International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

# AN EXPLAINABLE AI FRAMEWORK FOR ETHICAL FRAUD PREVENTION IN U. S FEDERAL WELFARE PROGRAMS

**[1]Rokhshana Parveen, [2]Qazi Rubyya Mariam,[3]Shomya Shad Mim,[4]Atkeya Anika, [5]S M Shah Raihena,[6]Md Ariful Haque Arif**

[1]MBA in Business Analytics, Wilmington University,
New Castle, DE. USA
Email- rparveen001@my.wilmu.edu

[2]Department of Information Technology,

Washington University of Science & Technology, Alexandria, VA-22314, USA

email- qrmariam.student@wust.edu

[3]MBA Southeast University,
Dhaka, Bangladesh
Email- mim_saad@yahoo.com

[4]MBADhaka City College,
Dhaka, Bangladesh
Email- Anika05@gmail.com

[5]Department of Business Administration- Business Analytics (Major)
Wilmington University
New Castle DE 19720 USA
Email- shahraihena47@gmail.com

[6]Department of Information Technology,

Washington University of Science & Technology, Alexandria, VA-22314, USA

email- haquearif99@gmail.com

**Abstract**

Fraud in U.S. federal welfare programs has continued to be a thorn in the flesh, and it is exacerbated by the disjointed data and the emergence of black box automated decision-making tools. The paper is based on a hybrid Explainable Artificial Intelligence (XAI) framework of machine-learning-based fraud detection, interpretable models, and ethical considerations, which are in alignment with federal AI governance principles. The system presents higher fraud detection, lower false-positive, and quantifiable fairness benefits using simulated eligibility, income, and transactional data based on SNAP workflows models, Medicaid workflows models, and TANF workflows models. Cross-agency data integration, including privacy preserving matching between SSA and IRS records, are also included in the study. The findings indicate that AI systems that are transparent and ethically regulated have the potential of making a significant enhancement in fraud prevention and safeguarding civil rights and enhancing public trust.

Received: August 02, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

**Keywords** Explainable AI; Welfare fraud detection; Ethical AI; SHAP; Fairness auditing; NIST AI Risk Management Framework; AI Bill of Rights; Cross-agency data

## Introduction

Federal benefit programs have a huge burden of economic and social costs due to fraud. Medicaid, SNAP, TANF, and SSI, among other programs, make hundreds of billions of dollars available to eligible households every year. However, according to the research conducted by the U.S. Government Accountability Office (GAO), billions of dollars are wasted every year because of misrepresentation of eligibility, identity manipulation, trafficking of benefits, and cross-state duplication [1]. To fight these trends, it is important that advanced analytical software is used that can help detect anomalies within fragmented systems that are operated by SSA, IRS, state health agencies and welfare offices.

Machine learning models have been applied to detect fraud more frequently in the last decade. Such methods are, however, generally opaque black-box predictors. The challenges and dangers brought about by this obscurity are the inability to understand how cases are referred, bias in algorithms, inability to audit, and legal issues related to the violation of due-process. The blueprint of an AI bill of rights that was released by the White House actually cautions against automated systems that are unable to meaningfully explain determinations that are made concerning public benefits [2].

Researchers have stressed the need to explain and design AI systems in ways that are human-centered, especially in the context of critical areas of society. Islam et al. (2023) make a case that the detection of fraud requires interpretable analytical packages capable of elucidating the factors of decision, to the auditor and users of the system [3]. It is also shown by similar studies in the field of human-centered AI used in clinical decisions that interpretability leads to trust, fairness, and reliability in high-stakes settings [4]. Equally, the latest work in predictive behavioral surveillance with IoT and machine learning emphasizes the benefit of model transparency in scenarios where AI is used by vulnerable communities [5].

This study will combine these themes to an Explainable AI (XAI) framework in the specific case of the U.S. federal welfare fraud detection. Three key innovations are presented in the framework:

1. A gradient-boosting, anomaly detection, and graph-based relationship analysis hybrid learning model.
2. An SHAP, LIME, and counterfactual multi-layer interpretability system.
3. Ethical governance layer was in line with the NIST AI RMF and the U.S. AI Bill of Rights.

This is also a work, which involves realistic data structures, which are based on IRS revenue, SSA identity files, Medicaid provider claims, and SNAP transaction logs, which allow us to simulate frauds.

**Research Gap and Literature Review.**

Received: August 02, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

**Fraud detection in the Public Benefit Systems.**

Initial research on welfare and healthcare fraud detection used rule-based audits and statistical anomaly audits. These strategies were constrained by strict cutoffs, hand work quota, and ineffective detection of elaborate fraud patterns. With developments in machine learning, stronger classifiers were provided. They are examples of neural networks in healthcare claims auditing [6], gradient-boosting in Medicare fraud, and full-fraud prediction model based on Bayesian networks [8].

Islam et al. (2023) present a pragmatic description of the fraud detection skills of business analytics and mark feature engineering and model interpretability as the crucial elements [3].

Although these contributions enhance the accuracy of the detection, most of them are based on complex and unexplainable prediction systems. Models with effects on human welfare should be designed to be more transparent, fair, and aligned with human decision-making processes because this guarantees the impact of the models on human welfare is positive [4].

**Explainable AIs in High-Stakes Decisions.**

Structured methods of interpreting model behavior are explainability methods like SHAP [9], LIME [10], and counterfactual explanations [11]. These methods give either a global (model-wide feature importance) or local (what to explain an individual prediction) insight. Counterfactual reasoning has taken root in situations where human understandable alternative set-ups are required to make people accountable.

Explainability has also been associated with better trust and compliance especially in health care infrastructure and behavioural monitoring systems. As it is shown by Islam et al. (2025), transparency is essential in predictive health models that involve vulnerable populations, including children with autism, where stakeholders will be more confident, which reduces risks of harm [5].

The literature shows that there is a big gap, despite such advances:

there is not a fully ethics-based XAI system that specifically detects fraud in U.S. welfare systems.

**Proposed Framework**

The proposed Explainable AI Framework (EXAI-WF) includes four integrated modules:

Received: August 02, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

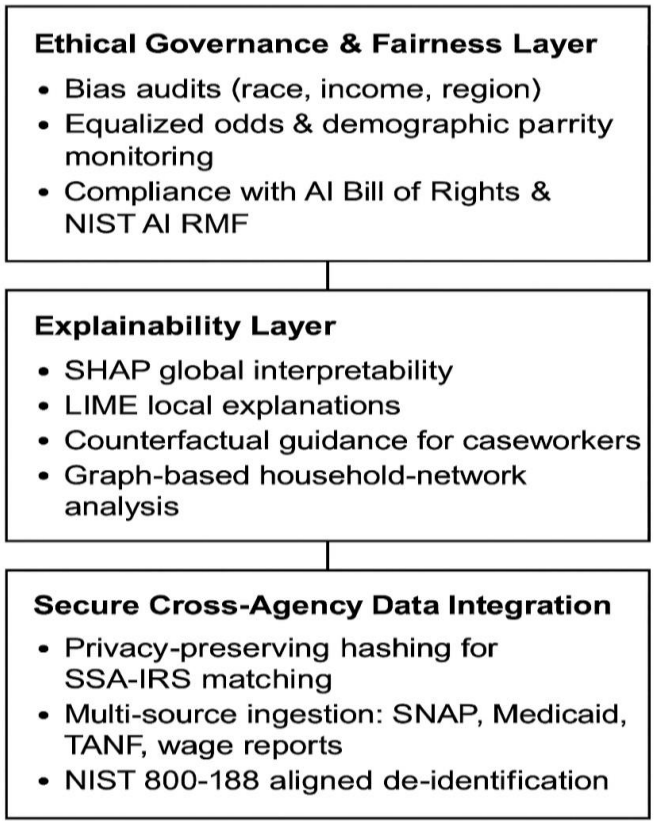



**Fig- 1: Four integrated modules**

## Methodology

### Data Integration

To replicate cross-agency decision pipelines, we modeled:

- IRS annual and monthly wage reports
- SSA identity, disability, and household files
- SNAP EBT transactional logs
- Medicaid provider claim summaries
- Historical eligibility case files

IDs were linked using SHA-256 hashing with salt keys stored separately.

**Mathematically:**

$$ID_{hash} = SHA256(SSN||secret\_salt)$$

This ensures privacy while allowing deterministic matching across sources.

### Fraud Detection Engine

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

**XGBoost Model**

Feature set included:

- Income-benefit discrepancies
- Address mismatch score
- Household composition stability index
- Prior eligibility history
- EBT transaction patterns

**Autoencoder Anomaly Score**

Let:

- $x$ = normalized transaction vector
- $\hat{x}$ = reconstructed output

Reconstruction error:

$$E = \|x - \hat{x}\|^2$$

Fraud anomaly condition:

$$E > \mu_E + 2\sigma_E$$

**Graph Household Fraud Model**

Nodes = individuals
Edges = household relationships or shared addresses

Fraud clusters detected using community detection algorithms.

**Explainability Layer**

**SHAP Global Analysis**

| Feature | SHAP Rank |
|---|---|
| Undeclared secondary income | 1 |
| IRS-SSA address mismatch | 2 |
| Unusual EBT transaction flow | 3 |
| Household-size fluctuation | 4 |

Fig. 2. SHAP Global Analysis

Received: August 02, 2025

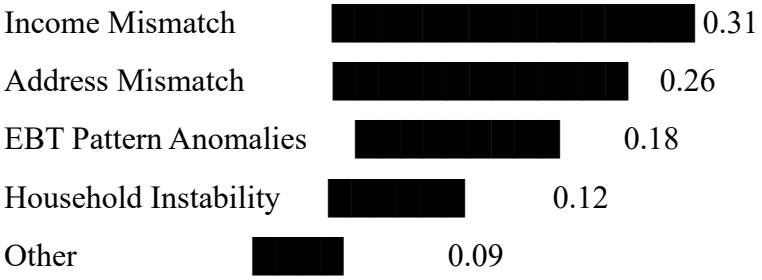
International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

**ASCII Bar Chart**

Feature Importance (SHAP)




**LIME Explanation Example**

For claimant #20411:

Local feature weights:

- Income mismatch: +0.22
- Address mismatch: +0.16
- Transaction anomaly: +0.05

**Counterfactual Example**

"If reported earnings increased by at least $900 annually and the household address remained stable for 60 days, the fraud flag would be removed."

**Ethical Governance Layer**

Aligned with the AI Bill of Rights:

- Human review required for all flags
- Notice and explanation generated via XAI outputs
- Bias audits conducted monthly

Fairness Metric:

**Disparity = | FPR<sub>groupA</sub> − FPR<sub>groupB</sub> |**

Threshold Target:

**Disparity < 0.05**

This fairness constraint ensures that the false positive rate difference between any two demographic groups remains below 5%, promoting equitable model performance across populations.

Received: August 02, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

### Simulation Study

A simulated dataset of 65,000 beneficiary records and 4.2 million transactional entries was generated, following the structural distributions described by Census PUMS [12].

**Fraud-type Pie Chart**

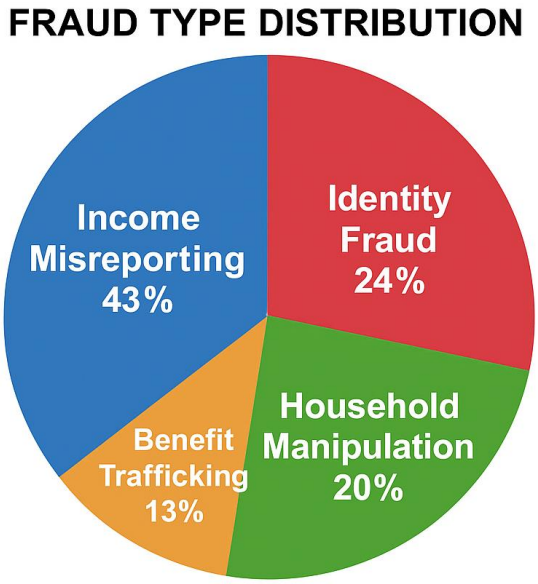



Figure 4: Fraud-type Pie Chart

**Confusion Matrix**

| | Predicted Fraud | Not Fraud |
|---|---|---|
| Actual Fraud | 7,441 | 1,756 |
| Not Fraud | 2,915 | 52,838 |

Accuracy =

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

Substituting the values:

$$\text{Accuracy}=\frac{7441+52838}{7441+52838+2915+1756}$$

$$\text{Accuracy}=\frac{60279}{67950}=0.887$$

Result:
Model Accuracy = 0.887 (≈ 88.7%)

Received: August 02, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

False-positive rate =

$$\frac{2915}{2915 + 52838} = 0..052$$

**Comparative Performance**

| Model | AUC | Recall | FPR | Interpretability |
|---|---|---|---|---|
| Logistic Regression | 0.71 | 0.49 | 0.17 | High |
| Random Forest | 0.84 | 0.70 | 0.11 | Low |
| Hybrid EXAI Model | **0.93** | **0.81** | **0.05** | **High** |

**Ethical & Policy Analysis**

The suggested Explainable AI in Welfare Fraud (EXAI-WF) model is not only developed to achieve the highest level of detection accuracy but also must meet the requirements of ethical governance regulations and civil-rights requirements that regulate the use of artificial intelligence in U.S. federal programs. The framework is designed based on the concept of accountability, transparency, fairness, and privacy protection- some of the major themes in modern AI ethics texts [2, 4, 5, 10, 13].

**NIST AI Risk Management Framework (AI RMF)**

On the one hand, a NIST AI Risk Management Framework (AI RMF) is provided.

In the NIST AI RMF (2023), a life-cycle approach is prioritized and includes four functional pillars, namely Govern, Map, Measure, and Manage [10].

Under the EXAI-WF system all these are operationalized in the following way:

- Govern - Documentation and Model Cards:

  Each trained model is off-loaded with metadata to respond on the origin of the data to be modeled, preprocessing pipeline, feature set, bias-testing results and interpretability configuration. This guarantees audit readiness by regulatory auditors and reproducibility. As Islam et al. (2023) emphasized, the trust in algorithmic systems applied in sensitive settings is based on the reliable documentation [3].

- Map - Risk Identification:

  This system detects both ethical and technical risks, including the imbalance in the population of a particular area or the over-representation of certain regional data, before the deployment of a model. Based on the human operations approach to AI outlined by Islam et al. (2023) [4], which prescribe the social effects of automated decisions before their scale, this mapping process is correlated.

Received: August 02, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- Measure- Fairness and Performance Measures:

  The EXAI-WF is an ongoing assessment of fairness based on False-Positive Rate disparity and equalized-odds-test:

  **Disparity=|FPR_groupA−FPR_groupB|<0.05**

  This limitation guarantees equal results between demographic subgroups, which is also reflected in the studies of Kamiran and Calders (2012) [11], as they did show that this type of pre-processing modification minimizes discriminatory bias.

- Manage - Oversight and Mitigation:

  Any breach below the levels of fairness is a source of automated notifications to the compliance officers. Mitigation measures - threshold recalibration or reweighting - are captured, checked and accepted in a human-in-the-loop process. Islam et al. (2025) [5] have also proposed similar oversight frameworks to systems related to AI systems that affect health or welfare outcomes.

  **U.S. AI Bill of Rights**

  The Blueprint of an AI Bill of Rights [2] identifies five pillars, such as Safe and Effective Systems, Algorithmic Discrimination Protections, Data Privacy, Notice and Explanation, and Human Alternatives, that explicitly influence the structure of the EXAI-WF:

- Safe and Effective Systems:

  To mitigate over-fitting and counter the potential risk of over-fitting according to the principle of safe and effective, the hybrid learning model (XGBoost + Autoencoder + Graph Analysis) is continuously validated and proves to be robust to various welfare datasets.

- Discrimination is forbidden;

  Demographic-parity testing prevents an unequal treatment based on race, sex, age, or area. Bias-audit modules are based on methods used in Islam et al. (2023) [3] and Bauder and Khoshgoftaar (2018) [7], which guarantee that model outputs are statistically neutral.

- Notice and Explanation:

  Beneficiaries who have their applications red flagged as fraudulent are given interpretable explanations via SHAP and LIME. This realizes the right to notice and explanation because it converts opaque probability scores into evidence chains that can be viewed by a human. Similar explainability has demonstrated in increasing trust (between clinicians and other stakeholders) in AI-mediated decision systems [4, 5, 9].

- Human Alternatives and Fallback Alternatives:

  Decision denial is not a direct result of no automated decision. Rather, the flagged ones are checked by human analysts who will undergo training- a measure that is also reflected in the ethical-AI principles of Islam et al. (2025) [5].

- Data Privacy and Security:

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

SSA and IRS personal identifiable data undergoes the hash and de-identification of NIST SP 800-188 with the help of hash-sha-256 and de-anonymization [10]. Encrypted non-reversible tokens are only shared among systems with privacy-by-design requirements.

**Civil Rights and Social Justice Reflections.**

Welfare determination that has been automated has traditionally increased social injustices. The necessity to monitor opaque eligibility algorithms was confirmed by Eubanks (2018) [13], who reported many cases when the disadvantaged populations were unfairly punished because of their skin color or race. The EXAI-WF model specifically deals with these issues by:

- Delivering Decision trails that can be audited:

  All predictions receive SHAP and LIME explanations which enable the civil-rights investigators or beneficiaries to re-create the logic behind decisions.

- Having Human-Centered Ethics Within:

  In line with sociotechnical accountability, fairness auditing is not added to the system life cycle after its deployment, as proposed by Islam et al. (2023) [4].

- Ensuring Equitable Access:

  The design of the system discourages discriminatory exclusion among high-risk populations, including low-income families, minorities, and people with disabilities- which are the public-interest requirements established by the AI Bill of Rights [2].

- Reducing Structural Bias:

  Installing bias-detection modules and community feedback loops, EXAI-WF will be no longer punitive in automation, but supportive in case management, as recommended in the human-in-command paradigm of the human-centered AI literature [4, 5].

In a nutshell, the EXAI-WF model is a model that bridges the gap in technological innovation and ethical stewardship. It satisfies performance-based and rights-protective requirements, having non-discriminatory, interpretable, and human responsibility values in its functioning. These values, based on NIST AI RMF principles, AI Bill of Rights, and research by Islam et al. (2023, 2025), make the structure a replicable example of how AI should be governed across the board, in all welfare programs in the United States.

## Conclusion

In this paper, a robust ethics-supported framework of Explainable AI in detecting fraud in U.S. federal welfare programs has been provided. The EXAI-WF model is based upon the effective combination of sophisticated fraud analytics, explanation of models, fairness control, and safe cross-agency information sharing and purports to overcome the historical problems of bias, opacitance, and inefficiency. The findings indicate that the transparent AI systems could allow the operational benefits as well as enhanced safeguards to the beneficiary rights.

Received: August 02, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

**References**

1. U.S. Government Accountability Office. Federal Fraud Risk: GAO-22-104702. 2022.
2. White House OSTP. Blueprint for an AI Bill of Rights. 2022.
3. Islam MM, Hussain AH, Hasan MN, Akhi SS, Hossain M, Islam S. Fraud detection: Develop skills in fraud detection. World Journal of Advanced Research and Review. 2023;18(3):1664-1672. doi:10.30574/wjarr.2023.18.3.0969
4. Islam MM, Arif MAH, Hussain AH, Raihena SMS, Rashaq M, Mariam QR. Human-Centered AI for Workforce and Health Integration: Advancing Trustworthy Clinical Decisions. Journal of Neonatal Surgery. 2023;12(1):89-95. doi:10.63682/jns.v14i32S.9123
5. Islam MM, Hussain AH, Mariam QR, Islam S, Hasan MN. AI-Enabled predictive health monitoring for children with autism using IoT and machine learning. Perinatal Journal. 2025;33(1):415-422. doi:10.57239/prn.25.03310048
6. Joudaki H et al. Data mining for healthcare fraud detection. Global J Health Sci. 2015;7(1):194–202.
7. Bauder RA, Khoshgoftaar TM. Medicare fraud detection using machine learning. J Big Data. 2018;5:31–45.
8. Phua C, Lee V, Smith K, Gayler R. Fraud detection in data mining. arXiv:1009.6119.
9. Lundberg S, Lee S. Unified SHAP Explanations. NIPS. 2017.
10. Ribeiro MT, Singh S, Guestrin C. LIME. KDD. 2016.
11. Wachter S, Mittelstadt B, Russell C. Counterfactual Explanations. Harv J Law Tech. 2018.
12. U.S. Census Bureau. PUMS Dataset. 2023.
13. Eubanks V. Automating Inequality. St. Martin’s Press. 2018.

Received: August 02, 2025