BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

EXPLAINABLE AI FOR STOCK MARKET FRAUD DETECTION USING MACHINE LEARNING : *A Proposed Framework*

[1]Shravan Sachin Bhandari, [2]Prof. Kirti Gajbe, [3]Dr. Vishal Puranik

[1]Student, [2]Mentor, [3]Coordinator
[1,2,3]Department of Artificial Intelligence and Data Science,
[1,2,3]Ajeenkya DY Patil School of Engineering, Pune, India
[1]bhandarishravan760@gmail.com, [2]kirti.gajbe@dypic.in, [3]vishalpura@gmail.com

Abstract

Stock market fraud, including insider trading, market manipulation and deceptive trading, erodes investor confidence and is difficult to separate from legitimate market fluctuations. Machine learning (ML) can analyse large volumes of trading data and flag suspicious patterns, but many accurate models behave as black boxes, which limits their usefulness to analysts and regulators who must justify any follow-up action. This paper proposes an explainable ML framework for stock market fraud detection that couples fraud prediction with post-hoc explanations from SHAP and LIME. The framework covers data collection, preprocessing, feature engineering for suspicious trading patterns, comparison of supervised and anomaly-detection models, explanation generation, and an evaluation protocol based on precision, recall, F1-score, PR-AUC and false-positive rate, with human review of every alert. The work is a seminar-stage proposal built on a literature survey: no dataset has been assembled and no experiment has been run, so no performance results are reported. The paper instead specifies what must be built and measured to test whether explainability improves the trustworthiness of alerts.

Keywords: explainable AI, stock market fraud detection, machine learning, SHAP, LIME, market manipulation, insider trading, interpretable machine learning, FinTech

I. Introduction

Stock markets channel savings into investment, and the growth of digital trading platforms now produces very large streams of transactions and market data every day. The same growth creates openings for insider trading, market manipulation, deceptive trading and pump-and-dump schemes, which cause financial losses, distort fair price formation and reduce investor confidence. Such behaviour is hard to detect because it is dynamic, may involve several interacting factors, and can resemble ordinary market volatility. Conventional rule-based surveillance can be slow and can miss patterns that change over time.

ML offers a way to process large trading datasets and learn patterns associated with unusual behaviour. Its practical weakness is opacity: high-performing models such as ensembles and deep networks often give a score without a reason. In a financial setting a bare score is rarely enough, because analysts, institutions and regulators must understand which factors drove an alert before they act on it. Explainable

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

artificial intelligence (XAI) addresses this need through methods that describe model behaviour in human-understandable terms [1]–[4]. Two widely used post-hoc methods are LIME [1], which fits a local interpretable surrogate around one prediction, and SHAP [2], which attributes a prediction to features using Shapley values.

Existing XAI-in-finance work is substantial. A systematic review of XAI in finance reports that SHAP, LIME and feature importance are the most widely used methods, while stock-market fraud receives comparatively little attention [5]. A study of financial statement fraud combined gradient boosting with SHAP to expose influential features [6]. Neither addresses explainable detection of trading-based fraud directly, which motivates this work.

The contribution of this paper is a coherent, testable framework rather than an implemented system. Following the seminar synopsis, Section II states the problem and objectives, Section III reviews related work, Section IV states the research gap, and Sections V–VIII describe the methodology, architecture, explanation techniques and evaluation plan. Section IX discusses limitations and ethics, and Section X concludes. Fig. 1 previews the proposed architecture, which Section VI describes. Background on learning algorithms is drawn from standard texts [7], [8].

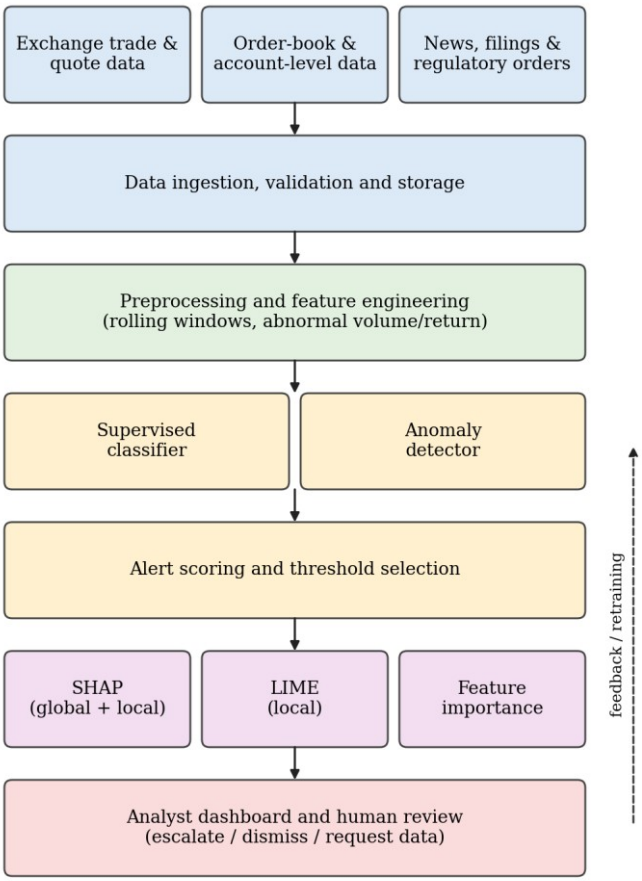

Fig. 1. Proposed explainable AI-based stock market fraud detection architecture (conceptual design; not implemented).

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

II. Problem Statement and Objectives

Modern ML models can detect suspicious trading with high accuracy, yet their decisions are often hard to interpret. Regulators, financial institutions and investors need transparent decisions in order to build trust, meet compliance expectations and take proportionate action. The problem addressed here is therefore how to detect potentially fraudulent stock-market activity with ML while giving analysts a clear, feature-level account of why each activity was flagged.

The objectives, adapted from the synopsis and seminar presentation, are:

- to design an AI-based approach for detecting stock market fraud;
- to apply ML algorithms to identify fraudulent or abnormal trading patterns;
- to improve transparency by explaining predictions with XAI techniques;
- to help analysts and regulators understand the reasons behind each alert;
- to improve the reliability of detection while reducing false alarms; and
- to move toward a trustworthy and interpretable detection system for financial markets.

The aim of reducing false alarms is a hypothesis to be tested in future experiments, not an established property of the proposed framework.

III. Literature Review and Related Work

A. Explainability Foundations

Arrieta et al. present a taxonomy of XAI concepts, approaches, opportunities and challenges under the heading of responsible AI [3], and Guidotti et al. survey methods for explaining black-box models [4]. Both are general surveys and do not address stock-market fraud. Ribeiro et al. introduced LIME, which explains an individual prediction of any classifier by fitting a locally faithful interpretable model; the authors note that such local explanations may not capture the global behaviour of a complex model [1]. Lundberg and Lee unified several attribution methods under SHAP, which measures each feature's contribution using Shapley values; the exact computation can be expensive on large datasets [2].

B. XAI in Finance and Fraud Detection

Černevičienė and Kabašinskas reviewed XAI in finance and found that SHAP, LIME and feature importance dominate, with limited research on stock-market fraud [5]. Fukas et al. applied XGBoost with SHAP to financial fraud and identified the features that influenced predictions, but their focus was financial statement fraud rather than market trading [6]. Table I summarises the surveyed works.

Table I: Summary of Surveyed Literature

Work	Method	Key finding	Limitation
Ribeiro et al. [1]	LIME, model-agnostic local surrogate	Explains single predictions of any classifier	Local view may not reflect global behaviour
Lundberg & Lee [2]	SHAP, Shapley values	Consistent per-feature attributions	Costly on large datasets

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Work	Method	Key finding	Limitation
Arrieta et al. [3]	XAI taxonomy and survey	Maps concepts, opportunities, challenges	General XAI, not stock fraud
Guidotti et al. [4]	Survey of black-box explanation	Classifies explanation methods	General, not finance-specific
Černevičienė & Kabašinskas [5]	Systematic review, XAI in finance	SHAP, LIME, feature importance most used	Little work on stock fraud
Fukas et al. [6]	XGBoost + SHAP	Identifies features behind fraud predictions	Financial statement fraud only

The synopsis and presentation also list reviews by Zafar and Wu (2026), Weber et al. (2024) and Chen et al. (2023). Because their bibliographic details are not given in the supplied reference list and could not be independently verified, they are not cited here. A Literature Insights paragraph in the synopsis refers to XAI in healthcare; it is outside this paper's scope and is not used.

IV. Research Gap and Contributions

Three gaps emerge from the survey and the synopsis. First, most ML fraud-detection research targets banking, credit-card, insurance or financial statement fraud, leaving fraud in stock markets comparatively under-studied [5], [6]. Second, many detection models are black boxes that classify suspicious behaviour without indicating which factors produced the result, which makes it hard for analysts and regulators to verify and act on outputs. Third, few studies integrate SHAP and LIME with stock-market fraud models, and the presentation notes a lack of real-time explainable detection for market manipulation and insider trading.

The proposed contributions are: (i) a layered architecture that couples detection with explanation and human review; (ii) a feature-engineering scheme tied to specific manipulation and insider-trading patterns; (iii) a protocol that compares supervised and anomaly-detection models under realistic temporal validation; (iv) a plan for evaluating the explanations themselves, not only the predictions; and (v) an explicit statement of the experiments still to be conducted. None of these has yet been validated empirically.

V. Proposed Methodology

A. Problem Formulation

Let each observation x be a feature vector describing a security, account or time window, and let y indicate whether the activity was later confirmed as fraudulent. Where reliable labels exist, detection is a supervised binary classification task with severe class imbalance. Where labels are sparse or absent, which is common because confirmed cases are rare, detection becomes an anomaly-detection task in which unusual windows receive a high anomaly score [9]. The framework supports both, and Fig. 2 shows the branching decision.

B. Data Collection

Candidate data include exchange trade and quote histories (price, volume, timestamps), order-book events such as submissions, modifications and cancellations, account-level trade records where lawfully

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

accessible, corporate announcements and news, and, for labels, published regulatory enforcement orders and litigation releases that identify manipulated securities or periods. Availability, licensing, privacy and legal access differ across markets and must be confirmed before data collection; no dataset has yet been selected or obtained.

C. Preprocessing and Cleaning

Preprocessing removes duplicate and erroneous records, aligns timestamps across sources, handles missing values with documented rules, adjusts prices for splits and dividends, and treats extreme values carefully because genuine manipulation can itself appear as an outlier. To prevent information leakage, all scaling and encoding is fitted on training periods only. Class imbalance is handled with class weights, cost-sensitive thresholds or resampling such as SMOTE [10], applied inside training folds only.

D. Feature Engineering

Features are designed around the behaviours the framework aims to expose (Table II). They are computed over rolling windows and, for insider trading, over windows preceding material announcements. These are proposed candidates, not validated indicators.

Table II: Candidate Features for Suspicious Trading Patterns

Pattern	Candidate features	Rationale
Pump-and-dump	Volume vs. rolling median; abnormal return; post-spike price reversal; social-media mention surge	Sharp rise on unusual volume, then reversal
Spoofing / layering	Order-to-trade ratio; cancellation ratio; order-book imbalance; quote lifetime	Large orders placed and cancelled quickly
Wash / circular trading	Self-match ratio; round-trip trade count; counterparty concentration	Trades with little change in beneficial ownership
Insider trading	Cumulative abnormal return and volume in a window before announcements; account trading before events	Unusual positioning before public news

E. Model Selection and Comparison

The supervised branch compares an interpretable baseline (logistic regression), a random forest [11] and gradient-boosted trees such as XGBoost [12]; deep networks [7] are a later option if data volume justifies their opacity and cost. The anomaly branch uses methods such as isolation forest [13] when labels are scarce. All models are tuned on validation data, and the final model is chosen on a trade-off between detection quality, stability and explainability rather than accuracy alone. Fig. 2 summarises the workflow from raw data to a model ready for explanation.

F. Alert Generation and Human Review

Model scores are converted to alerts by a threshold chosen against analyst capacity, since the number of alerts a team can review is a real constraint. Each alert is accompanied by an explanation (Section VII)

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

and routed to an analyst, who may escalate, dismiss or request more data. Analyst decisions can later serve as additional labels, subject to bias checks.

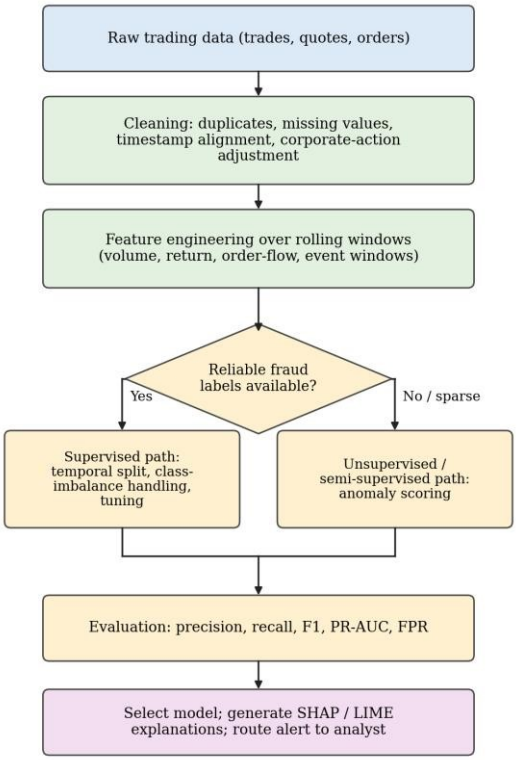

Fig. 2. Data preprocessing and fraud detection workflow of the proposed framework (conceptual design; not implemented).

VI. Proposed System Architecture

Fig. 1 shows the proposed architecture as seven layers. Data sources feed an ingestion and storage layer; preprocessing and feature engineering prepare model inputs; a detection layer holds supervised and anomaly models; alert scoring applies the threshold; an explainability layer produces SHAP, LIME and feature-importance outputs; and an analyst dashboard supports human review. A feedback path returns confirmed outcomes for monitoring and retraining. The design separates the detector from the explainer, so the model can be replaced without redesigning the explanation layer, because both SHAP and LIME can be applied to any trained classifier.

The seminar scope also lists real-time detection as a challenge to consider. The architecture is compatible with near-real-time scoring over streaming windows, but explanation methods add latency, so a practical design would score every window and generate full explanations only for alerts. The achievable latency has not been measured.

VII. Explainable AI Techniques: SHAP and LIME

A. SHAP

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

SHAP assigns each feature a contribution to a prediction, derived from Shapley values in cooperative game theory, so that contributions sum to the difference between the prediction and a baseline expectation [2]. For tree ensembles an efficient exact algorithm exists, and local explanations can be aggregated into global summaries [14]. In this framework SHAP provides two views: a global view of which features drive the model overall, used for model validation, and a local view showing why one alert was raised, for example that a volume spike and a high cancellation ratio pushed the score upward. Its main costs are computation on large data and sensitivity to correlated features, which can split credit among them.

B. LIME

LIME perturbs the instance being explained, queries the model on the perturbed samples, weights them by proximity, and fits a sparse linear model whose coefficients form the explanation [1]. It is model-agnostic and suits any final model, including non-tree models. Its explanations are local and can vary with the sampling and kernel settings, so their stability must be checked.

C. Combined Use

Using both methods allows cross-checking. Fig. 3 shows that the flagged instance and trained model are passed to SHAP and LIME in parallel, and a consistency check compares the top-ranked features and the stability of the explanation under perturbation. Agreement raises confidence in an explanation; disagreement is itself flagged to the analyst rather than hidden. The output is an interpretable alert report containing the risk score, leading features with direction of effect, and the time window concerned.

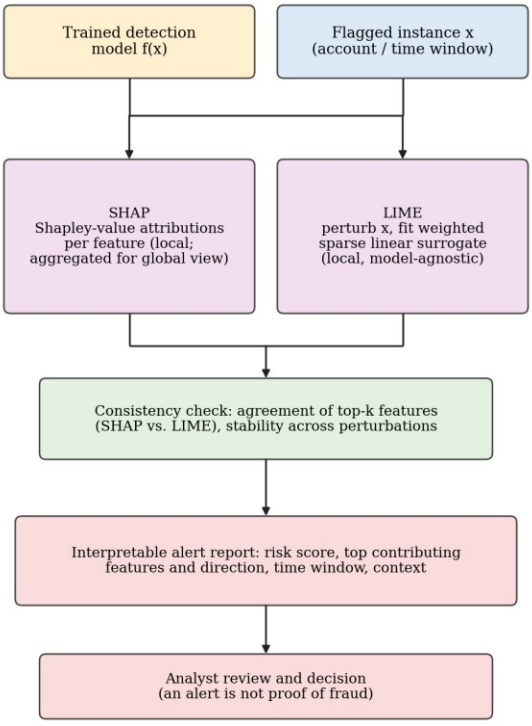

Fig. 3. Integration of machine learning, SHAP and LIME for interpretable fraud alerts (conceptual design; not implemented).

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

An important caution applies to every explanation. SHAP and LIME describe how a model reached its output; they do not establish that the behaviour is fraudulent or that a feature causes fraud. They support analyst judgement and do not replace it.

VIII. Evaluation Methodology and Performance Metrics

A. Validation Design

Because market data are time-ordered, random splits would leak future information. The plan uses chronological splits and walk-forward validation, with hyperparameters tuned on a validation period and a final untouched test period. Where possible, results are repeated across several periods and securities, and reported with confidence intervals so that differences between models are not over-interpreted.

B. Prediction Metrics

Accuracy is misleading when fraud is rare, so evaluation centres on measures sensitive to the minority class (Table III). The precision–recall curve is generally more informative than the ROC curve on imbalanced data [15], so PR-AUC is the primary summary metric, with ROC-AUC secondary.

Table III: Proposed Evaluation Measures

Measure	Definition / purpose	Why it matters
Precision	TP / (TP + FP)	Share of alerts worth investigating
Recall	TP / (TP + FN)	Share of true cases found
F1-score	Harmonic mean of precision and recall	Single balance measure
PR-AUC	Area under precision–recall curve	Primary metric under imbalance
False-positive rate	FP / (FP + TN)	Burden of false alarms
Precision@k	Precision among the k highest-scored alerts	Fits fixed analyst capacity
Detection delay	Time from onset of activity to alert	Relevance to real-time use

C. Explanation Quality

Explanations are evaluated separately from predictions. Proposed checks are: fidelity (does removing the features named as important actually lower the score); stability (do similar instances and repeated runs yield similar explanations); agreement between SHAP and LIME rankings; plausibility against known manipulation patterns; and computation time. A small structured review with domain experts, if access can be arranged, would assess whether the explanations help analysts decide faster or more accurately.

D. Experiments to Be Conducted

The following remain future work and no outcome is assumed: (E1) build and label a dataset from the sources in Section V-B; (E2) compare the baseline, random forest, XGBoost and anomaly models on the metrics above; (E3) test whether threshold selection and explanation-assisted triage change false-positive burden relative to score-only triage; (E4) measure SHAP–LIME agreement and stability; and

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

(E5) measure scoring and explanation latency. Only after E1–E5 can claims about accuracy, false-alarm reduction or real-time suitability be made.

IX. Limitations and Ethical Considerations

Label quality. Confirmed fraud cases are rare and biased toward cases that regulators detected and pursued, so a model may learn the detection process rather than fraud itself; undiscovered fraud appears as legitimate. **Drift and adaptation.** Market regimes change, and manipulators may alter tactics once patterns are known, so models need monitoring and retraining. **Explanation limits.** SHAP and LIME are approximations; they can be unstable, are affected by correlated features, and do not imply causation. **Cost.** SHAP can be computationally heavy on large datasets [2], and real-time use requires engineering not yet assessed.

Ethics and governance. A high score is an alert for investigation, not proof of wrongdoing; wrongly flagging a legitimate trader can cause reputational and financial harm. Final decisions must rest with trained humans, with documented review, the ability to contest outcomes and audit trails. Account-level data raise privacy and data-protection obligations, and models should be checked for systematic bias, for example against small-cap securities or particular investor groups. Explanations should be presented without implying certainty.

X. Conclusion and Future Work

This paper proposed an explainable ML framework for stock-market fraud detection that combines supervised and anomaly-based detection with SHAP and LIME explanations and a human-review stage. The work completed so far, as recorded in the seminar presentation, comprises topic finalisation, problem and objective definition, a literature survey, and identification of research gaps. The framework, architecture and evaluation plan presented here are a design proposal, and no dataset, implementation or experimental result exists yet.

Future work follows experiments E1–E5 in Section VIII-D: assembling and labelling a dataset, implementing and comparing the models, evaluating the explanations, and measuring latency. Further directions include studying explanation methods suited to sequential and order-book data, and evaluating the framework with domain experts. Only these steps can show whether explainability improves analyst trust and reduces unnecessary alerts in practice.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.

[3] A. B. Arrieta *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

[4] R. Guidotti *et al.*, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, Art. 93, pp. 1–42, 2018, doi: 10.1145/3236009.

[5] J. Černevičienė and A. Kabašinskas, "Explainable artificial intelligence in finance: A systematic literature review," *Artif. Intell. Rev.*, 2024. [Volume, article number and DOI to be verified.]

[6] P. Fukas *et al.*, "Explainable artificial intelligence for financial fraud detection," in *Proc. Int. Conf. Artificial Intelligence Applications and Innovations*, 2022. [Cited as given in the synopsis; exact title, venue and pages to be verified.]

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[9] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, Art. 15, 2009, doi: 10.1145/1541880.1541882.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[11] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[13] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining (ICDM)*, 2008, pp. 413–422.

[14] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020.

[15] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, Art. e0118432, 2015.

Source note: [1]–[8] come from the reference list in the uploaded synopsis (DOIs and page numbers added where known). [9]–[15] were added from general scholarly knowledge to support methodological statements; they could not be checked against an online database in this environment and should be verified before submission.