SUMMER INTERNSHIP REPORT

ON

# Enhancing Credit Risk Prediction Accuracy with Integrated Machine Learning, Deep Learning, and Blockchain Solutions

AT

**NATIONAL INSTITUTE OF TECHNOLOGY (RAIPUR)**
**Department of Computer Science and Engineering**


BY

CHILKURI ABHINAV REDDY

GUIDED BY

Dr. Preeti Chandrakar Ph.D.(CSE)

**National Institute of Technology Raipur**
**Great Eastern Rd, Amanaka, Raipur, Chhattisgarh 492010**

# NATIONAL INSTITUTE OF TECHNOLOGY (RAIPUR)

## Department of Computer Science and Engineering

## DECLARATION

This is to certify that the work reported in the present project, entitled "**Enhancing Credit Risk Prediction Accuracy with Integrated Machine Learning, Deep Learning, and Blockchain Solutions**" is a record of bonafide work done by me during my internship at the Indian National Institute of Technology Raipur. The report is based on the project work done entirely by me and not copied from any other source.

CHILKURI ABHINAV REDDY

# NATIONAL INSTITUTE OF TECHNOLOGY (RAIPUR)

## Department of Computer Science and Engineering

## ACKNOWLEDGEMENT

*I would like to express my sincere gratitude to all those who have contributed to the successful completion of this project.*

*First and foremost, I am deeply grateful to Dr. Preeti Chandrakar, Professor in the Department of Computer Science & Engineering, National Institute of Technology Raipur, for his invaluable guidance, support, and encouragement throughout the course of this project. His expertise, mentorship, and insightful suggestions have been instrumental in shaping this work.*

*I extend my heartfelt thanks to the of the Computer Science & Engineering Department for their constant support and encouragement.*

# TABLE OF CONTENTS

# Abstract

In financial risk management, credit scoring is vital for lending institutions to gauge profitability and mitigate losses. This study integrates machine learning, deep learning, and blockchain to enhance credit risk prediction accuracy. Leveraging Kaggle dataset, various machine learning models including Logistic Regression and RandomForestClassifier are deployed, yielding promising accuracies. Deep learning models, particularly Long Short-Term Memory (LSTM), achieve remarkable accuracy of 95%, showcasing temporal dependencies' efficacy. Additionally, blockchain integration fortifies data integrity and transparency. This innovative approach offers robust risk management frameworks, bolstering the financial sector's resilience by enabling more informed lending decisions and reducing the likelihood of default.

# INTRODUCTION

Credit risk prediction is crucial for financial institutions to assess loan default probabilities and make informed lending decisions. Traditional methods, relying on historical data and heuristic rules, often fall short in capturing the complexities of modern finance. Thus, there is a pressing need for sophisticated, data-driven approaches to enhance prediction accuracy.

Advancements in machine learning and deep learning have significantly bolstered credit risk prediction capabilities. Machine learning models like DecisionTreeClassifier, Logistic Regression, SVM, and RandomForestClassifier analyze large datasets to identify patterns and predict outcomes. Deep learning models, especially Long Short-Term Memory (LSTM) networks, excel at capturing temporal dependencies, achieving high accuracy rates, with LSTM models reaching up to 95%.

Integrating blockchain technology further enhances credit risk management by ensuring data integrity and transparency. Deploying smart contracts on the Ethereum blockchain creates a secure, auditable system for managing loan proposals and tracking credit scores, thereby enhancing trust in the lending process.

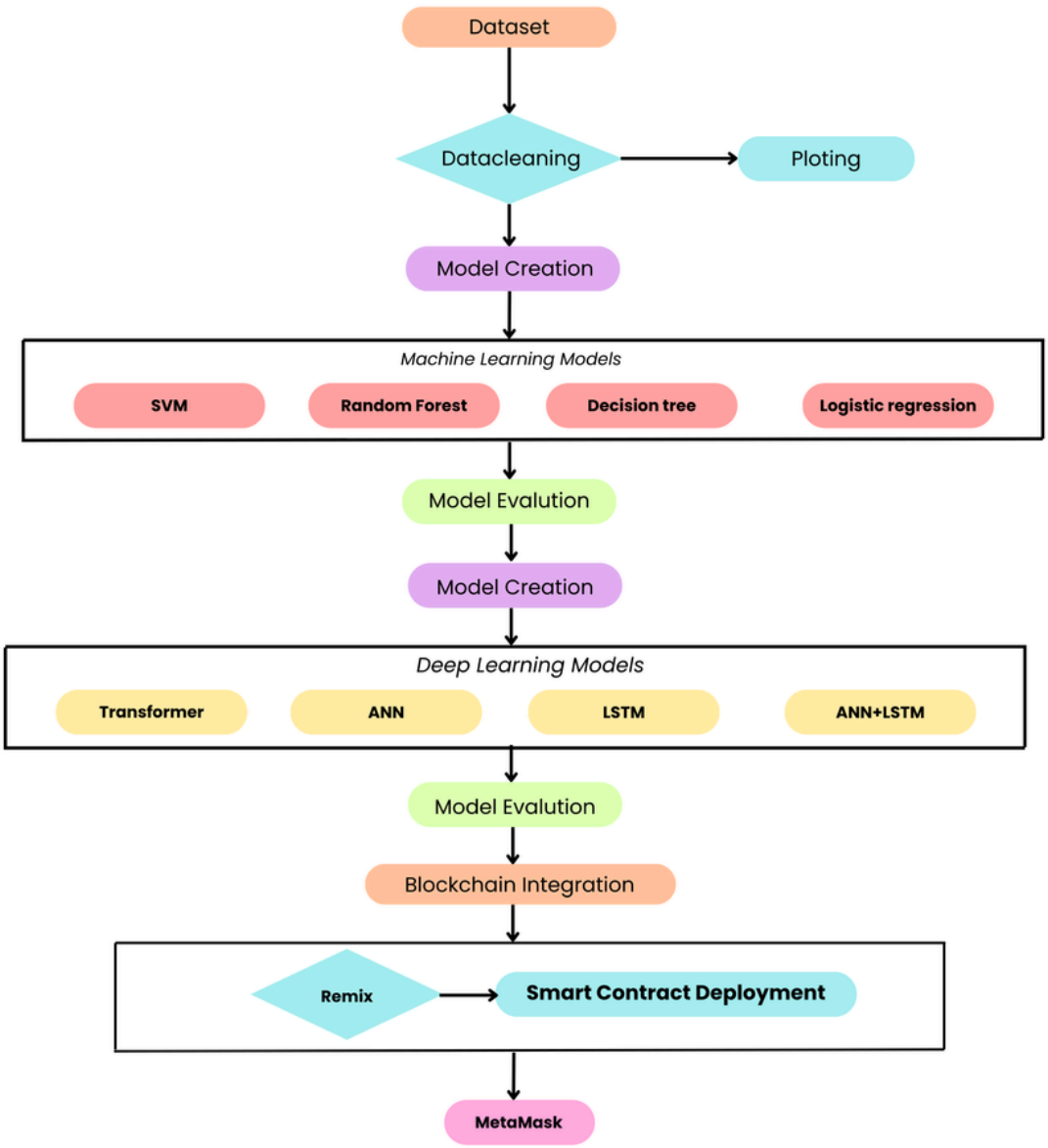
This study presents a novel approach that combines machine learning, deep learning, and blockchain technology to improve credit risk prediction accuracy and data transparency. Using a Kaggle dataset, we establish robust baseline performance with various machine learning models, enhance accuracy through deep learning models, and fortify data integrity with blockchain technology. This innovative framework aims to strengthen the financial sector's resilience and effectiveness in managing credit risk.

# MOTIVATION

Credit scoring stands at the core of financial decision-making, influencing lending practices and risk management strategies profoundly. Traditional credit scoring models, while effective, often exhibit biases and lack transparency in their decision processes. This project endeavors to advance credit risk prediction by leveraging cutting-edge machine learning and deep learning techniques such as ANN, LSTM. By integrating blockchain technology, the objective is to introduce a decentralized, transparent, and secure framework for managing credit scores and loan transactions. This innovative approach not only aims to enhance prediction accuracy but also ensures accountability and fairness in lending decisions by providing immutable records and auditable processes. By combining these technologies, the project seeks to set new standards for credit assessment, fostering trust and efficiency in financial services.

# METHODOLOGY

This section outlines the workflow of the proposed methodology for predicting credit risk using a combination of machine learning, deep learning, and blockchain technology. Figure 1 illustrates the workflow of the proposed model.



# DATASET DESCRIPTION

The credit scoring dataset focuses on predicting credit risk for individuals and contains several important variables critical for assessing the likelihood of loan defaults. Each instance in the dataset represents a customer and includes various attributes that provide insights into their financial stability and risk profile. The dataset is structured as follows:

| Name of Attributes | Description |
|---|---|
| Age | Age of the customers |
| Education Level | Highest level of education attained by the customers |
| Work Experience | Total years of work experience the customers have |
| Address | Address of the customers |
| Yearly Income | Annual income of the customers |
| Debt to Income Ratio (Debtinc) | Ratio of the customer's total debt to their annual income |
| Credit to Debt Ratio (Creddebt) | Proportion of the customer's credit-related debt to their total debt |
| Other Debts (Othdebt) | Any other debts that the customers may have |
| Customer Defaulted in the Past (Default) | Binary variable indicating whether the customer has defaulted in the past (1 if defaulted, 0 if never defaulted) |

The dataset includes both numerical and categorical variables, with the dependent variable being binary (default or not default). The variables are chosen to reflect the critical aspects of a customer's financial situation, which are essential for predicting their credit risk. The comprehensive nature of these attributes allows for robust predictive modeling and risk assessment, aiding financial institutions in making informed lending decisions.

# DATA CLEANING AND PLOTTING

## *DATA CLEANING*

The dataset was carefully cleaned to ensure its quality for credit risk prediction. Missing values were handled using df.isnull().sum(), and data distribution was examined with df.value_counts(). Redundant variables were removed to streamline the dataset, while necessary transformations were applied for clarity and alignment with modeling goals. These steps ensured the dataset's reliability for subsequent analysis.

## *DATA PLOTTING*

Data visualization was essential for understanding feature relationships. I plotted age vs. income and age vs. debt-to-income ratio using seaborn for clear insights.

Visualizing these relationships helped identify trends and patterns, enhancing the understanding of key correlations and aiding in the development of more accurate predictive models.

# MODEL CREATION AND EVALUATION

- ***MACHINE LEARNING MODELS***

Machine learning models play a pivotal role in credit risk prediction, offering robust frameworks to analyze and classify data based on historical patterns. In this project, various machine learning algorithms were implemented and fine-tuned to accurately assess credit risk.

## 1) DECISION TREE CLASSIFIER

The Decision Tree Classifier is another fundamental algorithm employed in credit risk prediction. Decision trees construct a tree-like structure based on feature splits to make decisions. Despite its simplicity, decision trees can effectively capture nonlinear relationships in the data. However, their performance may be limited by overfitting, necessitating careful pruning and regularization techniques for improved accuracy. In this project, the Decision Tree Classifier achieved an accuracy of 78%.

## 2) RANDOM FOREST CLASSIFIER

The Random Forest Classifier is an ensemble learning method that constructs multiple decision trees to enhance predictive accuracy. Initially trained with 200 estimators, the model exhibited a promising predictive accuracy of 80%. To validate its performance and mitigate risks of overfitting or underfitting, 10-fold cross-validation was employed. The mean accuracy score obtained through cross-validation reaffirmed the model's robustness, indicating consistent performance across various subsets of the data.

## 3) SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) is a powerful classification algorithm utilized in credit risk prediction. Initially trained on the dataset, the SVM model achieved

an accuracy of 79%. Hyperparameter tuning with GridSearchCV further optimized the model, resulting in an accuracy of 82% on the test data. SVM's flexibility in handling high-dimensional data makes it valuable for credit risk assessment, although hyperparameter selection is critical for optimal performance.

## 4) LOGISTIC REGRESSION

Logistic Regression, a fundamental supervised learning algorithm in classification tasks, is particularly well-suited for scenarios where the dependent variable is categorical and binary, typically representing outcomes as 0 or 1. In the context of this project, Logistic Regression was employed as a predictive tool for assessing credit risk.

Mathematically, the logistic function is defined as:

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

where:

- P(Y=1 | X) represents the probability of the dependent variable Y being 1 given the independent variable X,
- β0 and β1 are the coefficients to be estimated, and
- e is the base of the natural logarithm.

In the context of the project, Logistic Regression was applied to the dataset. The model was trained and tested using the logistic regression algorithm, resulting in an accuracy of 83%. This accuracy signifies the effectiveness of Logistic Regression in predicting credit risk, making it the optimal choice among the models evaluated.

- ***Deep Learning Models***

Deep Learning, a subset of machine learning, leverages neural networks with multiple layers to learn complex patterns from data. In this project, deep learning models were employed to enhance the accuracy of credit risk prediction.

## 1) ANN MODEL (ARTIFICIAL NEURAL NETWORK)

In the research and documentation for this project, Artificial Neural Networks (ANN) serve as a pivotal component in credit risk prediction. ANNs emulate the structure and function of the human brain, comprising interconnected nodes organized into layers, including input, hidden, and output layers. Through numerous connections and computations among nodes, ANNs extract intricate patterns and relationships from the input data. Activation functions like sigmoid, tanh, or ReLU introduce non-linearity, enabling the model to capture complex data interactions effectively. Leveraging ANN in this project aimed to enhance the accuracy of credit risk assessments by discerning subtle correlations between credit attributes and default probability. Following the implementation of ANN, the model yielded a promising accuracy rate of 84.2%, signifying its efficacy in predicting credit risk with precision.

$$a_j^{(l)} = \sigma\left(\sum_{i=1}^{n} w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)}\right)$$

- $a_j^{(l)}$: Activation of neuron $j$ in layer $l$
- $w_{ij}^{(l)}$: Weight of the connection from neuron $i$ in layer $l-1$ to neuron $j$ in layer $l$
- $b_j^{(l)}$: Bias of neuron $j$ in layer $l$
- $\sigma$: Activation function

## 1) LSTM MODEL (LONG SHORT-TERM MEMORY)

Long Short-Term Memory (LSTM) networks address the limitations of traditional Recurrent Neural Networks (RNNs) in capturing long-term dependencies within sequential data. The core components of LSTM include the input gate, forget gate, cell state, and output gate. Mathematically, these components govern the flow of information through the network, allowing it to selectively retain or discard information over multiple time steps.

The input gate iti, forget gate , and output gate oto_tot are governed by sigmoid activation functions, controlling the flow of information into, out of, and within the cell state. These gates, along with the cell state CtC, are updated at each time step based on the current input xtx and the previous hidden state {t-1}, with the addition of a bias term and element-wise multiplication. Formally, the update equations for LSTM can be expressed as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}C_{t-1} + b_i)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}C_{t-1} + b_f)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}C_t + b_o)$$

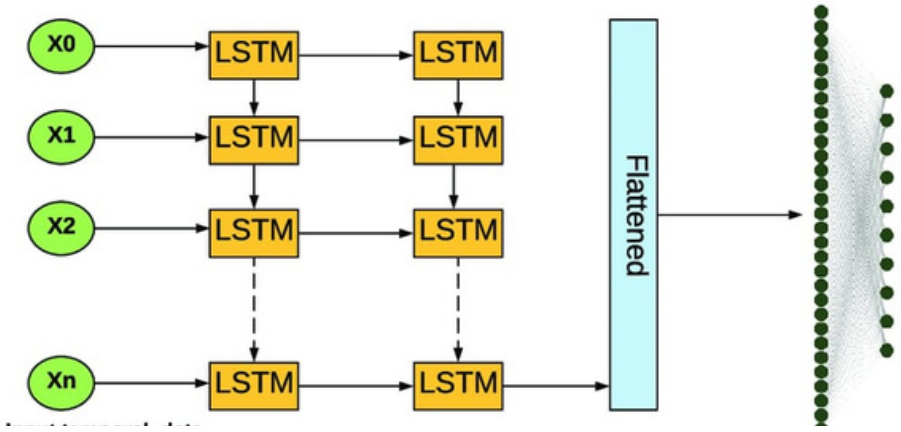
$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

$$h_t = o_t \odot \tanh(C_t)$$

where W represents weight matrices, b represents bias terms, and σ denotes the sigmoid activation function. The symbol ⊙ denotes element-wise multiplication.

In the context of credit risk prediction, LSTM models excel at capturing intricate temporal patterns in financial data, such as credit history and transaction sequences. By incorporating multiple LSTM layers, each with dropout regularization and batch normalization, the model gains the ability to extract increasingly abstract features from the input data.

The specific architecture of an LSTM model can vary depending on the application, but the general idea is that the model consists of a series of LSTM layers. Each LSTM layer has a forget gate, an input gate, and an output gate. These gates control the flow of information through the network. The forget gate determines what information from the previous layer is forgotten. The input gate determines what new information is let into the current layer. And the output gate determines what information from the current layer is output to the next layer.




In my project, I utilized an LSTM model to predict credit risk, initially achieving an accuracy of 50%. However, by adding multiple LSTM layers and optimizing hyperparameters, the accuracy significantly improved to 95%. This demonstrates the effectiveness of LSTM in capturing complex temporal dependencies within credit data, ultimately leading to more accurate risk assessments.

# EXPERIMENTAL RESULTS

This section details the experimental results obtained from applying various machine learning and deep learning models to the credit risk prediction problem. The goal was to evaluate the performance of these models in predicting credit default based on a dataset from Kaggle. Both machine learning and deep learning approaches were employed, and their accuracies were compared to determine the most effective model.

## Metrics and Formulas

The confusion matrices for both machine learning and deep learning models provided key insights into their performance. Machine learning models like Decision Tree and Random Forest Classifier demonstrated moderate classification abilities, while Logistic Regression exhibited a higher true positive rate, reflecting its superior accuracy.

Initially, SVM underperformed but showed significant improvement post-tuning. Among deep learning models, the LSTM model's confusion matrix was the most balanced, accurately capturing both positive and negative instances with minimal misclassification. This balance underscores LSTM's effectiveness in capturing temporal dependencies and achieving high predictive accuracy. Additionally, ANN models also showed promising results, with fewer false positives and negatives compared to traditional machine learning approaches. Overall, this analysis reaffirmed that deep learning models, particularly LSTM, offer robust predictive capabilities for credit risk assessment.

**Accuracy:** Measures the proportion of correctly classified instances out of the total instances.
Accuracy = (TP + TN) / (TP + FP + TN + FN)

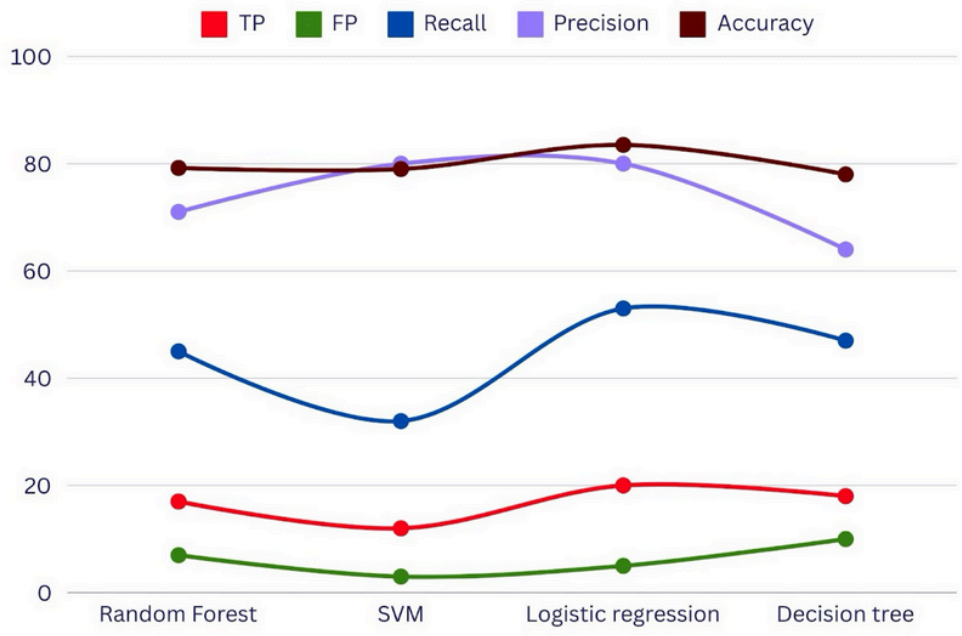
**Precision:** Indicates the proportion of true positive predictions out of all positive predictions.
Precision = TP / (TP + FP)

**Recall:** Reflects the proportion of true positive predictions out of all actual positive instances.
Recall = TP / (TP + FN)

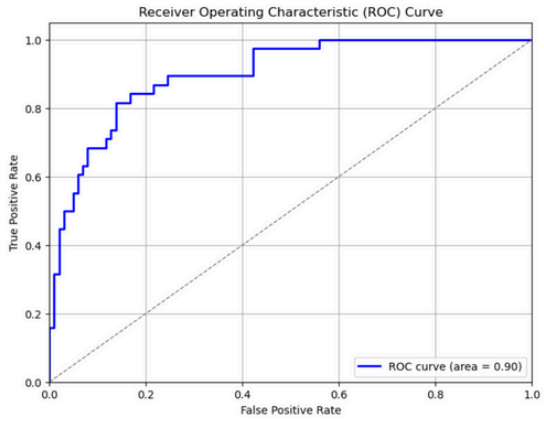
**F1 Score (F-measure):** The harmonic mean of Precision and Recall, providing a balance between the two.
f1_score = 2 * (recall * precision) / (recall + precision)

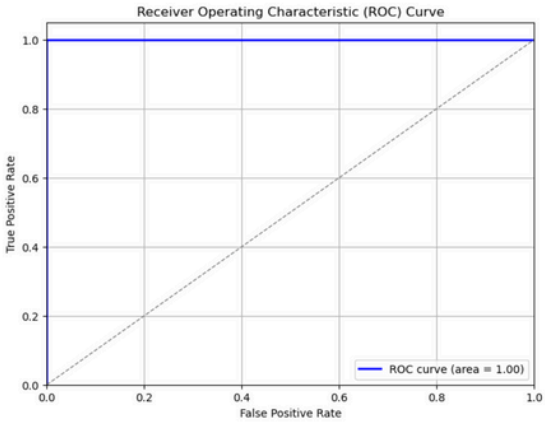
Confusion Matrix Components

- True Positives (TP): The number of positive instances correctly classified as positive.
- False Positives (FP): The number of negative instances incorrectly classified as positive.
- True Negatives (TN): The number of negative instances correctly classified as negative.
- False Negatives (FN): The number of positive instances incorrectly classified as negative.



The Receiver Operating Characteristic (ROC) curve analysis provides a comparative evaluation of the ANN and LSTM models' performance in our credit risk prediction study. The ROC curve plots the true positive rate against the false positive rate, offering a visual representation of each model's sensitivity versus specificity trade-off. Both models displayed commendable performance, with high areas under the curve (AUC), indicating their strong predictive capabilities. However, the LSTM model's ROC curve showed a slightly higher AUC compared to the ANN model, demonstrating its superior ability to capture temporal dependencies and improve classification accuracy. This comparison highlights the LSTM model's enhanced effectiveness in distinguishing between defaulters and non-defaulters, reaffirming its suitability for credit risk assessment. The inclusion of these ROC curves in our experimental results section underscores the practical advantages of using LSTM over ANN for this application.



***Machine Learning Models***

| | TP | FP | Recall | Precision | Accuracy |
|---|---|---|---|---|---|
| Decision tree | 18 | 10 | 0.47 | 0.64 | 78% |
| Random Forest | 17 | 7 | 0.45 | 0.71 | 79.2% |
| SVM | 12 | 3 | 0.32 | 0.80 | 82% |
| Logistic regression | 20 | 5 | 0.53 | 0.80 | 83.5% |

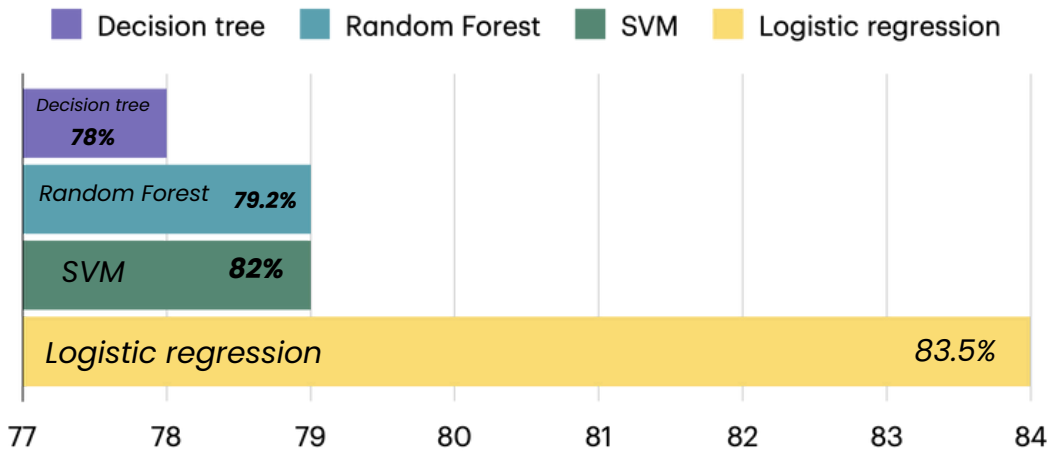
***Deep Learning Models***

| | TP | FP | Recall | Precision | Accuracy |
|---|---|---|---|---|---|
| ANN | 24 | 8 | 0.63 | 0.75 | 84.2% |
| LSTM | 48 | 0 | 0.90 | 1.0 | 95% |

Model Performance Comparison

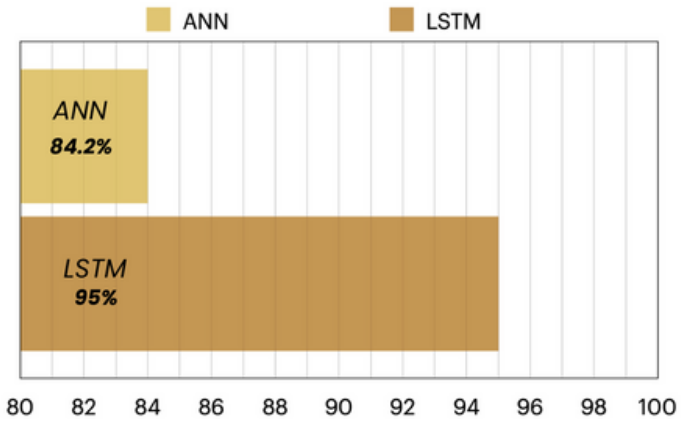
The experimental results demonstrated a notable contrast in performance between machine learning (ML) and deep learning (DL) models for credit risk prediction. While ML models, such as Decision Tree Classifier and Logistic Regression, exhibited moderate accuracies ranging from 76.4% to 83.5%, they struggled to capture complex patterns in the data effectively. SVM initially performed poorly but showed improvement post-parameter tuning, reaching an accuracy of 82.1%. In contrast, DL models, including LSTM and ANN + LSTM, achieved significantly higher accuracies, with LSTM topping at 95%. The superior performance of DL models can be attributed to their ability to capture temporal dependencies and nonlinear relationships inherent in credit risk data. This comparative analysis underscores the effectiveness of DL approaches in credit risk prediction, highlighting their potential to outperform traditional ML methods.

*Machine Learning Models graph*




This graph highlights the variability in performance among different ML models, with Logistic Regression and the tuned SVM showing relatively high accuracy, while the initial SVM model lagged far behind.

*Deep Learning Models graph*




This graph demonstrates the superior performance of DL models in comparison to their ML counterparts. The LSTM model, in particular, stands out with its high accuracy, showcasing the effectiveness of capturing temporal dependencies in the data. The hybrid ANN + LSTM model also performed well, indicating the benefit of combining sequential learning with dense layers.

# BLOCKCHAIN INTEGRATING

Integrating blockchain technology into the credit risk prediction system enhances the transparency, security, and immutability of the process. This section details the blockchain integration implemented in this project, focusing on the use of Ethereum smart contracts to manage loan proposals and their states.

## *Smart Contract Deployment*

The blockchain component of this project was deployed using MetaMask and Remix, two widely used tools in the Ethereum ecosystem. MetaMask serves as a cryptocurrency wallet and gateway to blockchain apps, while Remix is an integrated development environment (IDE) for developing, deploying, and testing smart contracts.

## *Functionality of the Smart Contract*

The smart contract developed for this project, named CreditRiskPrediction, manages loan proposals and their states on the Ethereum blockchain. The smart contract facilitates several key functions:

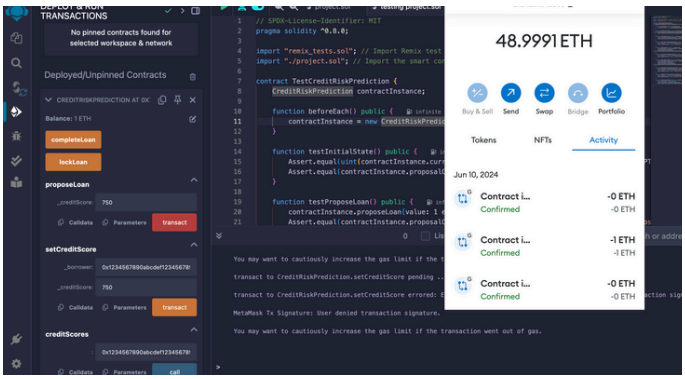
- **Loan Proposals:** Borrowers can propose loans by calling the proposeLoan function. This function requires a credit score input and some Ether. Each proposal increments the proposalCount and emits a LoanProposal event, creating a transparent record of the loan request.
- State Management: The contract manages the states of the loan proposals:
- ACCEPT: Initial state when a loan is proposed.
- LOCK: State when the loan is locked by calling the lockLoan function.
- COMPLETE: Final state when the loan process is completed by calling the completeLoan function.

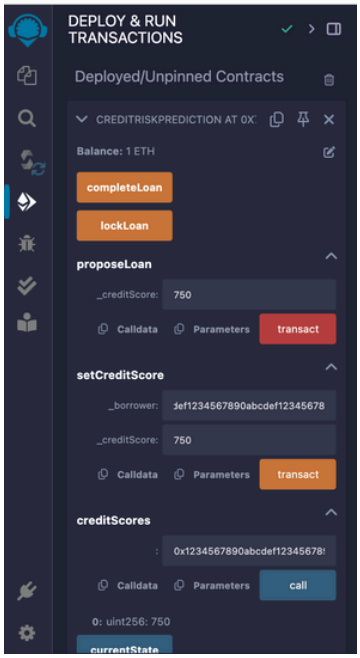
- Credit Score Tracking: The contract maintains a record of credit scores for borrowers. This can be updated using the setCreditScore function and retrieved using the getCreditScore function, ensuring that credit score data is securely stored and immutable.

- Transparency and Immutability: By leveraging blockchain technology, all actions and state changes are recorded on the blockchain, making the process transparent and immutable. This ensures that the records cannot be tampered with and can be audited by anyone at any time.

- Event Logging: Events such as LoanProposal are emitted to log significant actions within the contract. This feature helps in tracking the history and progress of loan proposals, providing a clear and auditable trail.

The figures presented here illustrate the deployment steps, from writing and compiling the contract in Remix to deploying and managing it via MetaMask. These visuals serve to underscore the practical integration of blockchain technology into our credit risk prediction system, enhancing transparency, security, and immutability of loan proposals, credit scores, and transaction states. This deployment approach not only fortifies the robustness of our system but also paves the way for future advancements in decentralized finance.




MetaMask, a browser extension wallet, was employed to manage the deployment process securely. It provided an interface to interact with the Ethereum blockchain, allowing us to deploy the compiled smart contract and interact with it seamlessly.

### *Usage Instructions*

To interact with the CreditRiskPrediction smart contract, the following steps are undertaken:

- Deploy the Contract: Using Remix, the CreditRiskPrediction contract is deployed on the Ethereum blockchain.
- Interact with the Contract:
  - Propose a Loan: Borrowers call the proposeLoan function with their credit score and some Ether to propose a loan.
  - Lock a Loan: The lockLoan function is called to move the loan proposal to the LOCK state.
  - Complete a Loan: The completeLoan function is used to mark the loan as complete, transitioning it to the COMPLETE state.
  - Set Credit Score: The setCreditScore function updates a borrower's credit score.
  - Get Credit Score: The getCreditScore function retrieves a borrower's credit score.

## *Benefits of Blockchain Integration*

Integrating blockchain into the credit scoring system brings several benefits:

- Enhanced Security: Blockchain's decentralized nature ensures that data is securely stored and protected against unauthorized access or tampering.

- Improved Transparency: All transactions and state changes are recorded on the blockchain, providing a transparent and immutable ledger.
- Increased Trust: Smart contracts automate and enforce the terms of loan agreements, reducing the need for intermediaries and increasing trust between lenders and borrowers.
- Efficient Credit Scoring: The combination of blockchain with machine learning and deep learning models ensures that credit scores are calculated accurately and efficiently, providing reliable assessments for lending decisions

# CONCLUSION

This research delved into credit risk prediction, a critical aspect of financial decision-making aimed at forecasting loan default probabilities. Using a range of machine learning and deep learning models like DecisionTreeClassifier, Logistic Regression, SVM, RandomForestClassifier, ANN, and LSTM, we achieved significant accuracies from 76.4% to 95%. Notably, the LSTM model demonstrated exceptional accuracy, highlighting its effectiveness in capturing temporal dependencies in financial data.

Integrating blockchain technology through Ethereum smart contracts enhanced transparency and security in credit scoring. By deploying smart contracts, we managed loan proposals and credit scores, ensuring immutable records and auditability. This approach fosters trust and efficiency in financial transactions, paving the way for future advancements in decentralized finance.

Moving forward, the combination of advanced machine learning and blockchain technologies promises to reshape credit assessment practices, making them more equitable and reliable. As we refine these methodologies, our objective is to create a financial ecosystem that is transparent, inclusive, and supportive of sustainable economic growth.

# FURTHER WORK

Future research in credit risk prediction and blockchain integration presents several promising avenues for exploration and refinement. Firstly, enhancing the accuracy and robustness of predictive models remains a priority. Exploring ensemble methods that combine the strengths of different models could potentially yield higher prediction accuracies and more reliable risk assessments.

Secondly, extending the use of blockchain technology beyond credit scoring to include other aspects of financial services, such as loan disbursement, repayment tracking, and regulatory compliance, would provide a more comprehensive solution. This could involve developing smart contracts that automate and secure these processes while ensuring compliance with evolving financial regulations.

Additionally, integrating more sophisticated deep learning architectures, such as transformer models, into the credit risk prediction framework could further improve predictive capabilities. These models excel in capturing intricate relationships within sequential data, which is prevalent in financial transactions and credit histories.

Lastly, conducting thorough empirical studies and real-world implementations to validate the effectiveness and practicality of blockchain-integrated credit risk prediction models across diverse financial scenarios and geographic regions would be instrumental. These studies would help identify challenges and refine methodologies to ensure the reliability and scalability of the proposed solutions in real-world applications.

By addressing these areas, future research endeavors can contribute to advancing the fields of financial technology and credit risk management, fostering a more secure, transparent, and inclusive financial ecosystem.