International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

MACHINE LEARNING AND LLM HYBRID SYSTEM FOR CREDIT RISK PREDICTION

[1] U. ARAVIND, [2] CHALUVADI RAVI TEJA, [3] JUNUTHULA SIVA SANKAR ACHARI, [4] CHALLA VENKATA SAI KIRAN, [5] KANDUKURI JAYARAMA CHARI, [6] BANDARU SRINIVAS

[1] ASST., PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, KRISHNA CHAITANYA INSTITUTE OF TECHNOLOGY & SCIENCES, DEVARAJUGATTU, MARKAPUR

[2,3,4,5,6] STUDENT, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, KRISHNA CHAITANYA INSTITUTE OF TECHNOLOGY & SCIENCES, DEVARAJUGATTU, MARKAPUR

ABSTRACT

The rapid growth of digital financial services has significantly increased the need for accurate and intelligent credit risk assessment systems. Traditional credit scoring models rely heavily on structured financial data and statistical techniques, often failing to capture complex patterns and unstructured information such as customer behavior, transaction narratives, and textual financial records. To address these limitations, this project proposes a Credit Risk Prediction System using Large Language Models (LLMs) that leverages advanced artificial intelligence techniques to enhance prediction accuracy and decision-making.

The proposed system integrates structured data (e.g., income, credit history, loan records) with unstructured data sources (e.g., customer profiles, financial statements, and textual reports) using LLMs. These models are capable of understanding natural language, extracting meaningful features, and generating contextual embeddings that improve risk classification. By combining LLM-based feature extraction with machine learning classifiers such as Gradient Boosting or Neural Networks, the system provides a more holistic evaluation of borrower risk.

Keywords: Credit Risk Prediction, Large Language Models (LLMs), Machine Learning, Natural Language Processing (NLP), Financial Analytics, Loan Default Prediction, Risk Assessment, Deep Learning, Explainable AI (XAI), Data Integration, Credit Scoring, Predictive Modeling, Structured and Unstructured Data Analysis

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

I. INTRODUCTION

In the modern financial ecosystem, credit risk assessment plays a critical role in determining the reliability of borrowers and ensuring the stability of lending institutions. With the rapid expansion of digital banking, online lending platforms, and fintech services, the volume and diversity of financial data have grown significantly. Traditional credit risk prediction systems primarily rely on structured data such as income levels, repayment history, credit scores, and loan records. These systems often use statistical models or conventional machine learning techniques, which may fail to capture complex relationships and patterns present in large and heterogeneous datasets.

One of the major limitations of existing systems is their inability to effectively utilize unstructured data, such as customer interactions, financial reports, transaction descriptions, and social or behavioral information. This type of data contains valuable insights into a borrower's financial behavior and intent, which can significantly improve the accuracy of credit risk evaluation. However, extracting meaningful information from such data requires advanced Natural Language Processing (NLP) techniques and powerful computational models.

Recent advancements in artificial intelligence, particularly in Large Language Models (LLMs), have opened new possibilities for enhancing credit risk prediction systems. LLMs are capable of understanding, processing, and generating human-like text by learning contextual relationships within large volumes of data. By leveraging these capabilities, LLMs can analyze both structured and unstructured financial information, extract relevant features, and provide deeper insights into borrower profiles.

II. LITERATURE REVIEW

The field of credit risk prediction has been extensively studied using traditional statistical methods and, more recently, advanced machine learning and deep learning techniques. Early research primarily focused on models such as Logistic Regression, Decision Trees, and Support Vector Machines (SVM), which rely on structured financial data like credit history, income, and repayment behavior. These models provided a strong baseline for risk classification but often lacked the ability to capture complex nonlinear relationships and hidden patterns in large datasets.

With the advancement of machine learning, researchers began exploring ensemble methods such as Random Forest and Gradient Boosting Machines. These approaches significantly improved prediction accuracy by

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

combining multiple weak learners and handling high-dimensional data effectively. Studies have shown that ensemble techniques outperform traditional models in terms of precision and recall, especially in large-scale banking datasets. However, these methods still primarily depend on structured inputs and do not fully utilize unstructured data sources.

In recent years, deep learning models such as Artificial Neural Networks (ANNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks have been applied to credit risk prediction. These models are capable of learning complex patterns and temporal dependencies in financial data, particularly in transaction sequences. Research indicates that deep learning approaches can enhance prediction performance, especially when dealing with time-series data. Nevertheless, they require large amounts of labeled data and are often criticized for their lack of interpretability.

EXISTING SYSTEM

The existing credit risk prediction systems used by banks and financial institutions are primarily based on traditional statistical and machine learning approaches. These systems focus mainly on structured financial data such as income level, employment history, credit score, past loan records, and repayment behavior. Common techniques include Logistic Regression, Decision Trees, and Support Vector Machines (SVM), which classify borrowers into risk categories based on predefined features.

In many cases, credit scoring models such as FICO or rule-based systems are used to evaluate the creditworthiness of applicants. These models rely on fixed rules and historical data patterns, making them simple to implement and interpret. With the advancement of machine learning, some institutions have adopted ensemble methods like Random Forest and Gradient Boosting to improve prediction accuracy. These models can handle large datasets and capture nonlinear relationships better than traditional statistical methods.

However, the existing systems have several limitations. One major drawback is their dependency on structured data alone, ignoring valuable unstructured information such as customer feedback, transaction descriptions, financial reports, and behavioral data. As a result, these systems may fail to capture the complete financial profile of a borrower. Additionally, many traditional models lack adaptability and struggle to perform well with dynamic and rapidly changing financial environments.

Another challenge is the limited capability of these systems to provide deep insights and

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

context-aware predictions. While some machine learning models improve accuracy,

they often act as “black boxes,” making it difficult for financial institutions to understand the reasoning behind predictions. This lack of transparency can reduce trust and create regulatory challenges.

PROPOSED SYSTEM

The proposed **Credit Risk Prediction System using LLMs** is designed to improve the accuracy, transparency, and efficiency of credit risk assessment. Unlike traditional systems that mainly depend on structured data, the proposed system uses both structured and unstructured financial information such as income, loan history, repayment records, transaction descriptions, customer profiles, bank statements, and financial text documents.

In this system, **Large Language Models (LLMs)** are used to analyze textual and financial data. The LLM extracts meaningful features from unstructured data and converts them into useful embeddings. These extracted features are combined with structured data and passed to machine learning models such as Gradient Boosting, Random Forest, or Neural Networks for final risk prediction. The system classifies borrowers into categories such as **low risk, medium risk, and high risk**.

The proposed system also includes an explainability module that helps banks understand why a borrower is classified into a particular risk category.

METHODOLOGY

The methodology of the Credit Risk Prediction System using LLMs involves a systematic process that integrates data collection, preprocessing, feature extraction, model training, and evaluation to accurately predict the creditworthiness of borrowers.

Initially, data is collected from multiple sources, including structured financial data such as income, credit history, loan details, repayment records, and demographic information, along with unstructured data like bank statements, transaction descriptions, and customer-related textual information. This combination ensures a comprehensive understanding of borrower behavior.

In the preprocessing stage, structured data is cleaned by handling missing values, removing inconsistencies, and normalizing features. For unstructured data, Natural Language Processing (NLP) techniques are applied, including tokenization, stop-word removal, and text normalization. This prepares the data for further analysis.

Next, Large Language Models (LLMs) are utilized for feature extraction from textual data. The LLM converts unstructured text into

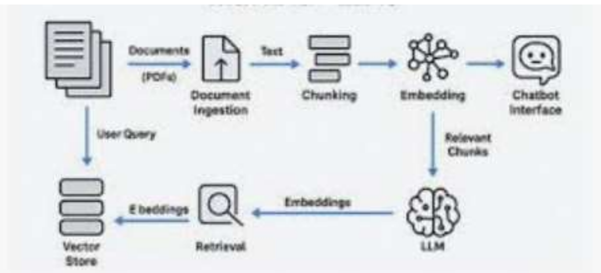
2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

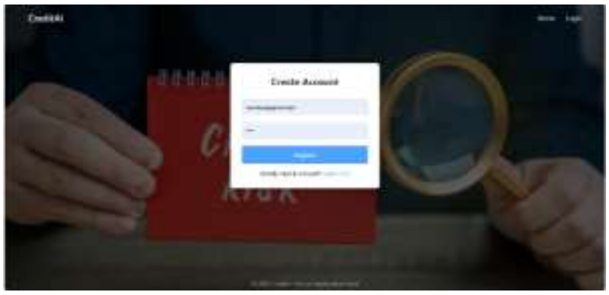
meaningful embeddings that capture semantic and contextual relationships. These embeddings are then combined with structured features to form a unified dataset.

VI. SYSTEM MODEL

System Architecture

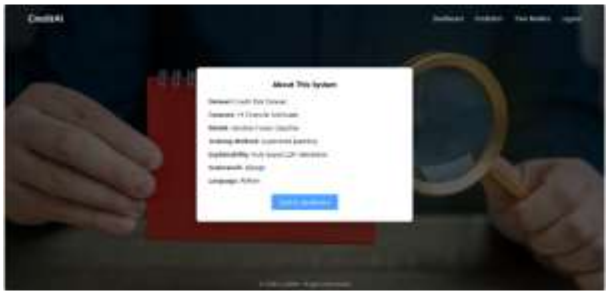


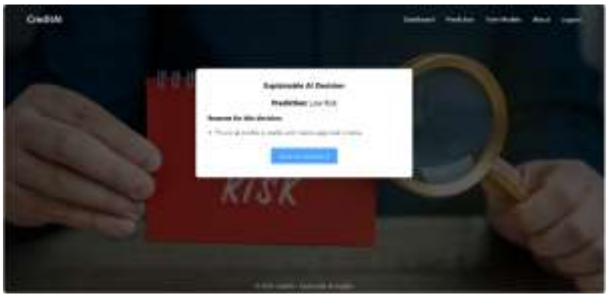
III. RESULTS AND DISCUSSIONS



2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026





VIII. CONCLUSION

The **Credit Risk Prediction System using LLMs** represents a significant advancement in financial risk assessment by integrating modern artificial intelligence techniques with traditional machine learning approaches. Unlike conventional systems that rely solely on structured data, the proposed system effectively utilizes both structured and unstructured data, enabling a more comprehensive understanding of borrower behavior and financial patterns.

By leveraging the capabilities of Large Language Models (LLMs), the system can extract meaningful insights from textual data such as transaction descriptions, financial documents, and customer profiles. This enhances the overall prediction accuracy and helps in identifying potential risks that may be overlooked by traditional models. The

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

integration of machine learning algorithms further strengthens the system's ability to classify borrowers into different risk categories with high reliability.

Moreover, the inclusion of explainable AI components ensures transparency in decision-making, which is crucial for gaining trust from financial institutions and complying with regulatory requirements. The system's ability to process data in real time and handle large-scale datasets makes it highly suitable for modern banking and fintech environments.

IX. FUTURE WORK: Future work for this

The proposed Credit Risk Prediction System using LLMs can be further enhanced in several ways to improve its performance, scalability, and real-world applicability. One important direction is the integration of more advanced and domain-specific LLMs trained specifically on financial datasets, which can provide deeper insights and more accurate feature extraction from complex financial documents.

Future work can also focus on incorporating real-time data streams such as live transaction data, social media signals, and market trends to enable dynamic and adaptive risk prediction. This would allow the system to respond quickly to changes in a borrower's financial behavior and external economic conditions.

Another area of improvement is enhancing model explainability using advanced Explainable AI (XAI) techniques. Providing more detailed and user-friendly explanations will help financial institutions better understand predictions and ensure regulatory compliance.

Additionally, efforts can be made to address data privacy and security concerns by integrating privacy-preserving techniques such as federated learning and secure data sharing mechanisms. This is especially important when handling sensitive financial and personal data.

XI. REFERENCES

➤ Thomas, L. C., Edelman, D. B., & Crook, J. N. – *Credit Scoring and Its Applications* – SIAM Publications, 2002.

➤ Hand, D. J., & Henley, W. E. – *Statistical Classification Methods in Consumer Credit Scoring: A Review* – Journal of the Royal Statistical Society, 1997.

➤ Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. – *Benchmarking State-of-the-Art Classification Algorithms for Credit*

Scoring – European Journal of Operational Research, 2015.

➤ Brown, I., & Mues, C. – *An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets* – Expert Systems with Applications, 2012.

➤ Chen, T., & Guestrin, C. – *XGBoost: A Scalable Tree Boosting System* – Proceedings of the ACM SIGKDD Conference, 2016.

➤ Breiman, L. – *Random Forests* – Machine Learning Journal, 2001.

➤ Hochreiter, S., & Schmidhuber, J. – *Long Short-Term Memory* – Neural Computation, 1997.

➤ Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. – *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* – NAACL, 2019.

➤ Vaswani, A., et al. – *Attention Is All You Need* – Advances in Neural Information Processing Systems (NeurIPS), 2017.

➤ Lundberg, S. M., & Lee, S. I. – *A Unified Approach to Interpreting Model Predictions (SHAP)* – NeurIPS, 2017.

➤ Goodfellow, I., Bengio, Y., & Courville, A. – *Deep Learning* – MIT Press, 2016.

➤ Baesens, B., Van Gestel, T., Stepanova, M., Van den Poel, D., & Vanthienen, J. – *Neural Network Survival Analysis for Personal Loan Data* – Journal of the Operational Research Society, 2005.

➢ J.V.ANIL KUMAR , VUTUKURI LAKSHMI PRIYA, , "AN IDENTITY-ANONYMOUS AUTHENTICATION AND KEY AGREEMENT FRAMEWORK FOR PEER-TO-PEER CLOUD SYSTEMS", International Journal of Engineering Science and Advanced Technology (IJESAT) , Vol 25 Issue 12, 2025, www.ijesat.com, https://doi.org/10.64771/ijesat.2025.039, Page 306 to 316, ISSN:2250-3676, 2025.

➢ J.V.Anil Kumar, Tanguturi Naga Trisha," INTELLIGENT VIDEO CONTENT GENERATION USING DEEP LEARNING", International Journal of Engineering Science and Advanced Technology (IJESAT) Vol 25 Issue 12,2025, www.ijesat.com, https://doi.org/10.64771/ijesat.2025.044, Page 357 to 364, ISSN:2250-3676, 2025.