International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-06 November- December 2026 ISSN NO: 3108-1312

AI-Based Phishing and UPI Fraud Detection Using Machine Learning, Deep Learning and Explainable AI

- Pooja S Suryawanshi
  FY MCA
  Sandip Foundation's Sandip Institute of Engineering & Management
  poojasuryawnshi2411@gmail.com

- Priyanka P Gupta
  FY MCA
  Sandip Foundation's Sandip Institute of Engineering & Management
  priyankagupta07286@gmail.com

AUTHOR GUIDE:

- Prof. Ganesh Ragade
  Sandip Foundation's Sandip Institute of Engineering & Management Nashik
  ganesh.ragade@siem.org.in

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-06 November- December 2026 ISSN NO: 3108-1312

Abstract

Online payments have become part of daily life in India, and UPI is now the first choice for most people when they pay a shopkeeper, split a bill or send money to a friend. The same convenience has given fraudsters a large and easy target. Fake bank links, cloned payment pages, fraudulent SMS messages, false collect requests and social engineering calls are used every day to trick ordinary users into sharing their PIN or approving a payment they never intended to make. Rule-based filters and manual complaint handling cannot keep up with the speed and volume of these attacks.

This paper proposes an AI-based framework that handles two closely related problems together: detecting phishing websites and messages, and detecting fraudulent UPI transactions. For phishing, the system looks at URL structure, page characteristics and message text. For UPI fraud, it studies transaction amount, timing, frequency, device and payee behaviour. The data is cleaned, balanced and converted into numerical features. Logistic Regression, Random Forest and XGBoost are compared with deep learning models, namely an LSTM network for transaction sequences and a BERT model for message text. SHAP and LIME are added so that every alert can be explained in simple terms to the user and to the bank analyst.

The models are evaluated with Accuracy, Precision, Recall, F1-Score and ROC-AUC. Because fraud cases are rare in real data, special attention is given to Recall and F1-Score instead of Accuracy alone. The results show that ensemble and deep learning models detect fraud noticeably better than simple baseline models, and that the explanation module makes the predictions easier to trust.

Keywords: Phishing Detection, UPI Fraud Detection, Machine Learning, Deep Learning, XGBoost, LSTM, BERT, Explainable Artificial Intelligence, Digital Payment Security.

1. Introduction

Digital payments have changed the way people in India handle money. The Unified Payments Interface, introduced by the National Payments Corporation of India, allows anyone with a bank account and a smartphone to send money in a few seconds. Today UPI handles billions of transactions every month, from small tea-stall payments to large business transfers. This growth is a success story, but it has also created a very attractive space for criminals.

Phishing is one of the most common ways in which people lose money. In a phishing attack, the user receives a message, e-mail or link that looks like it came from a bank, a delivery company, an electricity board or a government office. The link opens a fake page that asks for card details, net banking passwords, OTPs or the UPI PIN. Many victims realise what happened only after the money has left their account. UPI fraud also takes other forms, such as fake collect requests, QR code scams, remote access apps installed on the phone, and calls from people pretending to be bank staff.

Banks and payment apps do use rules to block suspicious activity, for example limits on amount or number of transactions. But fraudsters keep changing their methods, and fixed rules either

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-06 November- December 2026 ISSN NO: 3108-1312

miss new tricks or block genuine customers. Manual checking is too slow when millions of payments are happening every hour. Because of this, researchers have started using Machine Learning and Deep Learning, which can learn the patterns of normal and fraudulent behaviour from past data and apply them to new cases within milliseconds.

This research looks at phishing detection and UPI fraud detection as two parts of one problem, since a phishing link is very often the first step and a fraudulent UPI transaction is the last. We compare traditional machine learning algorithms with deep learning models and add Explainable AI methods, so that the system not only says that something is fraud but also shows why.

2. Literature Review

A lot of work has already been done on phishing and financial fraud detection. The main studies that guided this research are described below.

Sahingoz et al. built a phishing detection system that works only on URL features. They tested several classifiers and found that Random Forest gave the best results, which shows that the URL alone carries a lot of useful information.

Prasad and Chandra introduced the PhiUSIIL framework and released a large phishing URL dataset that combines URL-based, page-based and similarity-based features. Their work showed that mixing different kinds of features gives a more reliable detector than using only one kind.

Abdallah et al. reviewed fraud detection systems used in banking and payments. They pointed out that class imbalance, delayed labels and changing fraud patterns are the biggest practical difficulties, and that most systems struggle when fraudsters change their behaviour.

Dal Pozzolo et al. studied credit card fraud data and showed that under-sampling and probability calibration have a strong effect on how a model behaves when fraud is very rare. Chawla et al. proposed SMOTE, which creates synthetic samples of the minority class and is now widely used to balance fraud datasets.

Chen and Guestrin presented XGBoost, a gradient boosting method that has become a standard choice for tabular data such as transaction records because it is fast, accurate and handles missing values well. For sequence data, Hochreiter and Schmidhuber's LSTM network is commonly used to learn from a customer's transaction history over time.

Devlin et al. introduced BERT, which reads a sentence in both directions and understands context. It has been applied to SMS and e-mail phishing detection with good results. Lundberg and Lee proposed SHAP and Ribeiro et al. proposed LIME, two popular methods for explaining the output of complex models.

Most of these studies handle either phishing or transaction fraud, and many of them are built around credit card or international datasets. Very little work looks at the Indian UPI setting, where payments are instant, irreversible and often triggered by a message the user has just

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-06 November- December 2026 ISSN NO: 3108-1312

received.

3. Research Gap

The literature shows good progress in detecting phishing and payment fraud separately. Some problems are still open.

1. Most studies treat phishing detection and transaction fraud detection as two separate tasks, even though in real life one leads to the other.
2. Many fraud models are trained on credit card data and are not tested on UPI-style payments, where transfers are instant and the payee is often a person and not a merchant.
3. Real fraud data is extremely imbalanced, and many papers report high Accuracy that hides poor detection of the actual fraud cases.
4. Deep learning models act like black boxes. Banks and users are not told why a payment was blocked, which reduces trust and makes it hard to handle complaints.
5. Fraud methods keep changing, so a model that is trained once becomes outdated soon and needs a design that can be updated easily.

This gap motivates a combined and explainable framework for phishing and UPI fraud detection.

4. Problem Statement

The number of UPI users and transactions is growing every year, and so is the number of phishing attacks and payment frauds linked with it. Manual verification and fixed rules cannot handle this volume or adapt to new tricks, and many users lose money before they even realise they have been cheated.

There is a need for an intelligent system that can recognise phishing links and messages, spot suspicious UPI transactions in real time, work well even when fraud cases are very few compared to genuine ones, and give a clear reason for every alert it raises.

5. Objectives

The main objectives of this research are:

1. To build an AI-based system that detects phishing URLs and phishing messages.
2. To build a model that identifies fraudulent UPI transactions from transaction behaviour.
3. To clean and prepare the data and handle the class imbalance problem using techniques such as SMOTE.
4. To implement and compare Logistic Regression, Random Forest, XGBoost, LSTM and BERT models.
5. To evaluate the models using Accuracy, Precision, Recall, F1-Score and ROC-AUC.

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-06 November- December 2026 ISSN NO: 3108-1312

6. To add Explainable AI methods (SHAP and LIME) so that each decision can be understood.
7. To find the most effective approach for detecting phishing and UPI fraud.

6. Proposed Methodology

The proposed method has several stages. Raw URLs, messages and transaction records go in at one end, and a fraud or genuine decision with an explanation comes out at the other.

6.1 Data Collection

Two types of data are used. For phishing detection, a public phishing URL dataset (the PhiUSIIL dataset from the UCI repository) is used along with SMS and e-mail phishing text collections available on Kaggle. For transaction fraud, a publicly available simulated mobile money dataset (PaySim) and a synthetic UPI-style transaction dataset are used, because real UPI data from banks is confidential and cannot be shared for research.

The collected data contains:

- URLs and their legitimate or phishing labels
- Message text with spam or phishing labels
- Transaction amount, type, time and account balances
- Sender and receiver identifiers (anonymised)
- Fraud labels (Genuine or Fraud)

6.2 Data Preprocessing

Raw data is rarely ready for a model, so the following steps are carried out.

Handling Missing and Duplicate Values

Duplicate records are removed. Missing numeric values are filled with the median of the column, and records with too many missing fields are dropped.

Text Cleaning and Tokenization

Message text is converted to lowercase, extra symbols are removed, and the text is split into individual words or tokens. Links, phone numbers and amounts inside a message are kept as separate signals because they are very useful for detecting phishing.

Example:

Original: "Your SBI account is blocked. Click here to update KYC now"

Processed: "your sbi account is blocked click here to update kyc now"

Stop Word Removal, Stemming and Lemmatization

Very common words such as “is”, “the” and “to” are removed. Words are then reduced to their root form (for example, blocked → block) so that different forms of the same word are treated alike.

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-06 November- December 2026 ISSN NO: 3108-1312

Scaling

Numeric features such as transaction amount are scaled so that large values do not dominate smaller ones. This is especially important for Logistic Regression and the LSTM model.

Handling Class Imbalance

In a transaction dataset only a very small share of records is fraud. If the model is trained directly on such data, it learns to say "genuine" every time and still looks accurate. To avoid this, SMOTE is applied only on the training set. The test set is left untouched so that the evaluation reflects real conditions.

6.3 Feature Extraction

The models cannot read raw URLs or messages, so useful features are extracted from them.

URL Features

- Length of the URL and the domain name
- Number of dots, hyphens and special characters such as @
- Use of an IP address instead of a domain name
- Presence of HTTPS and the age of the domain
- Presence of words such as login, verify, update or bank
- Use of URL shortening services

Transaction Features

- Transaction amount compared to the customer's usual amount
- Time of the day and day of the week
- Number of transactions in the last hour and the last day
- Whether the receiver is new for the sender
- Sudden drop in account balance after a transaction
- Change of device or location

TF-IDF for Message Text

Term Frequency–Inverse Document Frequency shows how important a word is in a message compared to the whole collection of messages. Words like "urgent", "blocked" and "KYC" get high weight in phishing messages. It is simple, fast and works well with classical models.

6.4 Model Development

The system compares simple baseline models with stronger machine learning and deep learning models.

Logistic Regression

A basic classification algorithm that gives a probability for each class. It trains quickly, is easy to understand and works as a baseline for comparison.

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-06 November- December 2026 ISSN NO: 3108-1312

Random Forest

Random Forest builds many decision trees and takes a vote among them. It reduces overfitting and works well on URL and transaction features without much tuning.

XGBoost

XGBoost builds trees one after another, with each tree correcting the mistakes of the previous ones. It is very effective on tabular data and is widely used in fraud detection.

LSTM Network

A Long Short-Term Memory network can learn from the order of a customer's recent transactions. A sudden change from the usual pattern, such as several quick payments to new receivers, is picked up much more easily by this model.

BERT Model

BERT reads a message in both directions and understands the context of the words. It is used to classify SMS and e-mail text as phishing or genuine, even when the wording is different from the examples seen during training.

7. System Architecture

The proposed system works through the following stages:

URL / Message / Transaction Input
↓
Data Cleaning and Preprocessing
↓
Feature Extraction
↓
Class Balancing (SMOTE, on training data)
↓
Machine Learning / Deep Learning Models
↓
Classification
↓
Genuine or Fraud / Phishing
↓
Explainability Module (SHAP / LIME)
↓
Final Alert with Reason

8. Explainable AI Integration

A fraud system that only says "blocked" is frustrating for the customer and difficult for the bank to defend. Explainable AI methods are therefore added to show which factors pushed the

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-06 November- December 2026 ISSN NO: 3108-1312

model towards its decision.

SHAP (SHapley Additive exPlanations)

SHAP gives every feature a score that shows how much it increased or decreased the chance of fraud. For example, it can show that a high amount, a new receiver and a late-night time together caused a transaction to be flagged.

LIME (Local Interpretable Model-Agnostic Explanations)

LIME explains a single prediction by building a simple model around that one case. For a phishing message it can highlight the exact words, such as “urgent” or “verify”, that made the message look suspicious.

With SHAP and LIME, the analyst can check the reasoning, the customer can be told why a payment was stopped, and the bank can improve the model when the reasoning looks wrong.

9. Dataset Description

The quality of the data decides how good the final model will be. This research uses public datasets so that the work can be repeated by others.

9.1 Phishing Dataset

The PhiUSIIL phishing URL dataset contains around 235,000 URLs, labelled as legitimate or phishing, with a mix of URL-based and page-based features. This dataset is fairly balanced between the two classes. SMS and e-mail phishing text collections from Kaggle are used for the message classification part.

9.2 Transaction Dataset

The PaySim dataset is a simulated mobile money dataset created from real transaction logs. It contains about 6.3 million transactions, out of which only a very small fraction (around 0.1%) are fraud. A synthetic UPI-style dataset is also generated to include fields such as payer, payee, device and time of transaction.

Each transaction record contains:

- Transaction type and amount
- Sender and receiver balance before and after the transaction
- Time step of the transaction
- Fraud label

9.3 Data Distribution

Each dataset is divided as follows:

Training Set: 80%

Testing Set: 20%

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-06 November- December 2026 ISSN NO: 3108-1312

A stratified split is used so that the fraud ratio stays the same in both parts. SMOTE is applied only to the training part to avoid leaking information into the test data.

10. Mathematical Model

The detection framework can be written in simple mathematical form.

Let D = {d_1 , d_2 , d_3 , ..., d□} be the set of input records (URLs, messages or transactions).

After preprocessing: P(D) = Preprocessed Dataset

Feature extraction converts each record into a numerical vector: F = φ(P(D))

where φ is the feature extraction function (URL features, TF-IDF or transaction features).

Classification: C(F) → {Genuine, Fraud}

For Logistic Regression, the probability of fraud is calculated as:

P(y = 1 | x) = 1 / (1 + e^−(w·x + b))

The final prediction is: Y = f(X)

where X is the input record, f is the trained model and Y is the predicted label.

The explanation for a prediction is given by SHAP values φ_i such that $f(x) = \varphi_0 + \Sigma \varphi_i$, where φ_i is the contribution of the i-th feature.

11. Evaluation Metrics

The following measures are used to judge the models.

Accuracy

The share of all predictions that are correct.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

Precision

Out of all cases marked as fraud, how many were really fraud. Low precision means genuine customers get blocked.

Precision = TP / (TP + FP)

Recall

Out of all real fraud cases, how many were caught. Low recall means fraud goes undetected.

Recall = TP / (TP + FN)

F1-Score

Balances Precision and Recall.

F1 = 2 × (Precision × Recall) / (Precision + Recall)

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-06 November- December 2026 ISSN NO: 3108-1312

ROC-AUC

Shows how well the model separates fraud from genuine cases across all possible thresholds. A value close to 1 is better. For heavily imbalanced fraud data, Recall, F1-Score and ROC-AUC are more meaningful than Accuracy.

12. Experimental Results

All models were trained and tested on the same split of each dataset so that the comparison is fair.

12.1 Phishing Detection

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	93.1	92.4	91.8	92.1
Random Forest	97.2	96.8	96.3	96.5
XGBoost	98.0	97.6	97.2	97.4
BERT (message text)	98.6	98.2	98.0	98.1

12.2 UPI Fraud Detection

Model	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
Logistic Regression	78.5	71.2	74.7	0.93
Random Forest	91.3	84.6	87.8	0.97
XGBoost	93.4	88.9	91.1	0.98
LSTM	92.1	90.2	91.1	0.98

Result Analysis

The main observations from the experiments are:

- Logistic Regression is quick and easy but misses a good share of fraud cases, because fraud patterns are not linear.
- Random Forest and XGBoost give a large improvement, with XGBoost performing best among the classical models.
- The LSTM model catches more fraud than the other models because it looks at the sequence of transactions and not at a single payment alone.
- BERT gives the best results on message text because it understands the wording and context of the message.
- Recall improved clearly after balancing the training data, without a large fall in Precision.

These findings suggest that learning the behaviour of the user over time, and understanding the

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-06 November- December 2026 ISSN NO: 3108-1312

context of the text, are both important for catching modern phishing and UPI fraud.

13. Discussion

The results show that ensemble and deep learning models are clearly better than simple baseline models for both phishing and UPI fraud. Fraudsters change their messages, links and payment patterns, and models that only follow fixed rules or straight-line relationships cannot keep up. Tree-based models and sequence models are better at picking up these changes.

The explanation module added real value. In the SHAP analysis, features such as transaction amount, a new receiver, an unusual time and a sudden drop in balance had the largest effect on fraud predictions. For phishing messages, LIME highlighted words that create urgency or fear. This kind of feedback helps analysts to trust the system and can also be turned into a short, simple warning for the user.

There are costs too. Deep learning models need more data, more computing power and longer training time. A model that blocks too many genuine payments will annoy customers, so the decision threshold has to be chosen carefully.

14. Advantages of the Proposed System

- Detects both phishing and UPI fraud in one framework.
- High detection rate even when fraud cases are very few.
- Works automatically and fast enough for real-time checks.
- Explains every alert in simple terms.
- Can be retrained when new fraud patterns appear.
- Can be used by banks, payment apps and individual users.

15. Limitations

- Real UPI transaction data is confidential, so public and simulated data had to be used.
- Simulated data may not fully show how real fraudsters behave.
- Deep learning models need large datasets and high computing power.
- Regular retraining is needed as fraud methods keep changing.
- Messages in regional languages and mixed languages (for example Marathi written in English letters) are harder to analyse.

16. Future Scope

The following improvements can be explored in future work:

- Testing the system on real, anonymised bank data in cooperation with financial institutions.
- Support for Marathi, Hindi and other regional languages.

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-06 November- December 2026 ISSN NO: 3108-1312

- Detection of QR code scams and fake payment apps.
- Real-time warnings inside the payment app before the user confirms a payment.
- Detection of voice call scams and AI-generated fake messages.
- Federated learning, so that banks can train a shared model without sharing customer data.
- Graph-based methods to find groups of connected fraudulent accounts.

17. Conclusion

Phishing and UPI fraud are growing along with the use of digital payments, and they cause real financial and emotional harm to ordinary users. Rules and manual checking are not enough to deal with them.

This research presented an AI-based framework that detects phishing URLs and messages as well as fraudulent UPI transactions. It uses Machine Learning and Deep Learning models, handles the problem of imbalanced data, and adds SHAP and LIME so that decisions can be explained. The comparison shows that XGBoost, LSTM and BERT perform better than simple baseline models, and that looking at transaction behaviour and message context together gives a stronger defence.

With further testing on real data and support for more languages, a system of this kind can help banks and payment apps to protect their customers and increase trust in digital payments.

References

[1] Sahingoz, O. K., Buber, E., Demir, O., and Diri, B., "Machine Learning Based Phishing Detection from URLs," Expert Systems with Applications, 2019.

[2] Prasad, A., and Chandra, S., "PhiUSIIL: A Diverse Security Profile Empowered Phishing URL Detection Framework Based on Similarity Index and Incremental Learning," Computers & Security, 2024.

[3] Abdallah, A., Maarof, M. A., and Zainal, A., "Fraud Detection System: A Survey," Journal of Network and Computer Applications, 2016.

[4] Lopez-Rojas, E., Elmir, A., and Axelsson, S., "PaySim: A Financial Mobile Money Simulator for Fraud Detection," European Modeling and Simulation Symposium, 2016.

[5] Dal Pozzolo, A., Caelen, O., Johnson, R. A., and Bontempi, G., "Calibrating Probability with Undersampling for Unbalanced Classification," IEEE Symposium Series on Computational Intelligence, 2015.

[6] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, 2002.

[7] Chen, T., and Guestrin, C., "XGBoost: A Scalable Tree Boosting System," ACM SIGKDD, 2016.

[8] Hochreiter, S., and Schmidhuber, J., "Long Short-Term Memory," Neural Computation, 1997.

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-06 November- December 2026 ISSN NO: 3108-1312

[9] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, 2019.

[10] Lundberg, S. M., and Lee, S. I., "A Unified Approach to Interpreting Model Predictions," NIPS, 2017.

[11] Ribeiro, M. T., Singh, S., and Guestrin, C., "Why Should I Trust You? Explaining the Predictions of Any Classifier," KDD, 2016.

[12] Breiman, L., "Random Forests," Machine Learning, 2001.

[13] National Payments Corporation of India, "UPI Product Statistics," npci.org.in.