American Journal of Software Engineering, 2026, Vol. 9, No. 1, 1-9
Available online at http://pubs.sciepub.com/ajse/9/1/1
Published by Science and Education Publishing
DOI:10.12691/ajse-9-1-1

Identity Theft Detection at Data Ingestion Using AI: An Explainable Anomaly Detection Approach

Sachin Dattatreya Murthy[1,2,*]

[1]Independent Researcher
[2]United States
*Corresponding author: Sachin.damurthy@gmail.com

Received November 25, 2025; Revised December 27, 2025; Accepted January 03, 2026

Abstract The rise of identity theft has become one of the most dangerous growing cybercrimes today, particularly as individuals are now digitally on-boarding; therefore, with minimal information provided for identification/verification purposes, traditional rule-based systems cannot identify many of the sophisticated schemes used today such as Deepfakes, Document Forging, Synthetic Identities etc. Fraud detection has been the focus of much research but there is still a large void in the area of data ingestions, specifically in identifying and alerting Identity Theft prior to an account being created through a Real Time Explainable Solution. Fraud detection is a well-researched topic; however, fraud detection at the time of account creation (during the ingestion of data) remains a largely unexplored area where fraud detection is most important. In addition, current fraud detection systems do not have the capability to use hybrid models that can detect multi-modal, synthetic identities, and deepfakes as well as other cross-channel anomalies. Additionally, most current fraud detection systems do not provide an integrated approach of using both supervised and unsupervised methods for detection or include the ability to provide explanations for the decision-making process of the model to combat modern forms of synthetic and AI-based attacks. We present a **Hybrid AI Framework** which utilizes **Supervised Learning, Unsupervised Anomaly Detection, and Explanatory AI (XAI)**, to identify Identity Fraud prior to Account Creation. This Framework will combine multiple Data Sources (Documents, Biometric Information, Devices, Structured Attributes) to produce Interpretable Risk Scores, utilizing SHAP Values & Rule Based Explanation, allowing Analysts to Identify Alerts & Resolve Them Efficiently. Our End-To-End Design Offers a Scalable, Compliant Solution to Early-Stage Identity Theft Prevention in Financial Services.

Keywords: *Identity Theft, Anomaly detection, Deep Fakes, Explainable AI, Feature engineering, Data pre-processing, finance, Machine Learning (ML), Hybrid AI*

Cite This Article: Sachin Dattatreya Murthy, "Identity Theft Detection at Data Ingestion Using AI: An Explainable Anomaly Detection Approach." *American Journal of Software Engineering*, vol. 9, no. 1 (2026): 1-9. doi: 10.12691/ajse-9-1-1.

1. Introduction

A major type of digital crime today is Identity Fraud—the unauthorized collection and misuse of an individual's **Personal Identifying Information**. As a result of the theft of identities (real or created), individuals have been able to fraudulently open accounts, obtain credit and conduct other types of fraudulent activity that cause them significant financial loss [1]. Identity fraud resulted in **18 million victims**, and caused US consumers **$27 billion in losses in 2024**. However, the total loss due to all types of scams was greater than **$47 billion [1]**. The FTC reported that there were over **1.1 million identity theft complaints in 2024**; this shows how large the problem of identity fraud has grown [2].

While the majority of fraud is identified after the fact through transaction fraud, identity fraud is normally discovered before it reaches that stage of fraud as it is committed at the time of onboarding before the fraudster can establish behavior that can be used to identify the fraudulent account [3]. Fraudsters use personal identifying information obtained from data breaches or phishing attacks to successfully complete the verification process. Sophisticated fraud schemes that include creating synthetic identities, and using deepfakes to create false identification documents, continue to thwart traditional KYC processes [4]. Therefore, manual reviews, and rule-based systems are unable to provide adequate protection.

Artificial Intelligence (AI) provides a better way for organizations to protect themselves from identity fraud by being able to identify subtle and complex patterns in identity related data. Supervised and anomaly detection models are able to recognize and report suspicious patterns of identity data that may indicate fraud, even in its earliest stages [5,6]. However, many AI systems operate as **"black boxes"** causing problems with transparency and accountability, especially when making decisions regarding high stakes issues. As a result of these

concerns, regulatory bodies now require AI systems to be explainable; however, only 22% of organizations had implemented such systems by 2023 [7].

The goal of this study is to create an artificial intelligence (AI) based anomaly detection system for the purpose of detecting identity fraud at the earliest stages possible during a customer's digital onboarding process in the banking/financial sector. Additionally, our system will provide **human-readable explanations** for the results of the analysis. Our objective has led us to conceptualize a basic algorithm and overall system architecture, where we ingest user submitted identity information, detect and identify high-risk patterns (i.e., synthetic identities) using this information, and explain the logic behind our detection in terms that are understandable by humans. The **Explainable AI** portion of our system enables a connection between the predictive capability of AI technology and the degree of transparency that is necessary for regulatory compliance regarding AML and Risk Management in financial organizations [7,8].

This study makes the following contributions:

- We have proposed a new AI-based algorithm for the detection of identity fraud via supervised classification and unsupervised anomaly detection in the initial phases of a digital onboarding process.
- The system incorporates a component of Explainable AI which generates instance specific explanations for the detection of identity fraud for the purposes of enhancing regulatory compliance and minimizing false positive alerts that can complicate the interpretation of analysts [8,9].
- We described the end-to-end architecture of our system, from data ingestion and feature extraction to model training and real time inference.

This paper will be structured as follows: Section 2 addresses prior research. Section 3 describes the techniques and detection. Section 4 describes the regulatory compliance. Section 5 describes Algorithm and XAI design; Section 6 and 7 describes Data collection, preprocessing and Model design. Following with XAI Integration, Implementation Notes, Results and discussion and finally with conclusions along with the implications of the study.

2. Background and Related Work

2.1. Identity Theft in the Financial Domain

Definition and impact: As previously discussed, the use of a person's personally identifiable information for illicit purposes or the creation of fictitious identities is referred to as identity theft. Identity theft can be used to facilitate other types of financial fraud such as New Account Fraud, or Account Takeover fraud by allowing fraudsters to bypass identity authentication, withdraw money, or accumulate debt using a victim's identification credentials including financial and emotional stress suffered by victims, negative impact to their credit, and legal issues are also possibilities; institutions may suffer a variety of financial and non-financial losses including monetary losses, regulatory fines, and reputational damage [3]. Specifically, in 2024 New Account Fraud alone accounted for over $6.2 billion in losses, which was an increase from $5.3 billion in 2023, and this highlights how vulnerable online account opening systems continue to be [1]. Further, the COVID-19 pandemic has significantly accelerated the adoption of digital onboarding, and has exposed vulnerabilities in traditional in-person verification methods.

2.2. Identity Theft Techniques and Synthetic Identities

Identity thieves currently employ many different ways of conducting their fraud. Stolen personal data (i.e. Social Security Number, Driver License Numbers, Passwords) is commonly used for identity theft; this can be acquired from hacks, phishing attempts [3] or by simply obtaining it via other means (Social Engineering), while weak identity verification mechanisms allow identity thieves to obtain new identification with the least amount of resistance to the process of acquiring a new ID. Identity thieves are increasingly employing **Artificial Intelligence** including deepfakes in order to avoid being **detected during video-based identity verification** processes. These deepfakes are created to appear as a photo of the thief but are actually an artificially generated image of someone else whose identity was previously stolen (to avoid the identity verification mechanism if it is unable to detect “liveness”) [4]. **Document spoofer** (document spoofing) remains a problem as well since institutions are still unaware of how to determine when minor changes have been made to an ID which has gone undetected [4].

Synthetic Identity Theft is extremely difficult to detect. Thieves take advantage of combining both legitimate and fictitious information (for example, a legitimate SSN combined with a fictitious first and last name) to create an identity profile and history, then after creating these identities they execute high dollar transactions as the thief. Since there is no legitimate victim for the institution to flag as a potential problem, the synthetic identity passes all of the initial checks associated with opening an account [4].

2.3. Historical Methods of Defense

In the past, identity verification during the account opening process has utilized several methods including:

- Knowledge Based Authentication (KBA) - Asking security questions or One-time passcodes.
- Database Checks - Comparing provided data to Credit Bureau files/Government records.
- Manual Document Inspection - Staff reviewing and inspecting documents submitted by customers.

The United Nations and Law Enforcement Agencies differentiate between **Identity Theft (Stealing Information)** and **Identity Fraud (Using Stolen Information)**. In reality, organizations need to address both aspects of identity fraud - preventing both the theft of data and the fraudulent use of the data [3]. In our case, we focus solely on the detection of fraudulent use of the data at the point of entry to a financial system.

3. AI Techniques for Fraud and Anomaly Detection

3.1. Machine Learning in Fraud Detection

AI and Machine Learning (ML) are increasingly applied in fraud detection due to their ability to uncover complex patterns in large datasets. **Supervised models**—such as neural networks, logistic regression, and decision trees—have shown success in classifying fraudulent financial transactions using labeled historical data [5,7]. However, identity theft poses unique challenges due to the limited availability of confirmed fraudulent identity samples, especially with new fraud schemes. **Unsupervised and anomaly detection models** (e.g., One-Class SVMs, Isolation Forests, Autoencoders) are valuable in this context. These models learn what legitimate identity data looks like and flag deviations. Hybrid approaches, such as the CCAD model by Almalki and Masud (2025), combine outlier detectors and boosted learners to capture rare fraud cases by identifying inconsistent signals across models [7].

3.2. Feature Engineering for Identity Data

AI applications are successful when deployed as part of an organized set of features generated by Identity data. Examples of such features would be:

a. Validation of cross-checking results (SSN Validation; Phone/Address Consistency)
b. Signal identification based on patterns (Disposable Email Formats)
c. Confidence levels of matching biometrics (Selfie vs. ID Photo)
d. Features related to Device/Network characteristics (IP Geolocation; Device Fingerprint)
e. External risk assessment scores (Fraud Blacklists).

ID verification platforms employ a multi-modal approach to integrate document verification, biometric verification, and data verification layers to verify identities. Vaidya & Awasthi (2025) state that “Identity 3.0” is comprised of four AI-based components: **Document Verification** (using Optical Character Recognition [OCR] and computer vision); **Biometric Verification** (facial recognition, etc.; Liveness Checks); **Data Verification** (cross-checking against trusted sources); and **Risk Assessment** (Behavioral Analytics Engines) [10]. The layered nature of these systems has been shown to greatly outperform manual review processes. Examples of the benefits of using layered approaches to ID verification include identifying font anomalies in documents, detecting MRZ inconsistencies in documents and using liveness checks to resist Deepfakes [4,10]. Of note, approximately **79% of identity fraud detection studies published within the last year have used biometric methods**, primarily facial recognition [3].

3.3. Previous work on identity theft detection

There have been numerous studies focusing on financial fraud but few studies specifically focus on identity theft. Using historical U.S. public records to develop supervised models (i.e., decision trees, SVM), Mitchell & Sambasivam (2025) provided evidence of reasonability in terms of model accuracy. The authors also emphasize that there is a need for detailed information about cases when developing useful predictive methods for fraud prevention [6]. Agarwal (2021) investigated the use of machine learning for detecting online identity theft, and discovered that using random forests, logistic regression, and neural networks significantly improve fraud detection compared to rule-based systems for recognizing fraudulent use of identities [5,7].

Many financial organizations use AI-based identity verification software to help implement their KYC processes. According to reports from these organizations, the use of this type of software has resulted in an **increase of fraud identification of at least 40%** when comparing it to manual reviews of data based on digital signals [10]. Post onboarding, many banks have begun to utilize **User and Entity Behavior Analytics (UEBA)** for ongoing analysis of user behavior to identify suspicious activity which may be related to an account takeover—often a result of identity theft. Although this paper focused on the onboarding process for new users, implementing both a robust onboarding process and utilizing UEBA provides a multi-layered approach to fraud protection [3].

4. Explainable AI and Regulatory Compliance

4.1. The Need for Explain-ability

Complexity associated with AI technology has raised concerns in the banking world regarding justification of decisions made using the technology (like account opening denials). Banking decisions and regulatory requirements for transparency (as required in GDPR and proposed AI Act for high-risk AI technologies such as Credit and Identity Verification) [7,11] necessitate the use of XAI to provide explainable results of decisions made by a model.

XAI helps identify biases within a model and support fairness; it also aids in resolving disputes by providing information on how the model arrived at a decision [12]. For example, XAI could describe why a user was identified as “unusual” based on document(s) used for identification and/or device fingerprint(s).

4.2. Techniques used in XAI

A number of post-hoc techniques exist to provide model agnostic explanations such as SHAP and LIME. SHAP calculates the contribution score of each input feature (e.g., ID authentication, IP Mismatch) to a decision made by a model, whereas LIME creates a simplified representation of the local model behavior around a data point by creating an interpretable linear model. Both of these tools have been broadly accepted in fraud detection to convert opaque risk scores into understandable insights [7].

Although simpler models (e.g., decision trees) can be inherently interpretable, they often trade off this characteristic for accuracy. Therefore, using XAI to

provide explanations to complex models is generally preferable.

Benefits of Using XAI in Identity Theft Detection:

- **Transparency and Trust**: Explain-ability allows customers and analysts to understand the reasoning behind AI decisions which will *improve both transparency and trust in AI-based systems*, and allow organizations to comply with GDPR [9].
- **Compliance with Regulations**: XAI enables organizations to demonstrate their fairness obligations under US and EU regulations by providing clear rationales for decisions which negatively impact users.
- **Operational Efficiency**: Explanations enable analysts to rapidly verify alerts and **decrease false positive rates** (for example, identifying why a flag was triggered due to legitimate international usage) [8,13].
- **Model Validation**: Explanations enable developers to evaluate fairness and eliminate dependence on spurious features in their models which is particularly important in sensitive areas such as identity verification.

Using XAI in Identity Theft Detection Research and Practices:

Studies have shown that achieving high levels of interpretability does not come at the cost of performance. Almalki & Masud (2025) were able to achieve an area under curve (AUC) of .99 using SHAP in conjunction with a stacking ensemble and demonstrated that performance and interpretability can coexist. Other studies have also demonstrated high AUC values (near perfect) using LIME enhanced fraud detection models [7]. Banks and other financial institutions are increasingly implementing interpretable AI and are incorporating "reason codes" provided by vendors (e.g., "ID expired", "High-Risk Email Domain") to address their auditing and customer transparency needs [8].

In conclusion, the literature and current solutions demonstrate that an effective identity theft detection system in finance should employ **AI-based anomaly detection** to address the sheer scale and complexity of modern fraudulent activity, and should include **XAI** to make the decisions of the system transparent, accountable, and practical. The following section describes the proposed algorithm and system architecture that integrates both of these components, and employs best practices [10] in a practical workflow designed for real time ingestion of data.

5. Methodology: Proposed Algorithm and XAI Framework

A conceptual model is presented for detecting identity theft at the ingestion point (on-boarding) using AI-based anomaly detection along with **explainable AI** and modular verification pipelines to automatically evaluate identity risk in real-time during on-boarding. **Although hypothetical**, this model can be adapted for actual use with a variety of data sets and algorithms.

System Architecture Overview

The system architecture is depicted in Figure 1 and represents a "smart gatekeeper" that provides onboarding services through the execution of several basic functions:

The main components of the architecture are:

1. The **Data Ingestion Interface** is the first step in the workflow that allows users to enter their information such as by filling out an online application or uploading documents (the interface connects to the back-end AI processing system).
2. The **Preprocessing Module** is the second module of the workflow that validates and standardizes the user's input (for example it can be used to standardize names, check if social security numbers are valid etc.), this is to ensure all of the necessary fields have been filled in by the user and can optionally verify the user's information with outside sources (such as credit bureau services).
3. The **Document Verification Service** is the third module of the workflow that uses AI (for example, computer vision and optical character recognition) to validate the authenticity of the user provided documents, extract relevant fields (for example, the user's name and ID number), and identify potential anomalies (for example, possible photo-shopped images or poor-quality photos of faces that do not accurately match the selfie images taken of the applicant). The service will also provide confidence scores or flag anomalies (for example, an altered image, or a very low-confidence face-matching).
4. **Feature Extraction & Risk Engine** is the fourth module of the workflow that converts the user provided identity attributes into numerical vector representations of features for use in machine learning models. Some examples of the types of features that could be extracted include geographic location discrepancies between the submitted address and IP address, or domain risk associated with the submitted email address, or results returned from third party fraud databases.
5. The **Anomaly Detection Model** is the fifth module of the workflow that is a hybrid machine learning model that has been trained on both known fraud and legitimate samples. This model will output a fraud probability or anomaly score based on the user-provided data that will then be compared to predetermined threshold values to determine how the data should be processed.
6. **Decision Logic** is the sixth module of the workflow that will apply the predefined risk thresholds to the calculated fraud probability or anomaly score to determine whether to alert the user that they are at high-risk, route the case to a human investigator for manual review, or automatically approve the user's request.
7. The **Explain-ability Module** is the seventh module of the workflow that generates human readable justification for why a specific case was flagged (for example, "phone number", "Geographic Location Mismatch") using techniques such as SHAP or rule-based logic. The generated report assists compliance staff and investigators in reviewing flagged cases.
8. Finally, the **Analyst Dashboard is** the last module of the workflow that provides the investigator

access to the case(s) that were flagged, along with the generated explanation, and any additional data that supports the flagged case(s). Any feedback (fraud confirmed or false positive) from the investigators will be used to retrain and further improve the machine learning.

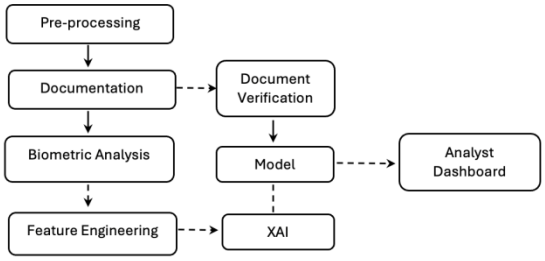

Figure 1. System Architecture for Identity Theft Detection at Data Ingestion, conceptually aligning with "Zero-to-One" framework's layers. [10]

6. Identity Data Collection and Pre-processing

6.1. Data Collection

During account creation or batch onboarding, the system collects multiple layers of identity attributes to prepare a complete risk profile:

1. **Personally Identifiable Information (PII)** Collected via form fields: Full Name, Date of Birth, SSN, Email, Phone Number, Address.
2. **Uploaded Documents** Scanned or photographed government-issued IDs (e.g., Driver's License, Passport). Multiple documents may be requested.
3. **Biometric Data** Live selfies or video used for facial comparison with the ID. Some systems require the user to hold their ID.
4. **Device & Environmental Data** Collected automatically (e.g., device fingerprint, IP geolocation, VPN usage). These help detect risk patterns (e.g., TOR network access).

6.2. Data Pre-Processing:

1. **Validation**: All necessary information fields should be populated with the appropriate format data to ensure that there are no obvious discrepancies (e.g., an age of 150).
2. **Normalizing Data**: Format all data in the same case (i.e., lower-case), remove leading/trailing white space from free-form fields, standardize date formats (e.g., MM/DD/YYYY), and convert address fields into geo-coordinates.
3. **Checks Against External Data Sources**: Use third-party APIs to perform background checks against public data sources (e.g., name and Social Security Number, email risk assessment by credit bureau).
4. Document Image Processing:

- Optical Character Recognition (OCR): Read the fields off of the documents (e.g., ID number).
- Authentication: Identify if the document has been altered or created as a template using Artificial Intelligence (AI) models designed to recognize the differences between real/fake IDs [14].
- Facial Comparison: Compare the applicant's self-submitted photograph to the photograph contained within their ID; generate a comparison score and determine whether the individual was live during the application process by performing blinking/motion or spoofing detection [14].

5. **Feature Vectors**: The last phase takes all of the generated signals and creates a structured vector features for machine learning models, which include:

- Signals derived from personal identifying information (e.g., age, the zip code/state combination).
- Verification result signals (e.g., whether or not the Social Security Number matched, whether or not the applicant is listed in a fraud database).
- Document scores/types (e.g., Document Authenticity Score (0-1), Document Type (Categorical: DL, Passport, Etc.), ID Expiration Validity.
- Signals indicating the degree of match/liveness of the biometrics used in the verification process.
- Signals associated with network/device risks (e.g., internet-based phone numbers flag).
- Signals related to behavior and time (e.g., the applicant submitted their application at an unusual hour, the applicant reused their device).

All of these features are then normalized/scaling (e.g., one-hot encoding), creating a numeric X representation that can be analyzed to detect anomalies.

7. Anomaly Detection Model Design

The main method to find out whether a new identity record, X, is likely **fraudulent or an outlier compared** to typical legitimate identities is a hybrid approach using supervised learning (recognizing established fraudulent patterns) and unsupervised anomaly detection (detecting both new and uncommon fraudulent patterns) along with Ensemble Model. Table 1 compares supervised, unsupervised, and ensemble models.

8. Explainable AI Integration

In order to increase transparency of the system, an XAI (Explainable Artificial Intelligence) module has been developed and integrated into the system to provide additional explanation when a flagged case does not have a clearly defined low risk designation. This module will include two levels of explanation:

1. **SHAP feature attribution**: SHAP is being used to interpret the outputs of supervised models. For example, a score of $p = 0.87$ could be explained by SHAP to include factors such as a face mismatch (+0.30), document authentication failure (+0.20), and geographic inconsistencies (+0.15). Each of these insights will then be translated into a clear

statement for review by the reviewer.

2. **Rules-based human-readable justifications**: In addition to using SHAP to explain supervised model outputs, this system includes a set of pre-defined rules that will generate human-readable messages.

For example:

- Failed ID document authenticity checks.
- The IP address (Nigeria) does not match the U.S. address provided."
- The phone number associated with the account was previously identified as fraudulent or is an internet-based phone number service.
- The Social Security Number (SSN) issuance date does not match the age claimed by the applicant.

Each of these rules can be easily mapped to common analysis steps performed by analysts and will therefore assist in expediting the analysis process. An example of how both types of explanations could be combined would be:

"Flagged due to: (1) Failed ID document authenticity checks (confidence level: 40%), (2) Biometric mismatch (Score: 50%), (3) IP address location discrepancy. The email domain also indicated that there were high risk uses.

Table 2 presents the different tools, strengths and limitation in explainable AI:

Table 1. Comparison of Fraud Detection Model Types

	Algorithm	Strengths	Limitations	Use in Framework
Supervised Classification	XGBoost, LightGBM	High accuracy; learns known fraud patterns	Needs labeled fraud data; less effective for new fraud	Primary fraud detector
Unsupervised Anomaly Detection	Isolation Forest	Detects rare/unseen fraud; no labels needed	May flag legitimate outliers; needs good legit sample	Secondary detector
Ensemble Model	Weighted average or dual rule	Balances strengths of both models	Requires calibration; adds complexity	Risk score generator

Table 2. *Explainable AI Techniques for Identity Fraud Detection* XAI Tool/Method

	Description	Applicable Features	Strengths	Limitations
SHAP (Tree SHAP)	Calculates feature contributions to model output	Numerical/categorical model inputs (e.g., document score, IP risk)	Accurate, additive, widely used in finance	Computationally intensive for large models
LIME	Local surrogate model explains individual predictions	Text/image/tabular (localized around one instance)	Model-agnostic, simple to implement	Can vary between runs; less stable than SHAP
Rule-Based Explanations	Predefined human-readable triggers	Document flags, location mismatches, device anomalies	Transparent; easy to audit	Not comprehensive; limited nuance
Hybrid (SHAP + Rules)	Combines both types of explanations	All features	Balanced interpretability and completeness	Requires design effort

9. Algorithm Pseudocode and Example

To summarize the detection approach, we provide a pseudocode encapsulating the entire pipeline:

```
Algorithm IdentityIngestionXAI_Detector(record):

    # Step 1: Preprocess and extract features
    X ← extract_features(record)
    # Includes identity fields, document metrics, biometric scores, device/network signals

    # Step 2: Model predictions
    p ← SupervisedModel.predict_proba(X) # Supervised fraud probability [0,1]
    s ← AnomalyModel.score(X) # Unsupervised anomaly score [0,1]

    # Step 3: Risk scoring (hybrid model fusion)
    w₁ ← 0.7 # Weight for supervised model
    w₂ ← 0.3 # Weight for anomaly model
    R ← w₁ * p + w₂ * s # Combined risk score

    # Thresholds (calibrated during model validation to balance precision/recall)
    Θ_high ← 0.8 # Threshold for automatic rejection
    Θ_low ← 0.5 # Threshold for manual review

    # Step 4: Decision logic
    if R ≥ Θ_high:
    outcome ← "Suspicious"
    else if Θ_low ≤ R < Θ_high:
    outcome ← "Review"
    else:
    outcome ← "Verified"

    # Step 5: Generate explanation if required
    explanation ← None
    if outcome ∈ {"Suspicious", "Review"}:
    explanation ← XAI_module.generate_explanation(X, p, s)
    log("Flagged record explanation: " + explanation)

    return outcome, explanation
```

Illustrative Example:

Alice Doe sends in a simple application for an account using only her name, address, etc., along with a picture of her driver's license and a self-portrait. Unfortunately, she doesn't know that the Social Security Number she provided belongs to someone who has died; and, that the ID she sent in is a very good forgery.

Her application will be processed by the system and the results are as follows: the document authenticity score is "low" (60%); the biometric (face) score is "moderate" (70%); the SSN verification score is "flagged" because it shows up as belonging to a deceased person. Her supervised fraud probability is "high" (.95) and the

Anomaly Detector flag the transaction as "unusual".

As such, the System will determine that her application has been flagged as "Suspicious", and include the following information in the explanation:

- ID document did not pass authenticity tests
- SSN is tied to a deceased individual
- Biometric verification was inconclusive

These explanations from algorithm will provide investigators with where to focus their attention, resulting in improved detection in early stage, and reduced need for manual review.

10. Implementation Notes

To show that the proposed hybrid identity theft detection method can be implemented in a full python-based reference implementation utilizing free libraries, we created a synthetic dataset as well as a full Python-based reference implementation using open-source libraries. The synthetic dataset mimics real-world signals for each type of signal such as: Document Authenticity, Biometric Match and Liveness Scores, Geolocation Consistency, VPN/TOR Usage, Device Reuse, Email Risk, internet-based phone number Detection and High-Risk SSN Patterns due to the sensitive nature of real onboarding datasets.

The full python reference implementation has three main layers:

- **Fraud Probability Supervised Model (XGBoost)**: Trained on engineered features to predict the fraud probability, we chose XGBoost for its compatibility with SHAP as well as its accuracy.
- **Anomaly Detection- Unsupervised Model (Isolation Forest)**: trained on legitimate identity samples to model typical identity characteristics, we used anomaly scores from isolation forest in conjunction with supervised outputs to generate a weighted sum:

R = w1·psupervised + w2·sanomaly

where $w_1 = 0.7$ and $w_2 = 0.3$.

- **Explanation Layer (Rules, SHAP, LIME)**: we utilized SHAP for feature attribution, LIME for providing local explanations, and rule-based tags (i.e. "VPN Usage" or "low document score") to provide compliant aligned, human readable explanations.

The resulting implementation provided several key results including: ROC Curve comparisons, SHAP plots, LIME reports, and analyst facing csv output reporting decision rationale and explanation code. All testing occurred on standard CPU's and resulted in an inference time less than 100ms, validating the system's ability to operate in real-time. Python code and data generation script examples are available in the project's GitHub repository.

GitHub Repository: https:// github.com/ AI-SyntheticData/IdentityTheft_XAI

11. Results and Discussion

The proposed identity fraud detection system is a hypothetical example of an architecture that has shown similar success in literature for fraud related tasks. Previous fraud related task examples of hybrid architectures combining both supervised and unsupervised machine learning methods have demonstrated AUC-ROC scores >0.95; with one example demonstrating an accuracy rate of 98.3% while maintaining the ability to use Explainable AI to enhance model interpretability [7].

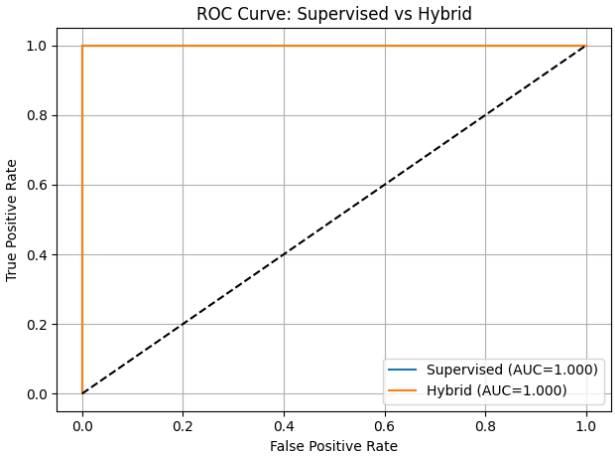

Figure 2. ROC Curve

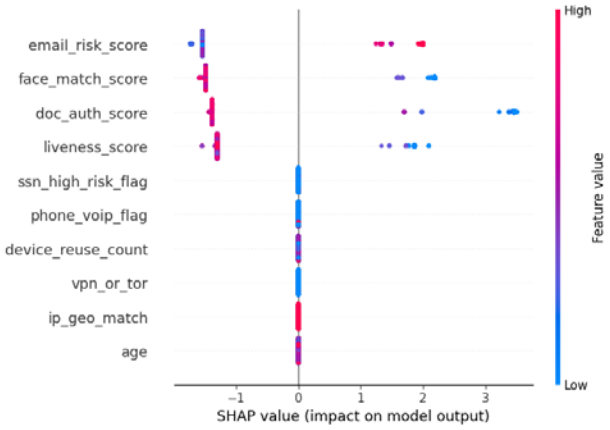

Figure 3. SHAP Summary Plot

Using the fraud detection system to intercept fraudulent accounts at the onboarding phase will help to mitigate future losses from fraud. There is industry evidence to support that using advanced ID verification to identify and prevent new account fraud will reduce it by approximately 50%. In addition, incorporating Anomaly Detection into the system will provide the ability to detect new forms of fraud, such as synthetic identities [14]. To address false positives generated by the fraud detection system, the system utilizes tiered results (Verified, Review, Suspicious), and a conservative threshold (i.e., <0.5% auto-reject, ~2-3% review). The incorporation of XAI provides additional insight into the model's decision-making process and will provide analysts the tools to quickly resolve alerts, as was demonstrated in previous studies which showed an improvement in trust and efficiency in fraud detection workflows [15].

Figure 2 illustrates the ROC performance of the hybrid model and Figure 3 presents the SHAP summary plot highlighting dominant risk features.

12. Regulatory and Business Impact

The implementation of the system we have developed meets the regulatory requirements as illustrated below in Table 3:

Table 3. Table mapping key features of your system to GDPR, AMLD5, and US CIP regulatory requirements [10,16,14,15]

System Feature	GDPR Alignment	AMLD5 (EU Directive) Alignment	US CIP (Bank Secrecy Act) Alignment [17]
Document Verification	Supports data minimization and accuracy; complies with identity verification under legitimate interest.	Fulfills electronic identification for CDD/KYC.	Supports 'reasonable belief' of identity verification.
Biometric Verification & Liveness	Requires explicit consent under Article 9; enhances identity assurance.	Reinforces customer identification in remote onboarding.	Enhances verification via non-documentary methods.
Anomaly Detection Engine	Contributes to automated decision safeguards (Article 22).	Aids in detecting high-risk clients for EDD.	Identifies suspicious profiles at onboarding.
Explainable AI Module	Supports 'right to explanation' and transparency obligations.	Supports traceability of due diligence efforts.	Provides rationale for decisions to regulators.
Consent Collection for Biometric Data	Ensures explicit consent (Article 7, 9).	Addresses biometric controls explicitly recognized.	Documents user awareness and permission for processing.
Audit Trail Generation	Meets documentation and accountability principles.	Supports regulatory reporting and case traceability.	Provides logs for audit and regulatory review.
Risk-Based Decision Logic	Aligns with proportionality for data processing and decision fairness.	Aligns with EDD and ongoing monitoring protocols.	Meets CIP mandate for adaptable identity verification procedures.

13. Challenges and Limitations

The fraud prevention system has its advantages, however, there are significant disadvantages and limitations.

- **Availability and Quality of Data**: A robust supervised model needs large amounts of fraud data that have been accurately labeled, and smaller financial institutions do not typically have this kind of data available. If the data is shared, then privacy becomes an issue. In addition, if the input data quality is poor, or if the outcomes were inaccurately labeled, then the overall accuracy of the model suffers. Anomaly detection allows for some independence from labeling data, as long as you have a good representation of legitimate user data to prevent false positives [3].
- **Fraudster's Rapidly Evolving Tactics**: Fraudsters quickly adapt their tactics, such as spoofing their IP location to avoid being detected. Training your models continuously will allow them to eventually detect these tactics, but new tactics (such as new forms of deepfakes), will likely take time before they are detected. Therefore, both initial onboarding checks, and post-login monitoring, should be used together to catch as many fraudulent transactions as possible [3].
- **Over-reliance on Artificial Intelligence**: There is always the possibility of a false negative when using artificial intelligence to flag potential fraud, not every fraudulent transaction will be flagged. Humans are still needed to analyze the flagged transactions, and provide a second level of review. Our design incorporates features that allow analysts to override the system's flagging decision, and provides a mechanism for analysts to feed back into the system so that it can continue to improve over time.
- **Privacy and Ethical Risks**: As we are dealing with sensitive information about individuals and using external sources (such as credit reporting agencies), there is an increased risk to individuals' privacy, and to the ethical standards of our organization. To address these concerns, we must implement strong security and privacy measures, such as encryption, access control mechanisms, etc. We also need to ensure that our models are fair and unbiased in their treatment of users. One way to measure fairness is to make sure that the explanations for why a particular pattern was identified as potentially fraudulent, are understandable and transparent. In addition, under GDPR, we must obtain explicit consent from our users to automatically process their biometric data.
- **Integration and Cost**: Implementing the system requires integration with existing banking systems and APIs, and ongoing maintenance and training costs. However, many of these costs are likely to be offset by the reductions in fraudulent activity that we anticipate.

14. Conclusion

Digital onboarding continues to be plagued by identity fraud as the lack of prior information and adaptive fraudsters continue to hinder all forms of verification (classical or AI-based). A number of authors have shown that it is possible to detect identity fraud at the very earliest point during digital onboarding (i.e., when the applicant provides their identifying information) and therefore prevent costly downstream issues related to the applicant's ability to access the accounts they create; operational burdens due to additional customer service requests; and regulatory exposures due to fraud-related reporting requirements.

A new hybrid framework was designed which combines supervised classification with unsupervised anomaly detection to allow for both well-known and previously unknown types of identity abuse. In contrast to many existing models that use either only supervised classification or only unsupervised anomaly detection, this combined framework will improve the robustness of fraud detection systems to data sparseness and improve the adaptability of these systems to emerging attack methods.

Additionally, the incorporation of explainable AI into the framework provides end-users of the system (e.g., analysts, regulators) with a rational explanation of why a particular account has been approved or denied and thus supports informed decision-making and satisfies regulatory requirements related to transparency and accountability.

Operationally, the proposed framework allows for the implementation of a risk-based onboarding process. As such, the framework will support the automation of approval processes for low-risk applicants, facilitate manual reviews of ambiguous identities, and provide for automatic denial of high-risk applicant profiles. The results indicate that explainability is not only beneficial to detection performance, but rather necessary to ensure the responsible deployment of AI within regulated identity verification systems.

Future work should include the empirical validation of the proposed framework using a large-scale anonymized dataset from onboarding, the development of a methodology for privacy preserving collaboration between organizations to share knowledge regarding fraudulent activity, the extension of the proposed framework to defend against emerging threats such as deepfakes and document forgery, and the extension of the framework to enable continuous identity assurance after initial onboarding. While the proposed framework is focused on identity verification for financial services, the general principles can be applied to numerous other industries that require secure and explainable digital identity verification.

ACKNOWLEDGMENTS

The author gratefully acknowledges the use of OpenAI's ChatGPT in assisting with sentence phrasing, language refinement, and grammar polishing during the preparation of this manuscript. All ideas, analyses, system designs, and technical contributions presented in this work are solely the author's own. No external funding or institutional support was received for this research, and the author declares full responsibility for the content herein.

References

[1] Ianzito, C. (2025). *Identity Fraud and Scams Cost Americans $47 Billion in 2024*. AARP. Available: https:// www.aarp.org/ money/ scams-fraud/javelin-identity-theft-report-2024/.

[2] Federal Trade Commission (FTC). (2025). *New FTC Data Show a Big Jump in Reported Losses to Fraud to $12.5 Billion in 2024*. Available: https:// www.ftc.gov/ news-events/ news/press-releases/2025/03/new-ftc-data-show-big-jump-reported-losses-fraud-125-billion-2024.

[3] Zhang, C. J. (2025). *AI-based Identity Fraud Detection: A Systematic Review*. arXiv preprint arXiv:2501.09239. Available: https://arxiv.org/abs/2501.09239.

[4] Sugavanam, A. (2025). *AI Threats Pose New Fraud Risks, But AI Can Also Defend Banks*. The Financial Brand. Available: https://thefinancialbrand.com/news/banking-technology/ai-threats-pose-new-financial-frauds-but-ai-can-also-defend-banks-192450/.

[5] Agarwal, V. (2021). *Identity Theft Detection Using Machine Learning*. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 9(8), 1943–1949.

[6] Mitchell, C. D., & Sambasivam, S. (2025). *Predictive Modeling for Identity Theft Detection: A Design Science Approach Using Machine Learning and Historical Data*. Issues in Informing Science and Information Technology, 22, 1–13.

[7] Almalki, F., & Masud, M. (2025). *Financial Fraud Detection Using Explainable AI and Stacking Ensemble Methods*. arXiv preprint arXiv:2505.10050. Available: https:// arxiv.org/ abs/ 2505.10050.

[8] Kaur, G., et al. (2025). *Explainable AI for Regulatory Compliance in Financial and Healthcare Sectors: A Comprehensive Review*. International Journal of Advances in Engineering and Management (IJAEM), 7(3), 489–494.

[9] Sai'd, Z. (2025). *Explainable AI (XAI) in Identity Access Management: Bridging Trust and Transparency in User Authentication*. TechRxiv preprint.

[10] Vaidya, A., & Awasthi, A. (2025). *Zero-to-One IDV: A Conceptual Model for AI-Powered Identity Verification*. arXiv preprint arXiv:2503.08734. Available: https:// arxiv.org/ abs/ 2503.08734.

[11] Palo Alto Networks. (2024). *What Is Explainable AI (XAI)?* Available: https:// www.paloaltonetworks.com/ cyberpedia/ explainable-ai.

[12] FinTech Global. (2025). *Is Explainable AI the Missing Link in Regulatory Compliance?* Available: https:// fintech.global/ globalregtechsummitusa/is-explainable-ai-the-missing-link-in-regulatory-compliance/.

[13] ESS Open Archive. (2025). *Interpretable Machine Learning in Financial Risk Systems*. Available: https:// essopenarchive.org/ browse-all ? tags=% 5B% 22interpretable +machine+ learning%22%5D.

[14] The Financial Brand — Deepfake Fraud Examples & Document Spoof Detection (2025). Available: https:// thefinancialbrand.com/ news/banking-technology/ai-threats-pose-new-financial-frauds-but-ai-can-also-defend-banks-192450.

[15] Deloitte Insights. (2024). *Explainable Artificial Intelligence (XAI) in Banking – Towards Transparency*. Available: https:// www. deloitte.com/us/en/insights/industry/financial-services/ explainable - ai-in-banking.html.

[16] Zero-to-One IDV Regulatory Appendix (AMLD5, GDPR). (2025). arXiv supplemental material. Available: https:// arxiv.org/ pdf/2503.08734.

[17] U.S. Bank Secrecy Act (Customer Identification Program Requirements). (Regulatory reference; no URL required.