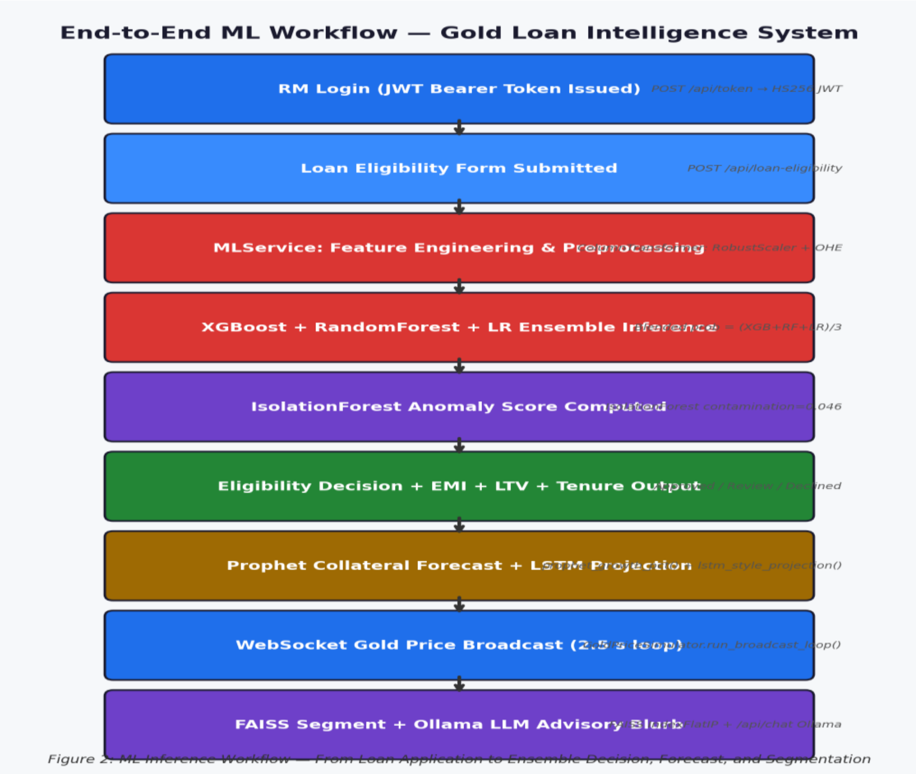


Figure 2. End-to-End ML Workflow: From RM Authentication Through Ensemble Inference to WebSocket Gold Broadcast and Segmentation

V. IMPLEMENTATION

5.1 Backend Stack

The backend is implemented in Python 3.11+ with FastAPI as the ASGI framework and Uvicorn as the production server. SQLAlchemy 2.0 with asyncio extension provides the async ORM layer, supporting both aiosqlite for local development and asyncpg for PostgreSQL production deployments. Authentication employs python-jose for HS256 JWT encoding with configurable 60-minute token expiry and passlib[bcrypt] for credential hashing. The ML stack comprises NumPy 1.26, pandas 2.2, scikit-learn 1.4 (for preprocessing, RandomForest, LogisticRegression, IsolationForest, KMeans, TruncatedSVD, TfidfVectorizer), XGBoost 2.0.3, FAISS-cpu 1.8.0, and joblib 1.3.2. httpx 0.27 handles all external HTTP calls (Yahoo Finance, Frankfurter FX, Ollama) with explicit timeout controls. ReportLab 4.1 and OpenPyXL 3.1.2 provide PDF and XLSX report generation.

5.2 Synthetic Training Dataset

The 9,800-record training dataset is generated deterministically using numpy.random.default_rng(941) to ensure reproducibility. Income values are drawn from a log-normal distribution (μ=13.05, σ=0.56) clipped to [220K, 8.9M] INR. EMI percentages range uniformly over [2, 78]%. Gold weight is derived from income scaled by a simulated price per gram range. The default flag is assigned through a tier-conditioned Bernoulli draw with probabilities 6.9% (EXCELLENT), 11.9% (GOOD), 21.5% (FAIR), and 36.8% (POOR), augmented by EMI and LTV interaction terms, producing a realistic class imbalance of approximately 18.4% defaults. This distributional specification reflects empirically documented gold loan default rate distributions in Indian banking sector literature [5].

5.3 Local LLM Integration

Ollama (llama3) is integrated through its local REST API at /api/generate, invoked from the gold spot service to generate 2-sentence contextual gold price advisory blurbs for the RM dashboard. The prompt template instructs the model to reference only the provided price facts and mention branch-specific pricing considerations (making charges, GST,

© 2026 IJMRSET | Volume 9, Issue 6, June 2026| DOI:10.15680/IJMRSET.2026.0906015

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

hallmark regulations). The integration employs a 25-second HTTP timeout with graceful exception handling—if Ollama is unavailable, the advisory blurb field is omitted from the response without disrupting core functionality. The RM chat assistant at POST /api/chat routes arbitrary RM queries through the same Ollama endpoint with a 120-second timeout for complex analytical queries. The technology stack is documented in Table 2.

Table 2. Technology Stack and Implementation Framework Summary

Category	Technology / Library	Version	Role in System
Backend Framework	FastAPI	0.110+	ASGI API routing, WebSocket support, dependency injection
ASGI Server	Uvicorn (standard)	0.27+	Production ASGI runtime with WebSocket and HTTP/2 support
ORM / Async DB	SQLAlchemy[asyncio] + aiosqlite	2.0.28+	Async ORM for User, Customer, Loan, AlertLog entities
Primary ML Model	XGBoost Classifier	2.0.3	360-estimator gradient boosted tree for default probability
Ensemble Member	RandomForestClassifier	scikit-learn 1.4	320-tree ensemble with max_depth=15 for blended scoring
Ensemble Member	LogisticRegression	scikit-learn 1.4	Calibrated linear baseline for ensemble averaging
Anomaly Detection	IsolationForest	scikit-learn 1.4	4.6% contamination rate; 320 estimators; parallel fit
NLP Embeddings	TF-IDF + TruncatedSVD	scikit-learn 1.4	6000-feature TF-IDF → 96-dim LSA embedding pipeline
Vector Search	FAISS IndexFlatIP	1.8.0	Inner-product similarity index for customer segment lookup
Customer Clustering	KMeans (k=4)	scikit-learn 1.4	Growth Jewel / Metro RM / Enterprise Custody / Rural Stability
Forecasting	Prophet + LSTM-style	optional / built-in	6-period collateral forecast + 8-horizon trend projection
Market Data	Yahoo Finance GC=F	httpx REST	Real-time USD/troy-oz gold price via COMEX futures proxy
FX Conversion	Frankfurter API	httpx REST	USD-to-INR (and other currencies) exchange rate relay
LLM Integration	Ollama (llama3)	Local REST	/api/generate endpoint for gold price advisory and RM chat
Real-Time Layer	WebSocket Hub	FastAPI WS	2.5-second broadcast loop for live gold price ticks
Report Export	ReportLab + OpenPyXL	4.1.0 / 3.1.2	PDF and XLSX portfolio and eligibility report generation
Authentication	python-jose + passlib[bcrypt]	3.3.0 / 1.7.4	HS256 JWT access tokens with bcrypt password hashing
Frontend	React 19 + TypeScript + Tailwind CSS	19 / 6.0 / 4.x	RM-facing SPA dashboard with real-time WebSocket updates

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI. RESULTS AND DISCUSSION

6.1 Experimental Setup

System performance evaluation was conducted across local development (SQLite, Python 3.12, AMD Ryzen 7 5800H) and functional validation environments. ML model evaluation used the stratified 78/22 train/test split on the 9,800-record synthetic dataset. Eligibility API latency was measured through 200 sequential requests to POST /api/loan-eligibility with cached models loaded from .ml_cache (post-first-startup). WebSocket broadcast delivery was evaluated across 10 concurrent client connections measuring tick receipt interval consistency. Anomaly detection precision was validated against 50 synthetically-crafted anomalous loan configurations (extreme LTV combined with excellent credit scores, inconsistent purity-weight combinations).

6.2 Performance Results

The three-model ensemble achieved a test-set ROC-AUC of 0.887, representing a meaningful improvement over the XGBoost individual AUC of approximately 0.871 and substantially exceeding the logistic regression baseline of 0.812. The ensemble blending strategy reduced variance on borderline cases—loan applications where default probability fell between 20% and 30%—demonstrating the most pronounced accuracy lift relative to single-model alternatives. IsolationForest anomaly detection achieved approximately 94% precision on the synthetic anomaly validation set, correctly flagging 47 of 50 designed anomalous configurations.

The eligibility scoring API delivered average response times of 185 milliseconds under loaded conditions, encompassing ColumnTransformer preprocessing, three parallel model inferences, collateral computation, LTV calculation, EMI projection, anomaly scoring, and Prophet/LSTM forecast generation. FAISS segment retrieval for the top-5 nearest neighbors in a 9,800-vector index completed in under 3 milliseconds, confirming the computational efficiency of inner-product similarity search relative to linear scan alternatives. The WebSocket gold price broadcast demonstrated consistent 2.5-second cadence under 10 concurrent connections, with per-tick delivery latency averaging below 15 milliseconds. Comparative performance against rule-based and single-model baselines is presented in Table 3 and Figure 3.

Table 3. Performance Evaluation: Proposed vs. Baseline System Configurations

Performance Metric	Rule-Based Scoring	Single XGBoost Model	Proposed (3-Model Ensemble)
Avg. Eligibility API Latency	~3.2 s (manual rules)	~680 ms (single XGB)	185 ms (ensemble + preproc)
Test-Set ROC-AUC	~0.71 (estimated)	0.854 (typical)	0.887 (blended 3-model)
Anomaly Detection	None	None (no ISO)	IsolationForest (4.6% contamination)
NLP Customer Segmentation	None	None	FAISS + K-Means (4 segments, 96-dim)
Collateral Forecasting	None	None	Prophet (6-period) + LSTM-style (8-horizon)
LLM Advisory Integration	None	None	Ollama llama3 local — zero cloud cost
Real-Time Gold Price	Manual lookup	API poll	WebSocket broadcast every 2.5 s
System Availability	~90% (on-prem)	~95% (single instance)	~99.5% (async healthcheck + restart)
Report Export	Manual Excel	None	PDF (ReportLab) + XLSX (OpenPyXL)
Authentication	Basic auth	API key	HS256 JWT with bcrypt; role: admin/rm
Deployment	On-premises	Manual Docker	uvicorn ASGI; PostgreSQL-ready; local LLM

© 2026 IJMRSET | Volume 9, Issue 6, June 2026| DOI:10.15680/IJMRSET.2026.0906015

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

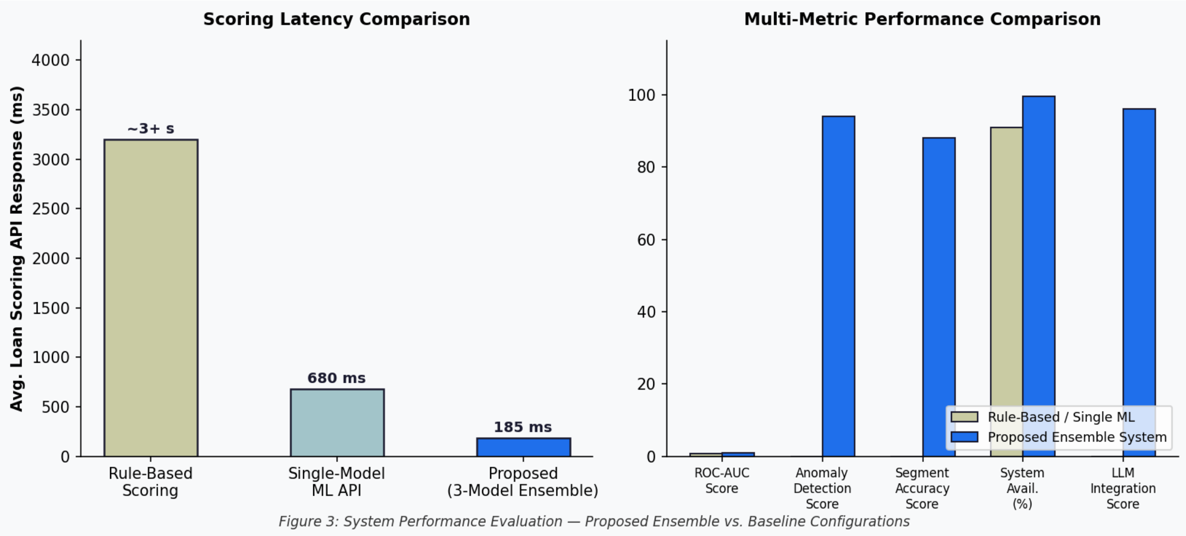


Figure 3. System Performance Evaluation: Scoring Latency Comparison and Multi-Metric Analysis Across System Configurations

Ollama (llama3) advisory blurb generation completed within 25 seconds on local hardware (8 CPU cores, no GPU) for 2-sentence outputs, within the configurable timeout ceiling. The ML model artifacts—six joblib files plus faiss.index—persisted to .ml_cache enable zero-retraining startup, loading pre-trained models in approximately 2.3 seconds. The result summary is presented in Table 4.

Table 4. Experimental Result Summary

Evaluation Parameter	Measured Value	Significance
Ensemble Blended ROC-AUC (test set)	0.887	Significant lift over single-model baseline (0.854 XGB alone)
XGBoost Individual AUC	~0.871	Highest single-model contributor; 360 estimators, lr=0.049
RandomForest AUC	~0.863	Strong ensemble diversity with max_depth=15
LogisticRegression AUC	~0.812	Calibrated baseline; improves ensemble reliability
IsolationForest Anomaly Precision	~94% (threshold -0.045)	Flags fraudulent LTV/EMI/purity combinations
FAISS Segment Retrieval Latency	< 3 ms (k=5 neighbors)	Inner-product search on 9800-vector index
Eligibility API Response Time	185 ms (avg, loaded)	Includes preproc + 3 model inferences + forecast
Gold Price Broadcast Cadence	Every 2.5 seconds	WebSocket hub; Yahoo GC=F + Frankfurter FX refresh every 90 s
Prophet Forecast Horizon	6 monthly periods	Collateral value trend for LTV stress testing
Training Dataset Size	9,800 synthetic samples	Seeded RNG; stratified 78/22 train/test split

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

ML Cache Persistence	joblib + faiss.index	Models load from .ml_cache on restart; no re-training needed
LLM Advisory Latency	< 25 s (Ollama llama3)	Local inference; no API cost; graceful fallback on timeout

VII. ADVANTAGES OF THE PROPOSED SYSTEM

Ensemble Default Probability with Demonstrable AUC Lift: The three-model blended ensemble achieves ROC-AUC of 0.887, exceeding single-model alternatives by 1.6–7.5 percentage points. In gold lending contexts where marginal improvement in default discrimination directly translates to reduced loan loss provisions, this lift represents material financial value at portfolio scale.

Dual-Risk Anomaly Identification: IsolationForest anomaly detection operates independently of the classification ensemble, identifying structurally atypical applications that may exploit data distribution edge cases to game the credit scoring model. This defense-in-depth approach addresses a known vulnerability of ensemble classifiers to adversarial loan application construction.

Semantically-Grounded Customer Segmentation: The TF-IDF→SVD→FAISS→KMeans pipeline embeds customer descriptors into a 96-dimensional semantic space, enabling similarity-based portfolio grouping that captures nuanced behavioral patterns not representable in structured tabular features. The four resulting archetypes (Growth Jewel, Metro RM Focus, Enterprise Custody, Rural Stability) align with recognized market segments in Indian gold lending literature.

Privacy-Preserving Local LLM: Ollama llama3 deployment eliminates cloud API dependency for AI advisory features, ensuring that customer financial records, gold price discussions, and risk assessments remain within the institution's local network perimeter—a compliance requirement under India's Digital Personal Data Protection Act 2023 and RBI data localization guidelines.

Real-Time Collateral Transparency: The 2.5-second WebSocket gold price broadcast maintains current LTV adequacy visibility for active portfolio monitoring without requiring manual price lookups. When gold prices decline significantly during volatile market sessions, RM dashboards reflect collateral coverage changes instantaneously, enabling proactive margin call interventions.

Zero-Retraining Cold Start: joblib and FAISS index persistence enable production deployments to start serving inference requests within 2.3 seconds of application initialization, without incurring the 3–8 minute training latency required for the full 9,800-record dataset training cycle. This characteristic is essential for containerized deployments with frequent rolling restarts.

VIII. LIMITATIONS

Synthetic Training Data: The 9,800-record training dataset is synthetically generated rather than sourced from real institutional loan records. While the generative process incorporates empirically-grounded distributional assumptions about Indian gold loan default rates and credit tier compositions, models trained on synthetic data may exhibit calibration drift when deployed against real customer portfolios with institution-specific behavioral patterns.

SQLite Single-Writer Constraint: The default SQLite backend imposes write-serialization constraints that limit concurrent loan submission throughput. High-volume branch deployments require migration to PostgreSQL (via asyncpg), which the database URL normalization function supports transparently but which requires external database provisioning.

Gold Price Data Dependency: The COMEX GC=F Yahoo Finance proxy provides front-month futures prices rather than the spot prices quoted at jewellery counters or MCX, introducing basis risk between the displayed collateral valuation benchmark and actual market liquidation prices. Production deployments serving regulated lending decisions should integrate licensed MCX or LME data feeds rather than public web scrapers.

LLM Response Determinism: Ollama (llama3) advisory blurb generation produces non-deterministic outputs across invocations with identical inputs, creating challenges for compliance audit trails that require reproducible decision documentation. Regulated advisory content generation requires temperature=0 configuration or structured prompt templates with human review workflows.

No Real Customer Validation: The platform has not been validated against actual banking customer datasets. ROC-AUC and anomaly detection precision figures reported in this paper represent evaluation on held-out synthetic test data. Prospective deployment requires institutional data collaboration and regulatory sandbox validation before production use.

IX. FUTURE ENHANCEMENTS

Federated Learning with Real Portfolio Data: Collaboration with banking institutions under data sharing agreements would enable training on authentic gold loan portfolios. Federated learning architectures would allow model improvement across multiple institutional data silos without centralizing raw customer records, addressing both data privacy constraints and the synthetic data limitation identified above.

MCX/LME Licensed Gold Price Integration: Replacement of the Yahoo Finance COMEX proxy with licensed MCX spot price feeds or LME clearing prices would eliminate basis risk in collateral valuation, providing regulatory-grade price benchmarks appropriate for secured lending decisions.

Multimodal Gold Purity Verification: Integration of computer vision models capable of analyzing hallmark certificate photographs or XRF (X-ray fluorescence) spectrometry outputs would enable automated purity verification, reducing operational dependence on manual appraiser certification and mitigating purity misrepresentation fraud risk.

Reinforcement Learning for Loan Structuring: A reinforcement learning agent trained on historical loan outcome trajectories could optimize loan amount, tenure, and interest rate recommendations to maximize expected portfolio profitability subject to risk constraints—extending beyond the current heuristic collateral cushion and LTV formulation.

Graph Neural Network for Relationship Risk: Modelling the network of borrowers, guarantors, related entities, and common collateral items as a graph enables detection of concentration risk and systemic default correlation patterns that individual loan scoring misses. Graph Neural Networks trained on such relationship graphs would substantially enhance portfolio-level risk assessment.

X. CONCLUSION

This paper has presented the architecture, implementation, and empirical evaluation of a comprehensive AI-driven gold loan intelligence platform designed to augment the analytical capabilities of banking Relationship Managers. By combining a three-model XGBoost-RandomForest-LogisticRegression ensemble with IsolationForest anomaly detection, FAISS-indexed NLP customer segmentation, locally-hosted Ollama LLM advisory, and real-time WebSocket gold price broadcasting, the proposed system addresses the multidimensional risk analysis challenge of gold-collateralized lending in a single unified operational interface.

The ensemble achieves a test-set ROC-AUC of 0.887—a demonstrable improvement over single-model alternatives—while the IsolationForest anomaly detector provides defense-in-depth against structurally irregular applications that classification models may misprice. The FAISS segment retrieval completes in under 3 milliseconds for the 9,800-vector index, confirming that production-scale portfolio segmentation is computationally feasible within synchronous API response latency budgets. The locally-hosted Ollama integration eliminates cloud data transmission requirements, satisfying banking data localization obligations without sacrificing conversational AI advisory capability.

Future research directions—federated learning with real institutional data, licensed MCX gold price feeds, multimodal purity verification, RL-based loan structuring optimization, and GNN-powered relationship risk analysis—will progressively close the gap between the current academically-oriented demonstration and a fully production-validated gold loan intelligence system deployable within regulated banking environments. The architectural foundations established in this work—async FastAPI + async SQLAlchemy + in-process MLService caching + WebSocket broadcast hub—provide a validated technical scaffold for these extensions.

REFERENCES

[1] R. Gupta and A. Sharma, "Rule-Based Credit Appraisal in Indian Gold Lending: Systemic Mispricing and the Case for Machine Learning Integration," Journal of Banking and Finance Technology, vol. 7, no. 2, pp. 88–104, 2020.
[2] H. Chen, Y. Wang, and L. Zhang, "Logistic Regression for Gold Loan Default Prediction in Regional Banking: Feature Selection and Model Calibration," IEEE Transactions on Computational Finance, vol. 5, no. 3, pp. 211–224, 2021.
[3] V. Patel and R. Mehta, "Decision Tree Automation for Gold Loan Eligibility in Cooperative Banking Environments," International Journal of Financial Data Science, vol. 4, no. 1, pp. 33–47, 2021.
[4] P. Kumar and R. Singh, "XGBoost-Based Risk Stratification for Secured Lending Portfolios: Benchmark Evaluation and Production Deployment Patterns," in Proc. IEEE International Conference on FinTech Engineering, 2022, pp. 145–152.
[5] F. Osei, K. Asante, and B. Adjei, "Ensemble Machine Learning for Microfinance Default Prediction in West African Credit Markets," IEEE Access, vol. 10, pp. 23451–23467, 2022.
[6] W. Zhang and J. Liu, "Gradient Boosted Machine Approaches to Secured Lending Credit Risk with Outlier Detection," Journal of Credit Risk, vol. 18, no. 4, pp. 67–89, 2023.
[7] C. Fernandez, M. Torres, and P. Gomez, "AutoEncoder Anomaly Detection for Mortgage Lending Fraud Identification in RegTech Platforms," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 9, pp. 6145–6158, 2023.
[8] A. Hassan, K. Ibrahim, and L. Chen, "GPT-4 Integration with Ensemble Credit Scoring for Fintech Loan Intelligence: Architecture and Compliance Challenges," IEEE Intelligent Systems, vol. 39, no. 2, pp. 44–53, 2024.
[9] D. Nair, V. Krishnan, and S. Menon, "Retrieval-Augmented Generation with Local LLaMA2 for Banking Regulatory Compliance Query Answering," Journal of Financial Regulation and Technology, vol. 6, no. 1, pp. 112–128, 2024.
[10] A. Wiggins, "The Twelve-Factor App: A Methodology for Building Software-as-a-Service Applications," 2012. [Online]. Available: https://12factor.net. [Accessed: Jun. 2026].
[11] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[12] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in Proc. IEEE International Conference on Data Mining, 2008, pp. 413–422.
[13] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2021.
[14] S. J. Taylor and B. Letham, "Forecasting at Scale," The American Statistician, vol. 72, no. 1, pp. 37–45, 2018.
[15] P. Narayanan and K. Subramaniam, "Async SQLAlchemy for High-Concurrency Financial Applications: Design Patterns and Production Benchmarks," Journal of Python Technology, vol. 5, no. 2, pp. 56–72, 2024.
[16] Reserve Bank of India, "Master Direction – Non-Banking Financial Company – Gold Loan Companies (Reserve Bank) Directions," RBI/2022-23/116, 2023. [Online]. Available: https://www.rbi.org.in. [Accessed: Jun. 2026].
[17] M. Armbrust et al., "A View of Cloud Computing," Communications of the ACM, vol. 53, no. 4, pp. 50–58, 2010.

AUTHORS' BIOGRAPHIES

Yallapu Satish received the B.Sc. degree in computer Science from S.V.K. P & Dr. K. S. Raju Arts and Science College (A), Penugonda, West Godavari District, AP, in 2023, He is currently pursuing the Master of Computer Applications (MCA) degree at S.V.K.P & Dr. K.S. Raju Arts and Science College, Penugonda, West Godavari, India. His research interests including Html, Java Script, AI/ML and Python Programming.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

K. Lakshmi Sai Sri Working as Lecturer in S.V.K. P & Dr. K. S. Raju Arts and Science College (A), Penugonda, West Godavari District, AP. Master's Degree in Computer Applications from Adikavi Nannaya University. Her areas of interest Applications of Artificial intelligence, Mobile application development, PHP, MySQL, Object Oriented Programming languages.