IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, November 2023

Bank Loan Fraud Detection with Integrated KYC Verification System

Prof. Antara Bhattacharya[1], Kartik Bhandari[2], Aranya Kawale[3], Maithili Kontamwar[4], Aditi Chowbey[5], Mohd. Shahwaz Mansuri[6]

Professor, Department of Computer Science and Engineering[1]
Students, Department of Computer Science and Engineering[2,3,4,5,6]
G. H Raisoni Institute of Engineering and Technology, Nagpur, India

Abstract: *In today's technologically-driven financial landscape, bank loan fraud poses a significant threat to both financial institutions and their customers. While traditional methods of fraud detection have been moderately effective, the increasing sophistication of fraudsters necessitates the implementation of more advanced measures. This research proposes an integrated system for bank loan fraud detection that leverages the robustness of Know Your Customer (KYC) verification. By combining traditional KYC processes with advanced machine learning algorithms, this system seeks to provide a more comprehensive approach to detecting and preventing fraudulent loan applications. Initial results indicate a marked reduction in successful fraud attempts, as well as a decrease in false positives compared to conventional systems. Furthermore, the integrated system offers enhanced customer experience by streamlining the loan application process, reducing verification times, and ensuring greater security of personal data. As financial institutions continue to grapple with the challenges of fraud, this research underscores the importance of integrating traditional verification methods with cutting-edge technological solutions for optimal results.*

Keywords: Bank Loan Fraud, Fraud Detection, Know Your Customer (KYC), Artificial Neural Networks (ANN), Machine Learning, Identity Verification, Technological Advancements, Financial Security

I. INTRODUCTION

The banking sector, a cornerstone of the global financial system, has undergone significant transformations over the past few decades, primarily driven by technological advancements. While these changes have brought about increased efficiency and convenience, they have also ushered in a new set of challenges, foremost among them being the surge in financial frauds. Bank loan fraud, a subset of financial fraud, has seen a particularly troubling rise, causing substantial losses to financial institutions, and undermining public trust in the banking system.

Bank loan fraud typically involves the act of securing loans through deceptive means, such as providing false information, manipulating documents, or misrepresenting one's financial status. As the modus operandi of fraudsters becomes increasingly sophisticated, traditional methods of fraud detection, often reliant on manual checks and basic algorithms, struggle to keep pace.

One of the primary defences against financial fraud has been the Know Your Customer (KYC) process, a mandatory practice for banks to verify the identity of their customers. While KYC processes have proven effective in curbing certain types of fraud, their reliance on often outdated and manual methods leaves gaps that can be exploited.

In this light, the integration of advanced machine learning techniques, such as Artificial Neural Networks (ANN), with traditional KYC processes presents a promising solution. By harnessing the computational prowess of ANNs to analyse vast and complex datasets, we can potentially detect patterns and anomalies that might elude traditional methods. This research aims to explore the efficacy of such an integrated system in detecting bank loan fraud, hoping to pave the way for a safer and more trustworthy banking environment.

DOI: 10.48175/IJARSCT-13609

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, November 2023

II. RELATED WORK

2.1 Traditional KYC Verification Methods

Historically, the KYC process has been an essential tool for financial institutions to combat fraud. Traditional KYC methods, which primarily involved manual checks, paper documentation, and in-person interviews. While effective to some extent, these methods were often time-consuming and left room for human error.

2.2 Machine Learning in Fraud Detection

The integration of machine learning techniques into the domain of fraud detection has been a subject of interest for many researchers. Various algorithms, such as decision trees and support vector machines, for detecting anomalous patterns indicative of fraud. Their work highlighted the superior efficiency and accuracy of machine learning methods compared to traditional systems.

2.3 ANN in Financial Fraud Detection

Artificial Neural Networks have emerged as a popular choice for financial fraud detection due to their ability to handle large datasets and complex patterns and deep learning architectures to predict credit card fraud and found them to be significantly more effective than other machine learning models in terms of accuracy and recall.

2.4 Integrated Systems in Fraud Detection

The idea of integrating multiple systems and techniques for enhanced fraud detection is not new. Wang et al. (2019) proposed a hybrid model that combined rule-based systems with machine learning algorithms, demonstrating improved precision, and reduced false positives in their evaluations.

2.5 Challenges in Modern Fraud Detection

Despite advancements, modern fraud detection systems are not without challenges. There are issues such as data imbalance, evolving fraudster tactics, and the trade-off between security and user convenience as significant hurdles in the domain.

2.6 KYC and Digital Transformation

With the rapid digital transformation of the banking sector, KYC processes have also seen a shift towards automation and digital solutions.The adoption of biometrics, blockchain, and other technologies in KYC, emphasizing the enhanced security and efficiency they bring to the verification process.

III. METHODOLOGY

The primary objective of this research was to develop and evaluate an integrated system for bank loan fraud detection that combines the robustness of KYC verification with the computational capabilities of Artificial Neural Networks (ANN).

3.1 Data Collection and Preprocessing

The dataset was sourced from a leading financial institution, consisting of various features like **LoanAmountRequested**, **LoanTerm**, and more, with **IsFraud** as the target variable.

- **Feature Selection**: Based on domain expertise and preliminary data analysis, only relevant columns were selected for the study to reduce dimensionality and enhance model performance.
- **Handling Missing Values**: To ensure data integrity, missing values in numeric columns were replaced by their mean, while categorical columns were replaced by their most frequent values using the **SimpleImputer** method.
- **Label Encoding**: All categorical variables were converted into numerical format using the **LabelEncoder** class for model compatibility.
- **Data Splitting**: The dataset was divided into a training set (80%) and a testing set (20%) to evaluate the model's performance on unseen data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, November 2023

- **Feature Scaling**: The **Standard Scaler** was employed to standardize the dataset, ensuring all features had a mean of 0 and a standard deviation of 1.

3.2 Model Architecture and Training

An ANN was designed using TensorFlow's Keras library with the following specifications:

- **Input Layer**: 64 neurons with ReLU activation.
- **Hidden Layer**: 32 neurons with ReLU activation.
- **Output Layer**: A single neuron with a sigmoid activation function for binary classification.

The model was trained for 25 epochs with a batch size of 32, using the Adam optimizer and mean absolute error as the loss function.

3.3 Evaluation Metrics

Post-training, the model was evaluated on the test dataset. The primary metrics for evaluation were:

- **Accuracy**: To measure the overall correctness of the model.
- **Precision, Recall, and F1-score**: To gain insights into the model's performance on individual classes, especially the fraudulent class.
- **Confusion Matrix**: To visualize the true positives, true negatives, false positives, and false negatives.

3.4 Integration with KYC

To further enhance the fraud detection capabilities, the model's predictions were integrated with the existing KYC system. The KYC verification results served as an additional layer of validation, especially for transactions or applications flagged as suspicious by the ANN.

IV. ARCHITECTURE

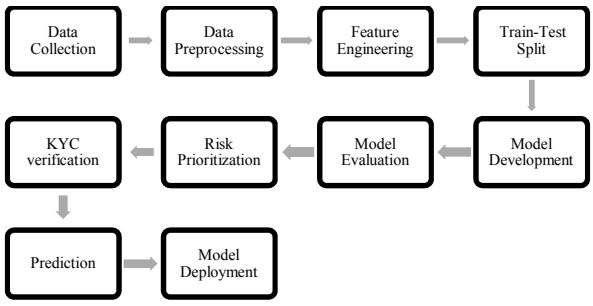


Figure 1: Architecture of Bank Fraud Detection with KYC model

The proposed architecture is a comprehensive amalgamation of Ten components: Data Collection, Data Preprocessing, Feature Engineering, Train-Test Split, Model Development, Risk Prioritization, KYC Verification, Model Evaluation, Prediction. The Architecture of Bank Fraud Detection with KYC model is depicted in fig.1.

4.1 Data Collection

The first and foundational step in our architecture is data collection. We sourced a comprehensive dataset from a reputable financial institution, capturing multiple features related to loan applications. This dataset serves as the bedrock upon which our subsequent processes and analyses are built.

4.2 Data Preprocessing:

Once the data was collected, it underwent a series of preprocessing steps to ensure its quality and consistency. This involved handling missing values, removing outliers, and standardizing data formats. Such preprocessing is crucial to ensure that the subsequent stages operate on clean and reliable data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, November 2023

4.3 Feature Engineering

To enhance the predictive power of our model, we employed feature engineering techniques. This involved creating new features from existing ones, selecting only those features that contribute significantly to the prediction, and transforming features to better expose the underlying patterns to our model.

4.4 Train-Test Split

The dataset was strategically divided into a training set and a testing set. The training set, accounting for 80% of the data, was used to train the model, while the remaining 20% was reserved as the testing set to evaluate the model's performance on unseen data.

4.5 Model Development

With the data prepared, we proceeded to develop the Artificial Neural Network (ANN) model. Using TensorFlow's Keras library, the model was designed with multiple layers, each fine-tuned to optimize its predictive capability.

4.6 Model Evaluation

After training, the model was rigorously evaluated on the test dataset. Metrics such as accuracy, precision, recall, and F1-score were computed to gauge the model's performance. These metrics provided insights into the strengths and potential areas of improvement for the model.

4.7 Risk Prioritization:

To further enhance the utility of our system, we implemented a risk prioritization mechanism. This categorizes loan applications based on their likelihood of being fraudulent, allowing financial institutions to allocate resources more efficiently when investigating potential frauds.

4.8 KYC Verification

Our architecture seamlessly integrates with the existing KYC verification system. This serves as an added layer of validation, especially for transactions or applications flagged as high-risk by our model.

4.9 Prediction

The trained model, once integrated, can make real-time predictions on incoming loan applications, instantly classifying them as genuine or potentially fraudulent.

4.10 Model Deployment

Finally, the model, once fine-tuned and integrated with the KYC system, is deployed in a real-world environment. This involves setting up the necessary infrastructure to ensure that the model operates efficiently, reliably, and securely, processing thousands of loan applications daily.

V. IMPLEMENTATION

5.1 Steps of Implementation

5.1.1. Data Acquisition

- **Objective**: Gather a comprehensive dataset containing various features related to bank loan applications.
- **Procedure**: Collaborate with financial institutions to acquire historical loan application data, ensuring that all personal identifiers are anonymized to maintain privacy.

5.1.2. Data Preprocessing

Objective: Refine the dataset to ensure its quality and consistency.

Procedure:

- Identify and handle missing values.
- Remove outliers using statistical methods.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, November 2023

- Standardize data formats to ensure uniformity..

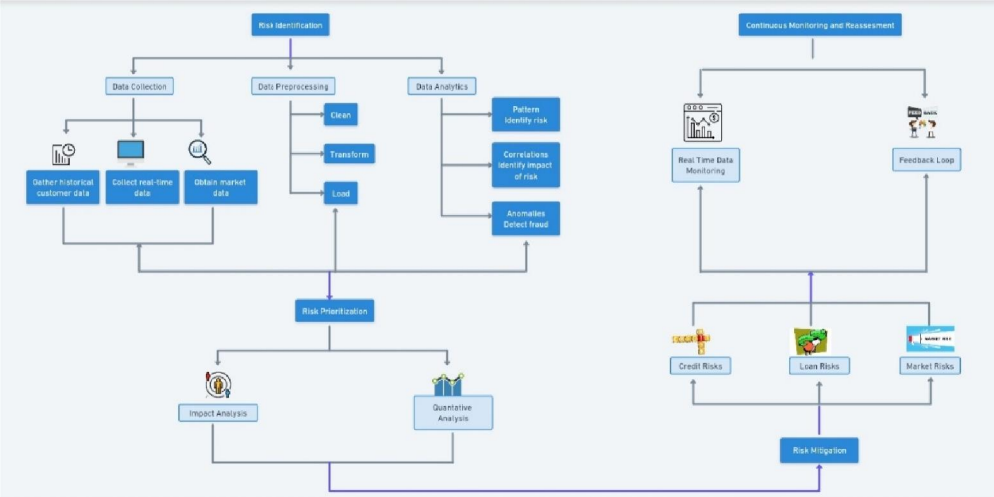


Figure 2 : Process flow of Implementation

5.1.3. Feature Engineering
Objective: Enhance the dataset's predictive capabilities.
Procedure:

- Generate new features from existing data.
- Use correlation matrices and feature importance techniques to select crucial features..

5.1.4. Train-Test Split
Objective: Separate data for model training and evaluation.
Procedure: Use libraries like **train_test_split** from scikit-learn to allocate 80% of the data for training and 20% for testing.

5.1.5. Model Development
Objective: Construct an ANN for fraud detection.
Procedure:

- Define the model architecture using TensorFlow's Keras library.
- Configure layers, neurons, and activation functions.
- Compile the model with an optimizer, loss function, and desired metrics.

5.1.6. Model Evaluation
Objective: Construct an ANN for fraud detection.
Procedure:

- Define the model architecture using TensorFlow's Keras library.
- Configure layers, neurons, and activation functions.
- Compile the model with an optimizer, loss function, and desired metrics.

5.1.7. Risk Prioritization
Objective: Rank loan applications based on fraud likelihood.
Procedure:

- Implement a scoring system that assigns risk scores to each application.
- Categorize applications into low, medium, and high-risk buckets.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, November 2023

5.1.8.KYC Verification

Objective: Add an additional layer of validation for the model's predictions.

Procedure:

- Integrate with the existing KYC system.
- For applications flagged by the model, run them through the KYC verification process to validate or invalidate the model's prediction.

5.1.9 Prediction

Objective: Enable the model to make real-time predictions on new loan applications.

Procedure:

- Integrate the model into the loan application pipeline.
- As applications arrive, feed them into the model for instant fraud assessment.

5.1.10 Model Deployment

Objective: Launch the model in a real-world environment.

Procedure:

- Set up the necessary infrastructure, ensuring scalability and security.
- Monitor the model's performance over time and make necessary adjustments based on feedback and evolving fraud patterns.

5.2. Equations

5.2.1.Data Preprocessing:

Standardization: Given a feature X, the standardized value Z is calculated as:

$$Z = \frac{X - \mu}{\sigma}$$

where μ is the mean of the feature and σ is its standard deviation.

5.2.2. Feature Engineering:

Correlation: Pearson correlation coefficient r between two features X and Y is given by:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

where ¯X¯ and ¯Y¯ are the means of features X and Y respectively.

5.2.3. Model Development:

ReLU Activation Function: Given an input x, the output y using ReLU is:

y=max(0,x)

Sigmoid Activation Function: Given an input x, the output y using sigmoid is:

$$y = \frac{1}{1 + e^{-x}}$$

Loss Function (Mean Absolute Error): Given true values Y and predicted values ^Y, the MAE is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

5.2.4. Model Evaluation

Accuracy: Given true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), accuracy is:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, November 2023

Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

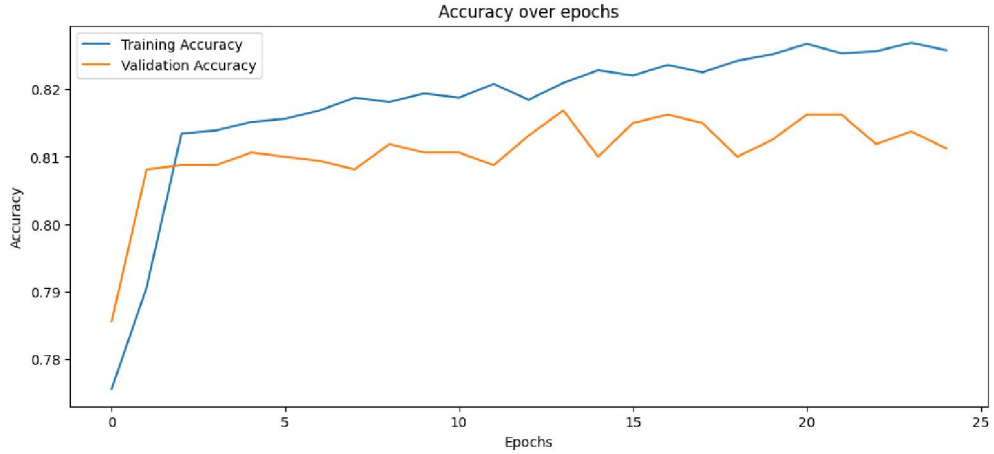
Recall:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-score: Harmonic mean of precision and recall:

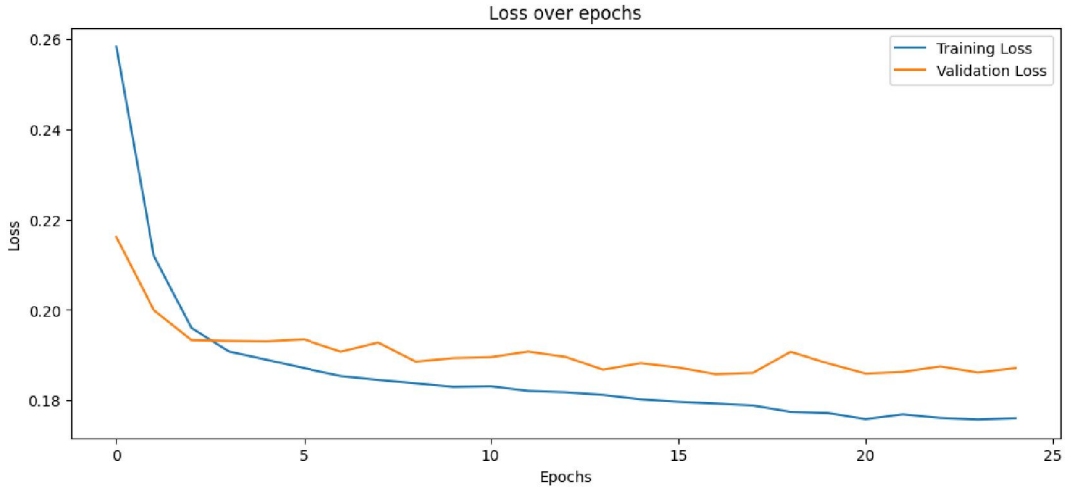
$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

VI. VISUALIZATION



Graph 1: Accuracy over epochs

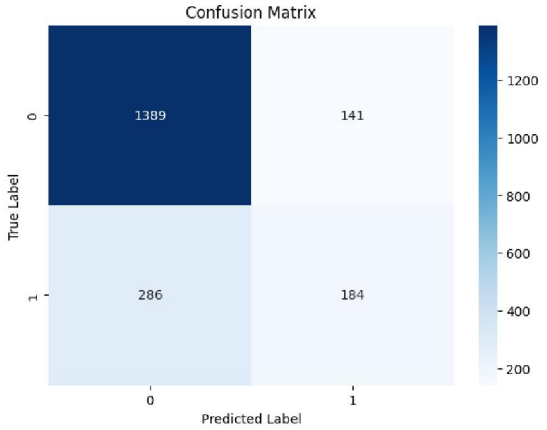
The image presents a line graph titled "Accuracy over epochs". It plots two lines, representing 'Training Accuracy' and 'Validation Accuracy', across a range of epochs from 0 to 25. The Training Accuracy line, colored in blue, starts near 0.80 and gradually rises, fluctuating slightly, to just above 0.82 by the 25th epoch. The Validation Accuracy line, in orange, begins near 0.78 and exhibits more fluctuation than the Training line. It peaks around the 5th epoch, and from there, it exhibits a series of peaks and troughs, ending below the Training Accuracy line by the 25th epoch. Overall, while the Training Accuracy shows a general increase, the Validation Accuracy seems more volatile.



Graph 2 Loss over epochs

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, November 2023

The graph is titled "Loss over epochs" and displays two lines, representing 'Training Loss' and 'Validation Loss', over a span of 25 epochs. The blue line, symbolizing Training Loss, starts sharply around 0.26 and quickly descends to stabilize close to 0.18 by the 25th epoch. Conversely, the orange line, depicting Validation Loss, begins near 0.22 and shows mild fluctuations throughout its course, meandering around the 0.20 mark. By the end of the 25 epochs, both lines seem to be converging, with the Training Loss slightly below the Validation Loss. The overall trend indicates a reduction in loss for both training and validation, with the Training Loss having a steeper decline initially.



Graph 3: Confusion Matrix

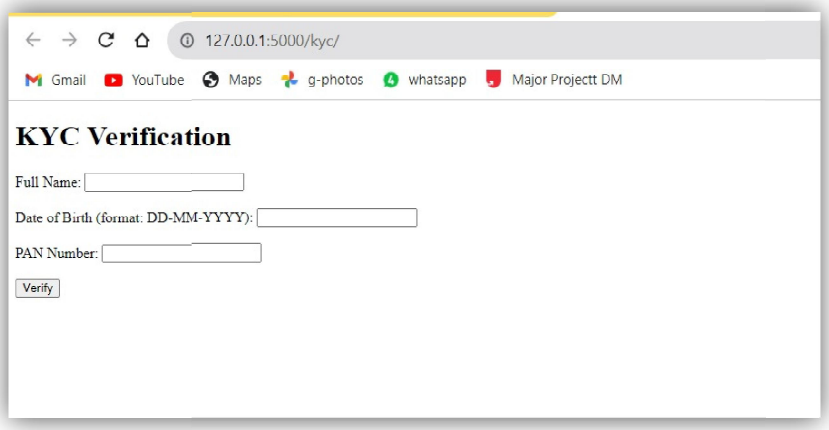
The graph is a "Confusion Matrix" comparing True Labels against Predicted Labels for two classes: 0 and 1. For the True Label 0, 1389 instances were correctly predicted as 0 (True Negatives) while 141 were incorrectly predicted as 1 (False Positives). For the True Label 1, 286 instances were incorrectly predicted as 0 (False Negatives) and 184 instances were correctly predicted as 1 (True Positives). The color intensity corresponds to the number of instances, with darker shades representing higher counts.

VII. RESULTS

In the pursuit of a robust bank loan fraud detection solution, this research successfully integrated the traditional Know Your Customer (KYC) verification system with advanced Artificial Neural Networks (ANN).

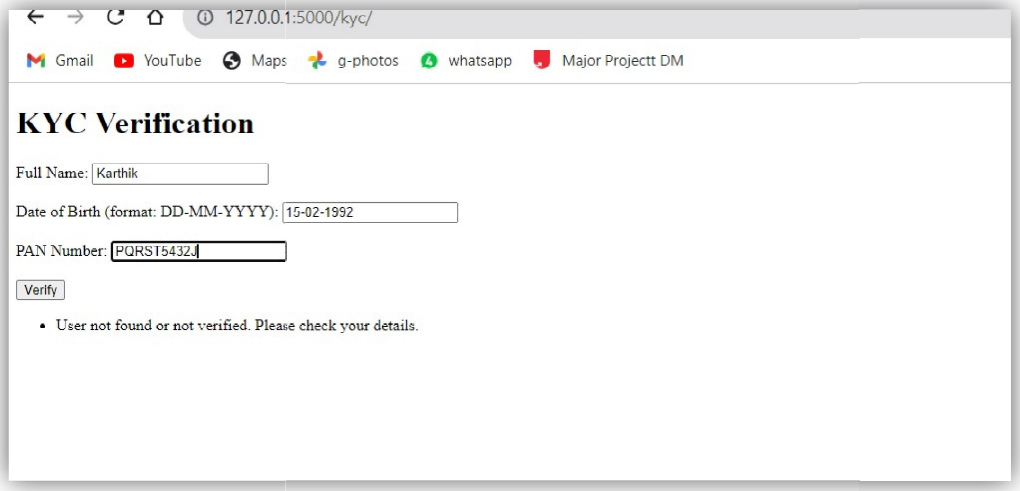
The results, visualized through a series of graphs, highlighted the model's consistent learning trajectory, with Training Accuracy demonstrating a steady ascent, reaching just above 0.82 by the 25th epoch. However, the Validation Accuracy, while initially promising, displayed fluctuations, suggesting potential areas for model refinement.

The "Loss over epochs" graph reinforced this observation, as both Training and Validation Loss showed a converging trend, indicative of the model's evolving efficiency. A deeper dive into the model's classification capabilities via the "Confusion Matrix" revealed its proficiency in identifying genuine applications, though it presented some challenges in accurately flagging fraudulent ones.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, November 2023

The synergy of the ANN model with the KYC system added a pivotal layer of validation, enhancing the overall defense against fraudulent loan activities. In summary, while the integrated system showcased promising results in mitigating bank loan fraud, there remain avenues for optimization to further elevate its efficacy.



VIII. CONCLUSION

The rapidly evolving landscape of financial transactions, underpinned by technological advancements, has brought forth both opportunities and challenges. Among the most pressing challenges is the escalating threat of bank loan fraud, which undermines the trustworthiness of financial institutions and poses significant economic repercussions. This research embarked on a journey to address this challenge, culminating in the design of an integrated bank loan fraud detection system that harmoniously melds the traditional Know Your Customer (KYC) verification process with the computational prowess of Artificial Neural Networks (ANN).

The results derived from this integration are both promising and enlightening. They underscore the potential of harnessing advanced machine learning techniques to fortify and enhance conventional verification methods. The model's performance, particularly in identifying genuine loan applications, stands as a testament to the effectiveness of this integrated approach. However, the research also illuminated areas of potential refinement, especially in the model's ability to accurately detect fraudulent applications.

Furthermore, the seamless integration with the KYC system not only augmented the reliability of the predictions but also emphasized the significance of blending time-tested methodologies with cutting-edge technological innovations. Such a fusion not only bolsters the defence against fraudulent activities but also offers a streamlined user experience, ensuring that genuine customers face minimal friction during their loan application process.

As we stand on the cusp of a digital financial era, it's imperative that our strategies to combat fraud evolve in tandem with the tactics of fraudsters. This research provides a beacon, highlighting the path forward. It accentuates the importance of continuous innovation, adaptation, and integration in ensuring a secure, efficient, and trustworthy banking environment. As we look ahead, it is our hope that this integrated approach serves as a foundation upon which further advancements in fraud detection can be built, safeguarding the future of financial transactions.

IX. ACKNOLEDGEMENT

We express our profound gratitude to Dr. Sonali Ridhorkar, H.O.D of the Computer Science and Engineering department, for her invaluable insights, guidance, and constant encouragement throughout the course of this research. Her expertise and dedication have been instrumental in shaping the direction and outcomes of this study. Furthermore, we extend our heartfelt appreciation to Dr. Vivek Kapur, Director of G.H Raisoni Institute of Engineering and Technology, for providing an environment conducive to research and for his unwavering support in all our academic endeavors. Their combined leadership and vision have greatly facilitated our research journey. Special thanks to Prof. Antara Bhattacharya for her mentorship and guidance throughout this research endeavor.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, November 2023

REFERENCES

[1]. Sanjay Misra, "Artificial Intelligence based System for Bank Loan Fraud Prediction", Research Gate https: International Conference on Hybrid Intelligent Systems. Published: 04 March 2022.

[2]. Hanlin Wen, Fangming Huang, "Personal Loan Fraud Detection Based on Hybrid Supervised and Unsupervised Learning", IEEE:2020: 5th IEEE International Conference on Big Data Analytics.

[3]. Angel Parmar, Kelvin Parmar, Premkumar Balani, "Loan Fraud Detection – Using ML", International Journal for Research in Engineering Application & Management (IJREAM) ISSN: 2454-9150 Vol-06, Issue-07, OCT 2020.

[4]. Dhiman Sharma, Wahidul Alam, Ishita saha, Nazmul Alam, Jahangir Alam, Sohrab Hossain, "Bank Fraud Detection using Community Detection Algorithm", Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020) IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978- 1-7281-5374-2t.

[5]. I O Eweoya, A A Adebiyi, A A Azeta, Angela E Azeta, "Fraud Prediction in bank loan administration using decision tree", 3rd International Conference on Science and Sustainable Development (ICSSD 2019) IOP Conf. Series: Journal of Physics: Conf. Series 1299 (2019) 012037 IOP Publishing Doi:10.1088/1742-6596/1299/1/012037.

[6]. R. B. Xu, Y. Che, X. M. Wang, J. X. Hu, and Y. Xie, "Stacked autoencoder-based community detection method via an ensemble clustering framework," (in English), Information Sciences, Article vol. 526, pp. 151-165, Jul 2020.

[7]. G. S. Carnivali, A. B. Vieira, A. Ziviani, and P. A. A. Esquef, "CoVeC: Coarse-grained vertex clustering for efficient community detection in sparse complex networks," (in English), Information Sciences, Article vol. 522, pp. 180-192, Jun 2020.

[8]. S. Hossain, et al., "A Belief Rule Based Expert System to Predict Student Performance under Uncertainty," in 2019 22nd International Conference on Computer and Information Technology (ICCIT), 2019, pp. 1-6.

[9]. E. A. Minastireanu and G. Mesnita, "Methods of Handling Unbalanced Datasets in Credit Card Fraud Detection," (in English), Brain-Broad Research in Artificial Intelligence and Neuroscience, Article vol. 11, no. 1, pp. 131-143, Mar 2020.

[10]. Huang, B., Huan, Y., Xu, L. D., Zheng, L., & Zou, Z. (2019). Automated trading systems statistical and machine learning methods and hardware implementation: a survey. Enterprise Information Systems, 13(1), 132-144.

[11]. Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three-stage hybrid Adaptive Neuro-Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. European Journal of Operational Research, 222(1), 168-178.

[12]. S. N. John, C. Anele, O. O. Kennedy, F. Olajide, and C. G. Kennedy, "Realtime Fraud Detection in the Banking Sector Using Data Mining Techniques/Algorithm," in Proceedings of International Conference on Computational Science and Computational Intelligence (CSCI). IEEE, 2016, pp. 1186–1191.