



GJAST

GLOBAL JOURNAL OF ADVANCED
SCIENCE AND TECHNOLOGY

INAUGURAL ISSUE

Volume 1 • Issue 1

October – December 2026

Research in science and technology, peer reviewed and free to read

Ph

Physics

Ch

Chemistry

Ls

Life sciences

Bt

Biotech

En

Engineering

Nt

Nanotech

Cs

Computing

Ma

Mathematics

Peer Reviewed • Open Access • Quarterly • Online

Published by Imaginative Ink Publications, Solapur, Maharashtra, India

iipublications.com/iipgjast



Global Journal of Advanced Science and Technology

JOURNAL INFORMATION

Journal	Global Journal of Advanced Science and Technology (GJAST)
Volume and issue	Volume 1, Issue 1 (October–December 2026)
Frequency	Quarterly (four issues a year)
Format	Online
e-ISSN	____-____
Language	Multilingual
Subject	Science and Technology
Review policy	Peer reviewed
Access	Open access under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)
Publisher	Imaginative Ink Publications, 37 Shridevi Nagar, Near Akashwani Kendra, Solapur 413006, Maharashtra, India
E-mail	imaginativeinkpublications@gmail.com
Phone	+91 87881 40066
Website	https://iipublications.com/iipgjast

Copyright and disclaimer

© 2026 The Authors. Articles in this issue are published under the CC BY-NC 4.0 license, which permits non-commercial use, distribution and reproduction in any medium, provided the original work is properly cited. The views expressed in the articles are those of the authors and do not necessarily reflect the views of the editors or the publisher.



Global Journal of Advanced Science and Technology

EDITORIAL BOARD

EDITOR-IN-CHIEF

Dr. Mohan Kalidas Kale

Assistant Professor, Department of Geography, Sangameshwar Night College (Sangameshwar Ratra Mahavidyalaya), Solapur, Maharashtra, India

EDITORIAL BOARD MEMBERS

Prof. Dr. S. S. Maske

Professor, Department of Geography, Sangameshwar College (Autonomous), Solapur, Maharashtra, India

Dr. Rahul M. Khobragade

Professor and Head, Department of Microbiology, Dr. Babasaheb Ambedkar Marathwada University, Sub-Campus Dharashiv, Maharashtra, India

Dr. Pandurang Choudhari

Associate Professor, Department of Geography, University of Mumbai, Mumbai, Maharashtra, India

Dr. S. S. Karande

Assistant Professor, Department of Physics, Sangameshwar College (Autonomous), Solapur, Maharashtra, India

Dr. Mrunalini Nilesh Desai

Assistant Professor, Department of Botany, Vivekanand College (An Empowered Autonomous Institute), Kolhapur, Maharashtra, India

Dr. Sneha Sampatrao Desai

Assistant Professor, Department of Zoology, Vivekanand College (An Empowered Autonomous Institute), Kolhapur, Maharashtra, India

Mr. Chandana Ruwan Kumara

Assistant Professor, Department of Fine Arts, University of Kelaniya, Sri Lanka

Dr. Phrapalad Soravit Aphipanyo

Assistant Professor, Department of Education Psychology and Guidance, Mahachulalongkornrajavidyalaya University, Thailand



Global Journal of Advanced Science and Technology

PREFACE

The Global Journal of Advanced Science and Technology (GJAST) is a peer-reviewed, open access journal published quarterly by Imaginative Ink Publications, Solapur, Maharashtra, India. It was founded to give researchers, teachers and students, particularly those working in colleges and universities across India, a carefully edited and openly available place to publish work in science and technology.

The journal publishes original research articles, review articles, short communications and book reviews across the physical, chemical, biological and environmental sciences, engineering, computer science and information technology, mathematics and statistics, and their applications. Interdisciplinary work that brings these fields together, or applies them to problems of society and the economy, is particularly welcome. Submissions are accepted in more than one language, provided they carry an abstract in English.

Every manuscript is checked by the editors for scope, originality and presentation, including a plagiarism check, and is then reviewed by subject experts before a decision is made. Where a member of the editorial board is an author, the manuscript is handled by another editor. Authors are asked to declare any conflicts of interest, sources of funding and the use of AI tools in preparing their work. Articles are free to read from the day they are published and are licensed under the Creative Commons Attribution-NonCommercial 4.0 International License.

Our aim is a journal that researchers can trust and readers can use: clearly written articles, fair and timely review, and complete, accurate information about every paper. We are grateful to our authors for choosing GJAST, to the members of the editorial board and our reviewers for their time and judgement, and to readers for their interest. We welcome submissions for forthcoming issues and invite suggestions that will help the journal serve its community better.

Dr. Mohan Kalidas Kale

Editor-in-Chief

Global Journal of Advanced Science and Technology



Global Journal of Advanced Science and Technology

CONTENTS

REVIEW ARTICLES	PAGE
Preface.....	iv
Machine Learning Methods for Credit Risk Assessment in Microfinance: A Review.....	1
<i>Parmeshwar Madhukar Patil, Rohini Rahul Anyapanawar</i>	
Data-Driven Detection of Online Investment Fraud: A Review of Statistical and Machine Learning Approaches.....	7
<i>Rohini Rahul Anyapanawar, Parmeshwar Madhukar Patil</i>	
Remote Sensing and GIS in Urban Land Use Change Studies: A Review of Methods and Indian Applications.....	12
<i>Mohan Kalidas Kale</i>	
Phishing Detection Techniques: A Review of Rule-Based, Machine Learning and Deep Learning Approaches.....	17
<i>B. S. Bukka</i>	
Sample Size Determination in Applied Research: A Critical Review of Cochran's, Yamane's and Power-Analysis Approaches.....	22
<i>Sarika B. Patil</i>	

REVIEW ARTICLE

Machine Learning Methods for Credit Risk Assessment in Microfinance: A Review

Parmeshwar Madhukar Patil^{a,*}, Rohini Rahul Anyapanawar^b^a Assistant Professor, Department of Statistics, Sangameshwar College (Autonomous), Solapur, Maharashtra, India^b Assistant Professor, Department of Management and Computer Science, Sangameshwar College (Autonomous), Solapur, Maharashtra, India

* Corresponding author

OPEN ACCESS

This article is licensed under CC BY-NC 4.0.

**ARTICLE HISTORY**

Received: 18 Sep 2026

Accepted: 01 Oct 2026

Published: 14 Oct 2026

CORRESPONDING AUTHOR

Parmeshwar Madhukar Patil, Department of Statistics, Sangameshwar College (Autonomous), Solapur, Maharashtra, India.

HOW TO CITEPatil, P. M., & Anyapanawar, R. R. (2026). Machine Learning Methods for Credit Risk Assessment in Microfinance: A Review. *Global Journal of Advanced Science and Technology*, 1(1), 1–6.

© 2026 The Author(s).

ABSTRACT

Microfinance institutions lend small amounts to borrowers who usually have no formal credit history, and they have long relied on group guarantees and the judgement of loan officers to manage the risk of default. As loan books grow and more data is captured digitally, statistical and machine learning models are increasingly used to support these decisions. This review examines the methods that have been applied to credit risk assessment, from logistic regression and discriminant analysis to decision trees, random forests, gradient boosting and neural networks, and considers how well they suit the conditions of microfinance. It also discusses three practical issues that often decide whether a model is useful: the imbalance between repaid and defaulted loans, the choice of evaluation measures, and the need to explain decisions to borrowers, loan officers and regulators. The literature shows that ensemble methods usually rank borrowers more accurately than single models, but the gains are often modest, and they come at a cost in transparency. For Indian microfinance and self-help group lending, the review argues that a well-specified logistic regression should remain the baseline, that tree ensembles are worth adopting where data volumes allow, and that explanation tools and careful validation matter as much as the choice of algorithm. Gaps include the scarcity of public microfinance datasets, limited use of group-level information, and few studies from India.

Keywords: Credit scoring; Microfinance; Machine learning; Random forest; Logistic regression; Self-help groups

1. Introduction

Microfinance grew out of a simple observation: poor households can be reliable borrowers even when they have no collateral and no credit record. Group lending, regular repayment meetings and close contact between loan officers and borrowers allowed early microfinance programmes to reach people whom

banks would not serve (Morduch, 1999). In India the same idea took the form of self-help groups (SHGs) linked to banks and of joint liability groups served by microfinance institutions.

These arrangements work well when lending is small and local, but they are expensive to scale. A loan officer can know a few hundred borrowers; a lender with several lakh accounts cannot rely on personal knowledge alone. Lenders therefore turn to credit scoring, the use of a statistical model to estimate the probability that a borrower will default from information available at the time of application or during the life of the loan (Thomas, 2000).

Credit scoring is well established in consumer and card lending, and the methods used there have been reviewed many times (Hand & Henley, 1997; Thomas, 2000). Its use in microfinance is newer and raises different questions. Data are thinner, borrowers are more homogeneous, repayment depends on group dynamics as well as on the individual, and the cost of a wrong rejection falls on households with few alternatives. Machine learning methods promise better predictions, but they also make decisions harder to explain.

This review brings together the main strands of this literature. It has three aims: to describe the statistical and machine learning methods that have been used for credit risk assessment; to examine the practical issues of data imbalance, evaluation and interpretability; and to identify what this means for microfinance and SHG lending in India.

2. Scope and approach of the review

The review is narrative rather than systematic. We concentrated on three kinds of work: benchmark studies that compare many classifiers on credit data, studies that build scoring models specifically for microfinance institutions, and methodological papers on techniques that are now standard in the field, such as random forests, gradient boosting, synthetic oversampling and model explanation. Studies were identified through Google Scholar, Scopus and the reference lists of the key papers. We give more weight to studies that report how models were validated, because accuracy figures without a clear validation design are difficult to compare.

3. Credit risk in microfinance

In conventional lending, a scorecard draws on a rich set of variables: income, employment history, past borrowing and repayment behaviour recorded by a credit bureau. Microfinance borrowers often have none of these on paper. The information available to the lender usually comprises basic demographics, the type and size of the household enterprise, the loan amount and purpose, membership of a group, and, for repeat borrowers, their repayment record with the same institution.

Two features of microfinance make risk assessment different. First, repayment is often a group outcome. Under joint liability, members cover for one another, so an individual default may be absorbed by the group and never appear in the data, while a shock to the group can lead several members to default together. Second, the lender's own staff hold a great deal of soft information about the borrower that never enters a database. Van Gool et al. (2012) examined whether credit scoring adds value in this setting and concluded that it can, provided the model complements rather than replaces the loan officer's judgement.

4. Methods for credit risk assessment

4.1 Statistical models

Linear discriminant analysis and logistic regression were the first methods used for credit scoring and remain the most common in practice (Hand & Henley, 1997). Logistic regression models the log-odds of default as a linear function of the borrower's characteristics. Its coefficients can be read directly, it produces probabilities that can be calibrated to observed default rates, and it can be converted into a

points-based scorecard that a branch manager can use without a computer. Its main limitation is that it captures non-linear effects and interactions only if the analyst builds them in, for example by grouping continuous variables into bands.

4.2 Decision trees and ensembles

A decision tree splits borrowers into groups using simple rules, such as loan amount above a threshold or fewer than two previous loans. Trees handle non-linearity and missing values naturally and are easy to explain, but a single tree is unstable: small changes in the data can produce a very different tree. Ensemble methods address this by combining many trees. A random forest grows each tree on a bootstrap sample of the data and a random subset of variables, and averages their predictions (Breiman, 2001). Gradient boosting builds trees one after another, each correcting the errors of the previous ones; XGBoost is an efficient and widely used implementation (Chen & Guestrin, 2016).

4.3 Neural networks and support vector machines

Neural networks and support vector machines can approximate complex relationships between borrower characteristics and default. In microfinance, Blanco et al. (2013) built multilayer perceptron models for a Peruvian microfinance institution and found that they classified borrowers more accurately, and at a lower cost of misclassification, than linear and quadratic discriminant analysis and logistic regression. Such models need more data and more careful tuning than logistic regression, and their predictions are harder to explain.

4.4 What benchmark studies show

Two large benchmark studies give the clearest picture of how these methods compare. Baesens et al. (2003) tested a wide range of classifiers on eight credit datasets and found that support vector machines and neural networks performed well, but that simple methods such as logistic regression and linear discriminant analysis were often close behind. More than a decade later, Lessmann et al. (2015) repeated the exercise with 41 classifiers and newer methods. They found that ensembles, especially heterogeneous ensembles that combine different types of model, gave the best results, and they recommended that random forest replace logistic regression as the standard benchmark for new studies.

Two cautions apply. The differences between the best methods are often small in absolute terms, and most benchmark datasets come from consumer lending in Europe and North America. Whether the same ranking holds for small, homogeneous microfinance portfolios is still largely untested. Table 1 summarises the strengths and limitations of the main methods.

Table 1. Main methods for credit risk assessment and their suitability for microfinance.

Method	Strengths	Limitations	Interpretability
Logistic regression	Stable; calibrated probabilities; easy scorecard	Misses interactions unless built in	High
Decision tree	Simple rules; handles missing data	Unstable; tends to overfit	High
Random forest	Accurate; robust to noise; little tuning	Many trees; less transparent	Medium (with SHAP)
Gradient boosting (XGBoost)	Often most accurate on tabular data	Needs careful tuning; can overfit small data	Medium (with SHAP)
Neural network	Captures complex patterns	Needs large data; hard to explain	Low
Support vector machine	Good with many variables	No direct probabilities; slow on large data	Low

Source: compiled by the authors from Hand and Henley (1997), Baesens et al. (2003), Lessmann et al. (2015) and Blanco et al. (2013).

5. Data, imbalance and evaluation

5.1 Class imbalance

Well-run microfinance portfolios have low default rates, often below five per cent. A model trained on such data can reach high accuracy simply by predicting that everyone repays. Common remedies include undersampling the repaid loans, oversampling the defaulted ones, adjusting class weights, or generating synthetic minority cases with SMOTE (Chawla et al., 2002). These techniques should be applied only to the training data; applying them before splitting the data inflates the reported performance.

5.2 Evaluation measures

Because accuracy is misleading on imbalanced data, credit models are judged mainly by how well they rank borrowers and how well their probabilities match observed default rates. The area under the ROC curve (AUC) and the Kolmogorov–Smirnov statistic measure ranking; the Brier score measures the quality of the predicted probabilities. Where the costs of approving a defaulter and rejecting a good borrower are known, a cost-based or profit-based measure is more informative than any statistical score (Lessmann et al., 2015). Table 2 lists the measures most often reported.

Table 2. Evaluation measures commonly used for credit scoring models.

Measure	What it tells the lender	Note
AUC (ROC curve)	How well the model ranks defaulters above good borrowers	0.5 = random; 1 = perfect
Kolmogorov–Smirnov	Maximum separation between good and bad score distributions	Common in industry practice
Brier score	Accuracy of predicted default probabilities	Lower is better
Sensitivity and specificity	Share of defaulters and good borrowers correctly identified	Depend on the cut-off chosen
Expected cost or profit	Financial effect of using the model at a given cut-off	Needs cost estimates

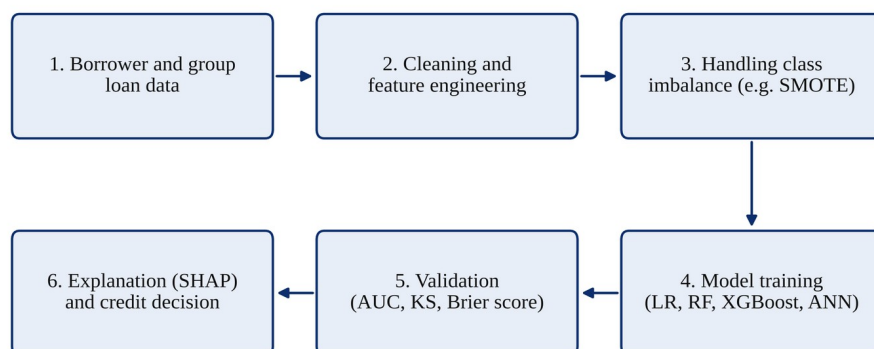


Figure 1. Typical workflow for building a credit scoring model for a microfinance portfolio.

Figure 1 sets these steps in order. In practice the validation step is where many published studies are weakest. A model should be tested on loans that were not used to build it, preferably loans issued later in time, because borrower behaviour and lending policy change. A model that performs well on a random hold-out sample may still fail when applied to next year's applicants.

6. Interpretability and fairness

A borrower who is refused a loan is entitled to know why, and a loan officer will not trust a score that contradicts what they know about a client without some explanation. This has made interpretability a central issue. Logistic regression and short decision trees explain themselves. For ensembles and neural networks, post-hoc methods are needed; SHAP values, which attribute each prediction to the contribution of each variable, are now the most widely used (Lundberg & Lee, 2017). SHAP can show, for example, that a particular applicant's score was lowered mainly by a large loan relative to the group's past borrowing.

Fairness is a related concern. Variables that predict default, such as location or household type, may also act as proxies for caste, religion or gender. A model that is accurate on average can still treat some groups unfairly. Checking performance and approval rates across groups, and excluding variables that cannot be justified, should be part of model development rather than an afterthought.

7. Implications for Indian microfinance and SHG lending

Four points follow from the literature. First, logistic regression remains a sound starting point. It is transparent, it performs close to more complex models in many settings, and it gives a benchmark against which any more complex model must justify itself. Second, where lenders have tens of thousands of loans and a reasonable number of defaults, random forests and gradient boosting are worth testing, provided they are validated on later loans and accompanied by explanations. Third, group-level variables, such as the group's age, its repayment history, attendance at meetings and the share of members with outstanding loans elsewhere, are likely to carry much of the predictive information in SHG and joint-liability lending, yet they are rarely used in published models. Fourth, scoring should support the loan officer rather than replace them: the most useful systems flag cases for closer review rather than issue automatic rejections.

8. Research gaps

- There are very few public microfinance datasets, which makes it hard to compare methods or reproduce results.
- Studies from India, and especially from SHG–bank linkage lending, are scarce compared with studies from Latin America and Europe.
- Group dynamics and joint liability are seldom modelled explicitly, for example with multilevel or network models.
- Most studies evaluate models statistically rather than by their financial effect on the lender or the borrower.
- Fairness across social groups is rarely tested in microfinance scoring models.

9. Conclusion

Machine learning has improved the ability of lenders to rank borrowers by risk, and ensemble methods such as random forests and gradient boosting now set the standard in benchmark studies. In microfinance, however, the gains from complex models are often small, data are limited, and the need for explanation is strong. The method matters less than the care with which a model is built, validated and used. For Indian microfinance institutions and banks lending to SHGs, the most productive path is likely to combine a transparent baseline model, well-validated ensembles where data allow, explanation tools such as SHAP, and the knowledge of field staff. Building shared, anonymised datasets would do more for research in this area than any new algorithm.

Author contributions: P. M. Patil: conceptualisation, methodology of the review, writing – review and editing. R. R. Anyapanawar: literature search, writing – original draft, preparation of tables.

Conflict of interest: The authors declare no conflict of interest.

Funding: This research received no external funding.

Declaration of generative AI use: During the preparation of this work the authors used an AI writing assistant (Claude, Anthropic) to help structure and draft the text. The authors reviewed, edited and verified the content and take full responsibility for this article.

References

- Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6), 627–635. <https://doi.org/10.1057/palgrave.jors.2601545>
- Blanco, A., Pino-Mejías, R., Lara, J., & Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Systems with Applications*, 40(1), 356–364. <https://doi.org/10.1016/j.eswa.2012.07.051>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A*, 160(3), 523–541. <https://doi.org/10.1111/j.1467-985X.1997.00078.x>
- Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136. <https://doi.org/10.1016/j.ejor.2015.05.013>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).
- Morduch, J. (1999). The microfinance promise. *Journal of Economic Literature*, 37(4), 1569–1614. <https://doi.org/10.1257/jel.37.4.1569>
- Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2), 149–172. [https://doi.org/10.1016/S0169-2070\(00\)00034-0](https://doi.org/10.1016/S0169-2070(00)00034-0)
- Van Gool, J., Verbeke, W., Sercu, P., & Baesens, B. (2012). Credit scoring for microfinance: Is it worth it? *International Journal of Finance & Economics*, 17(2), 103–123. <https://doi.org/10.1002/ijfe.444>

REVIEW ARTICLE

Data-Driven Detection of Online Investment Fraud: A Review of Statistical and Machine Learning Approaches

Rohini Rahul Anyapanawar^{a,*}, Parmeshwar Madhukar Patil^b^a Assistant Professor, Department of Management and Computer Science, Sangameshwar College (Autonomous), Solapur, Maharashtra, India^b Assistant Professor, Department of Statistics, Sangameshwar College (Autonomous), Solapur, Maharashtra, India

* Corresponding author

OPEN ACCESS

This article is licensed under CC BY-NC 4.0.

**ARTICLE HISTORY**

Received: 18 Sep 2026

Accepted: 04 Oct 2026

Published: 14 Oct 2026

CORRESPONDING AUTHOR

Rohini Rahul Anyapanawar, Department of Management and Computer Science, Sangameshwar College (Autonomous), Solapur, Maharashtra, India.

HOW TO CITEAnyapanawar, R. R., & Patil, P. M. (2026). Data-Driven Detection of Online Investment Fraud: A Review of Statistical and Machine Learning Approaches. *Global Journal of Advanced Science and Technology*, 1(1), 7–11.

© 2026 The Author(s).

ABSTRACT

Online investment fraud, in which victims are persuaded through websites, messaging groups or trading applications to invest in schemes that do not exist, has become one of the most damaging forms of cyber-enabled financial crime in India. Unlike card fraud, where the bank sees every transaction, investment fraud is usually carried out by the victim's own authorised payments, which makes it hard to detect with conventional controls. This review examines the statistical and machine learning approaches that have been used to detect financial fraud and considers how they apply to investment fraud. It covers rule-based systems, statistical models, supervised classifiers such as random forests and gradient boosting, unsupervised anomaly detection, and the analysis of text and payment networks. It then discusses the problems that limit these methods in practice: extreme class imbalance, changing fraud patterns, delayed and incomplete labels, and the difficulty of detecting fraud that the customer has been manipulated into authorising. The literature suggests that no single method is sufficient. Effective detection combines supervised models trained on confirmed cases, anomaly detection for new schemes, network analysis to trace mule accounts, and text analysis of advertisements and messages, all supported by human review. The review closes with priorities for research in India, including shared anonymised datasets drawn from cybercrime complaints and closer study of victims' behaviour.

Keywords: Online investment fraud; Fraud detection; Machine learning; Anomaly detection; Cybercrime; Financial technology

1. Introduction

Few crimes have grown as quickly with digital finance as investment fraud. The pattern is familiar to anyone who has read cybercrime complaints in India: a stranger adds the victim to a messaging group, shares screenshots of large trading profits, and invites them to an application or website that shows their

investment growing day by day. When the victim tries to withdraw, they are asked for fees or taxes, and eventually the contact disappears. The money has long since moved through a chain of accounts.

Fraud detection has a long history in statistics and computing. Bolton and Hand (2002) reviewed the statistical methods used against credit card, telecommunications and money-laundering fraud, and later surveys extended this work to data mining and machine learning (Ngai et al., 2011; Phua et al., 2010; West & Bhattacharya, 2016). Most of this research, however, concerns fraud committed against an account holder without their knowledge. Investment fraud is different: the victim usually makes the payments themselves, believing them to be genuine.

This review asks how well existing detection methods transfer to this setting. It describes the main families of methods, examines the practical problems that limit them, and identifies research priorities, with particular attention to India.

2. The nature of online investment fraud

Online investment fraud takes several forms: fake trading or cryptocurrency platforms, Ponzi-style schemes that pay early investors from later deposits, pump-and-dump promotions of thinly traded shares, and impersonation of real brokers or advisers. What these schemes share is persuasion. They exploit the appeal of quick gains and the reluctance to accept a loss, tendencies that prospect theory describes in general terms (Kahneman & Tversky, 1979). DeLiema et al. (2020) studied the mindsets and behaviours of investment fraud victims in the United States and found that exposure to sales pitches and willingness to consider risky opportunities distinguished victims from other investors.

From a detection point of view, three features matter. The payments are authorised by the victim, so they pass the authentication checks that stop most card fraud. The money is usually routed through accounts opened or rented for the purpose, often called mule accounts. And the scheme leaves traces outside the banking system, in advertisements, messaging groups, app stores and websites, that can be analysed before any money is lost.

3. Approaches to detection

Figure 1 groups the methods discussed in the literature into five families. In practice they are combined rather than used alone.

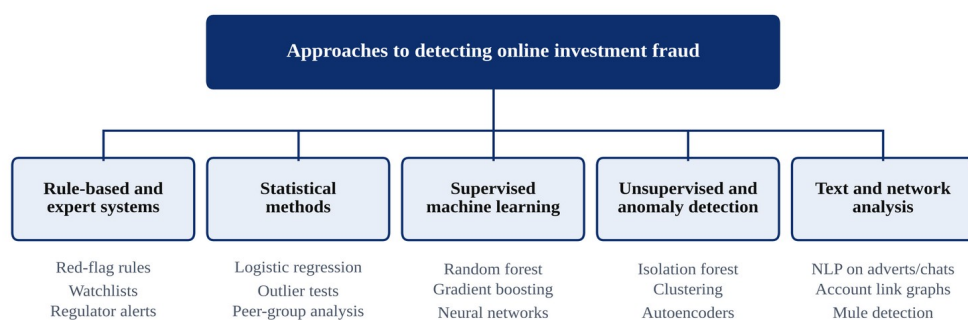


Figure 1. Families of approaches used to detect online investment fraud.

3.1 Rule-based and expert systems

The oldest approach encodes the knowledge of investigators as rules: a new account that receives many payments from unrelated individuals within a short time, transfers to beneficiaries named in regulator alerts, or payments described as trading fees. Rules are transparent and can be put in place quickly when a

new scheme is reported. Their weakness is that fraudsters learn them, and a large rule set produces many false alarms that investigators cannot follow up (Bolton & Hand, 2002).

3.2 Statistical methods

Statistical methods model what normal behaviour looks like and flag departures from it. Bolton and Hand (2002) described peer-group analysis, which compares an account with others that behaved similarly in the past, and break-point analysis, which looks for sudden changes in an account's own behaviour. Logistic regression remains widely used when confirmed fraud cases are available, because its coefficients show which factors raise the risk.

3.3 Supervised machine learning

When enough confirmed cases exist, supervised classifiers can learn the patterns that separate fraudulent from genuine activity. In their review of financial fraud detection, Ngai et al. (2011) found that logistic models, neural networks, Bayesian networks and decision trees were the most frequently used techniques. Since then, ensemble methods such as random forests and gradient boosting have become the usual choice for tabular transaction data. West and Bhattacharya (2016) note, however, that reported accuracies are hard to compare because studies use different data, different measures and often no common benchmark.

3.4 Unsupervised and anomaly detection

New schemes appear faster than labelled examples can be collected. Anomaly detection methods address this by learning what normal activity looks like and scoring how unusual each new case is, without needing fraud labels (Chandola et al., 2009). Isolation forest, which isolates unusual points with fewer random splits than normal ones, is efficient on large data and is now common in practice (Liu et al., 2008). Clustering and autoencoders are also used. The cost is a higher false-alarm rate, because unusual is not the same as fraudulent.

3.5 Text and network analysis

Investment fraud is advertised before it is paid for. Natural language processing can screen advertisements, social media posts and messages for the promises typical of fraud, such as guaranteed daily returns or pressure to act immediately. Network analysis takes the other end of the chain: by linking accounts that send money to the same beneficiaries, share devices or are opened with the same documents, it can reveal rings of mule accounts that individual transaction checks miss (Abdallah et al., 2016). Table 1 compares the five families.

Table 1. Comparison of approaches to detecting online investment fraud.

Approach	Data needed	Main strength	Main limitation
Rules and expert systems	Investigator knowledge; alerts	Transparent; quick to update	Easy to evade; many false alarms
Statistical models	Account history; some labels	Interpretable risk factors	Limited to patterns specified
Supervised machine learning	Large set of confirmed cases	High accuracy on known schemes	Weak on new schemes; needs labels
Anomaly detection	Normal activity only	Can flag new schemes	Unusual is not always fraud
Text and network analysis	Adverts, messages; payment links	Detects schemes and mule rings early	Needs data from outside the bank

Source: compiled by the authors from Bolton and Hand (2002), Ngai et al. (2011), West and Bhattacharya (2016) and Chandola et al. (2009).

4. Problems that limit detection in practice

4.1 Extreme class imbalance

Fraud is rare compared with genuine activity. A model can therefore appear accurate while missing most fraud, and evaluation must rely on measures such as precision, recall and the area under the precision–recall curve. Resampling methods and cost-sensitive learning help, but they do not remove the problem.

4.2 Changing patterns and delayed labels

Dal Pozzolo et al. (2014), drawing on a large card fraud system, showed that fraud patterns drift over time, that the true label of a transaction is often known only weeks later, and that the investigators' own choices determine which cases are labelled at all. Each of these problems is worse for investment fraud, where victims may report the loss months after the first payment, and many never report it at all.

4.3 Authorised payments

The hardest problem is that the victim wants the payment to go through. Even a correct alert may be overridden by a customer who believes the investment is genuine. Detection therefore has to be paired with intervention, such as a call from the bank or a cooling-off period for first payments to a new beneficiary, and with public awareness. Cross (2015) has shown how the blame directed at victims of online fraud discourages many of them from reporting, which in turn starves detection systems of labels. Table 2 summarises these problems and the responses suggested in the literature.

Table 2. Practical problems in investment fraud detection and suggested responses.

Problem	Effect on detection	Suggested response
Class imbalance	High accuracy but low recall	Precision–recall measures; cost-sensitive learning
Concept drift	Models age quickly	Regular retraining; drift monitoring
Delayed or missing labels	Training data incomplete and biased	Semi-supervised learning; link to complaint data
Authorised payments	Alerts overridden by victims	Human contact; delays on new beneficiaries
Fragmented data	Each bank sees only part of the chain	Secure data sharing between banks and police

5. Priorities for research in India

India has built an unusually rich source of data on cybercrime through the National Cyber Crime Reporting Portal and the 1930 helpline, which record complaints, payment details and the accounts involved. Little of this has reached researchers. We see four priorities.

1. Anonymised datasets drawn from cybercrime complaints, released under clear safeguards, would allow methods to be tested on Indian investment fraud rather than on card fraud from other countries.
2. Network methods that trace mule accounts across banks deserve more attention than further improvements to single-transaction classifiers.
3. Text classifiers for advertisements and messages need to work in Indian languages and in mixed Hindi–English and Marathi–English text, which most current tools do not handle well.
4. Detection research should be linked with studies of victims' behaviour, so that alerts and interventions are designed around how people actually make investment decisions.

6. Conclusion

The methods developed for card and payment fraud provide a strong base for detecting online investment fraud, but they cannot be transferred unchanged. Because victims authorise their own payments and report late, transaction-level classifiers see only part of the picture. The literature points towards layered systems that combine supervised models for known schemes, anomaly detection for new ones, network analysis for mule accounts and text analysis for the advertisements that start the process, with human review at the point of payment. For India, the most valuable next step is not a new algorithm but access to well-governed data from the complaints that victims are already filing.

Author contributions: R. R. Anyapanawar: conceptualisation, literature search, writing – original draft. P. M. Patil: methodology of the review, writing – review and editing.

Conflict of interest: The authors declare no conflict of interest.

Funding: This research received no external funding.

Declaration of generative AI use: During the preparation of this work the authors used an AI writing assistant (Claude, Anthropic) to help structure and draft the text. The authors reviewed, edited and verified the content and take full responsibility for this article.

References

- Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications*, 68, 90–113. <https://doi.org/10.1016/j.jnca.2016.04.007>
- Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, 17(3), 235–255. <https://doi.org/10.1214/ss/1042727940>
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), Article 15. <https://doi.org/10.1145/1541880.1541882>
- Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41(10), 4915–4928. <https://doi.org/10.1016/j.eswa.2014.02.026>
- Cross, C. (2015). No laughing matter: Blaming the victim of online fraud. *International Review of Victimology*, 21(2), 187–204. <https://doi.org/10.1177/0269758015571471>
- DeLiema, M., Shadel, D., & Pak, K. (2020). Profiling victims of investment fraud: Mindsets and risky behaviors. *Journal of Consumer Research*, 46(5), 904–914. <https://doi.org/10.1093/jcr/ucz020>
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291. <https://doi.org/10.2307/1914185>
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining* (pp. 413–422). <https://doi.org/10.1109/ICDM.2008.17>
- Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559–569. <https://doi.org/10.1016/j.dss.2010.08.006>
- Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv*. <https://doi.org/10.48550/arXiv.1009.6119>
- West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, 57, 47–66. <https://doi.org/10.1016/j.cose.2015.09.005>

REVIEW ARTICLE

Remote Sensing and GIS in Urban Land Use Change Studies: A Review of Methods and Indian Applications

Dr. Mohan Kalidas Kale*

*Assistant Professor, Department of Geography, Sangameshwar Ratra Mahavidyalaya, Solapur, Maharashtra, India*** Corresponding author***OPEN ACCESS**

This article is licensed under CC BY-NC 4.0.

**ARTICLE HISTORY**

Received: 21 Sep 2026

Accepted: 04 Oct 2026

Published: 14 Oct 2026

CORRESPONDING AUTHOR

Dr. Mohan Kalidas Kale, Department of Geography, Sangameshwar Ratra Mahavidyalaya, Solapur, Maharashtra, India.

HOW TO CITE

Kale, M. K. (2026). Remote Sensing and GIS in Urban Land Use Change Studies: A Review of Methods and Indian Applications. *Global Journal of Advanced Science and Technology*, 1(1), 12–16.

© 2026 The Author(s).

ABSTRACT

Indian cities have expanded rapidly over the last four decades, often faster than planning records can follow. Satellite remote sensing, combined with geographic information systems (GIS), offers a consistent way to map how land is used and how it changes over time. This review examines the methods used in urban land use and land cover (LULC) change studies and the way they have been applied in India. It covers the main sources of free satellite imagery, image pre-processing, classification methods from maximum likelihood to random forests, spectral indices for built-up areas, vegetation and water, accuracy assessment, change detection, and measures of urban sprawl such as Shannon's entropy. Indian studies have shown the value of these methods for cities such as Kolkata and Mangaluru, but many published studies share avoidable weaknesses: small or unreported validation samples, overall accuracy reported without area-adjusted estimates, and classification schemes that are not comparable between studies. The review recommends that urban LULC studies adopt current good-practice guidance on accuracy assessment, use cloud platforms such as Google Earth Engine to analyse long image series, and give more attention to the medium-sized cities, including Solapur, where most of India's future urban growth is expected.

Keywords: Remote sensing; GIS; Land use and land cover; Change detection; Urban sprawl; Landsat; Sentinel-2

1. Introduction

The growth of an Indian city is easy to see from its edges. Agricultural land on the outskirts is subdivided into plots, roads appear before houses, and villages that were once separate become part of the urban area. Tracking this change on the ground is slow and expensive. Census figures arrive once a decade, and municipal records rarely cover the fringe where most new building happens.

Satellite remote sensing changed this. Since the first Landsat satellite in 1972, the same areas of the earth have been imaged repeatedly, and since 2008 the entire Landsat archive has been free to use. With a GIS to combine images with maps, census data and field observations, a geographer can now reconstruct

how a city has grown over several decades. This review describes the methods involved and examines how they have been used in India.

2. Land use, land cover and classification systems

Land cover describes the physical material on the surface, such as buildings, vegetation, water or bare soil; land use describes what people do with it, such as housing, industry or agriculture. Satellite images record land cover directly, and land use must be inferred. Anderson et al. (1976) proposed the classification system that most later studies adapt, with broad first-level classes such as urban or built-up land, agricultural land, forest, water and barren land. Using a standard scheme matters: studies that define their own classes cannot easily be compared.

3. Data sources

Most urban LULC studies in India rely on free medium-resolution imagery (Table 1). The Landsat series offers the longest record at 30 m resolution (Roy et al., 2014), while Sentinel-2, launched by the European Space Agency in 2015, provides 10 m resolution and more frequent coverage (Drusch et al., 2012). Indian Remote Sensing satellites, available through ISRO's Bhuvan portal, are a useful national alternative.

Table 1. Free satellite data commonly used in urban LULC studies.

Sensor	Period of record	Resolution	Revisit	Remark
Landsat 5 TM	1984–2013	30 m	16 days	Basis for historical change
Landsat 7 ETM+	1999 onwards	30 m (15 m pan)	16 days	Data gaps after 2003 sensor fault
Landsat 8 OLI / Landsat 9	2013 / 2021 onwards	30 m (15 m pan)	16 days (8 combined)	Improved radiometry
Sentinel-2 MSI	2015 onwards	10, 20, 60 m	5 days (two satellites)	Best free detail for cities
IRS Resourcesat LISS-III	2003 onwards	23.5 m	24 days	Indian source via Bhuvan

Source: compiled by the author from Roy et al. (2014), Drusch et al. (2012) and agency documentation.

4. Methods

A typical urban LULC change study follows the sequence shown in Figure 1. Each step involves choices that affect the final result.

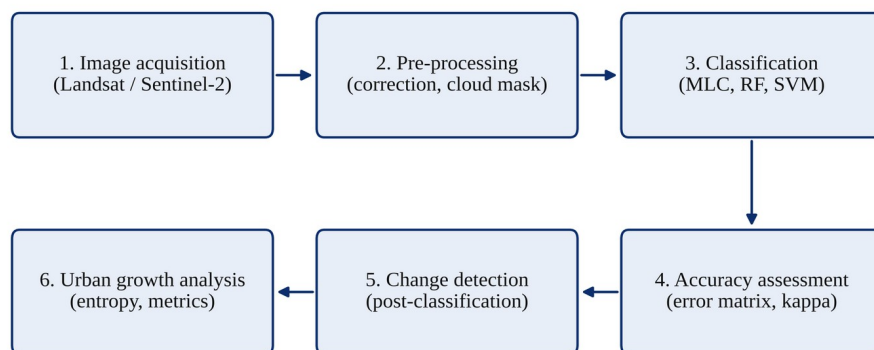


Figure 1. Typical workflow of an urban land use and land cover change study.

4.1 Pre-processing

Images from different dates must be comparable. This requires geometric registration so that the same pixel covers the same ground, atmospheric correction so that differences in haze do not appear as land change, and masking of clouds and their shadows. Analysis-ready surface reflectance products from Landsat and Sentinel-2 now carry out much of this work, which reduces errors that were common in older studies.

4.2 Classification

Supervised classification assigns each pixel to a class using training samples identified by the analyst. Maximum likelihood classification, which assumes that each class follows a normal distribution, was standard for decades and still appears in many Indian studies. Machine learning classifiers have largely replaced it in international work. Belgiu and Drăguț (2016) reviewed the use of random forests in remote sensing and found that they gave high accuracy, handled many input variables and were relatively insensitive to their settings. Support vector machines perform similarly. In fast-growing cities, the main source of error is often the mixed pixel at the urban edge, where a 30 m pixel contains buildings, soil and vegetation together.

4.3 Spectral indices

Spectral indices combine bands to enhance particular surfaces (Table 2). The normalised difference vegetation index (Tucker, 1979) highlights vegetation, the normalised difference built-up index (Zha et al., 2003) highlights built-up land, and the modified normalised difference water index (Xu, 2006) separates water from built-up surfaces. Indices can be used on their own for quick mapping or added as extra inputs to a classifier.

Table 2. Spectral indices commonly used in urban LULC mapping.

Index	Formula	Use in urban studies
NDVI	$(\text{NIR} - \text{Red}) / (\text{NIR} + \text{Red})$	Separates vegetation from built-up and bare land
NDBI	$(\text{SWIR} - \text{NIR}) / (\text{SWIR} + \text{NIR})$	Highlights built-up surfaces
MNDWI	$(\text{Green} - \text{SWIR}) / (\text{Green} + \text{SWIR})$	Maps water bodies; reduces confusion with buildings

NIR = near-infrared; SWIR = shortwave infrared. Sources: Tucker (1979); Zha et al. (2003); Xu (2006).

4.4 Accuracy assessment

A land cover map is only as useful as its accuracy is known. The standard tool is the error matrix, which compares mapped classes with reference data at sample locations (Congalton, 1991). From it come overall accuracy, user's and producer's accuracy for each class, and, in many studies, the kappa coefficient. Foody (2002) and later Olofsson et al. (2014) argued that kappa adds little and that studies should instead report class-wise accuracies, use a proper probability sample of reference points, and estimate the area of each class and of change with confidence intervals. These recommendations are still rarely followed in Indian urban studies.

4.5 Change detection

Lu et al. (2004) reviewed more than thirty change detection techniques, from simple image differencing to model-based methods. Most urban studies use post-classification comparison: each date is classified separately and the maps are overlaid to produce a matrix of changes, for example from agriculture to built-up land. The approach is easy to interpret, but errors in each map multiply in the change map, which is another reason for careful accuracy assessment.

4.6 Measuring urban sprawl

Change in built-up area says how much a city has grown but not how. Yeh and Li (2001) used Shannon's entropy, calculated from the share of built-up land in zones around the city centre or along roads, to measure whether growth is compact or dispersed: values near zero indicate concentration, values near one indicate sprawl. Landscape metrics, such as the number of built-up patches and their mean size, add detail on fragmentation.

5. Indian applications

Indian studies have used these methods to document both the speed and the form of urban growth. Sudhira et al. (2004) combined remote sensing and GIS to study sprawl along the Mangaluru–Udupi region, using Shannon's entropy and landscape metrics to show dispersed growth along highways. Bhatta (2009) used multi-date satellite images and GIS to analyse the pattern of Kolkata's growth and to measure the degree of sprawl. Many later studies of Indian cities follow the same pattern: two to four Landsat images a decade apart, maximum likelihood or random forest classification, post-classification change detection and an entropy measure of sprawl.

Three weaknesses recur in this literature. Validation samples are often small or not described. Overall accuracy and kappa are reported without class-wise accuracies or area-adjusted estimates. And the classes differ from one study to the next, so results from different cities cannot be compared. Most published studies also concern metropolitan cities, while the medium-sized cities where growth is now fastest, such as Solapur, have received comparatively little attention.

6. New directions

Two developments make better studies easier. Google Earth Engine provides the full Landsat and Sentinel archives and the computing power to process them in the cloud, so that an analyst can classify every year of a thirty-year record rather than two or three dates (Gorelick et al., 2017). And the arrival of Sentinel-2 at 10 m resolution has reduced the mixed-pixel problem at the urban fringe. Combining annual image series with good-practice accuracy assessment would make Indian urban LULC studies both more detailed and more comparable.

7. Conclusion

Remote sensing and GIS give geographers a consistent, repeatable record of how Indian cities grow. The methods are mature, the data are free, and cloud platforms have removed most computing barriers. What the field now needs is rigour rather than new tools: standard classification schemes, proper reference samples, class-wise and area-adjusted accuracy estimates, and attention to the medium-sized cities that will shape India's urban future. Studies of this kind would be of direct use to the town planners and municipal bodies that must manage that growth.

Conflict of interest: The author is the Editor-in-Chief of the Global Journal of Advanced Science and Technology. This manuscript was handled independently by another member of the editorial board, and the author took no part in its review or in the editorial decision.

Funding: This research received no external funding.

Declaration of generative AI use: During the preparation of this work the author used an AI writing assistant (Claude, Anthropic) to help structure and draft the text. The author reviewed, edited and verified the content and takes full responsibility for this article.

References

- Anderson, J. R., Hardy, E. E., Roach, J. T., & Witmer, R. E. (1976). A land use and land cover classification system for use with remote sensor data (Geological Survey Professional Paper 964). U.S. Government Printing Office. <https://doi.org/10.3133/pp964>
- Belgiu, M., & Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>
- Bhatta, B. (2009). Analysis of urban growth pattern using remote sensing and GIS: A case study of Kolkata, India. *International Journal of Remote Sensing*, 30(18), 4733–4746. <https://doi.org/10.1080/01431160802651967>
- Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37(1), 35–46. [https://doi.org/10.1016/0034-4257\(91\)90048-B](https://doi.org/10.1016/0034-4257(91)90048-B)
- Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., & Bargellini, P. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120, 25–36. <https://doi.org/10.1016/j.rse.2011.11.026>
- Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80(1), 185–201. [https://doi.org/10.1016/S0034-4257\(01\)00295-4](https://doi.org/10.1016/S0034-4257(01)00295-4)
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>
- Lu, D., Mausel, P., Brondizio, E., & Moran, E. (2004). Change detection techniques. *International Journal of Remote Sensing*, 25(12), 2365–2401. <https://doi.org/10.1080/0143116031000139863>
- Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148, 42–57. <https://doi.org/10.1016/j.rse.2014.02.015>
- Roy, D. P., Wulder, M. A., Loveland, T. R., Woodcock, C. E., Allen, R. G., Anderson, M. C., et al. (2014). Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, 145, 154–172. <https://doi.org/10.1016/j.rse.2014.02.001>
- Sudhira, H. S., Ramachandra, T. V., & Jagadish, K. S. (2004). Urban sprawl: Metrics, dynamics and modelling using GIS. *International Journal of Applied Earth Observation and Geoinformation*, 5(1), 29–39. <https://doi.org/10.1016/j.jag.2003.08.002>
- Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2), 127–150. [https://doi.org/10.1016/0034-4257\(79\)90013-0](https://doi.org/10.1016/0034-4257(79)90013-0)
- Xu, H. (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14), 3025–3033. <https://doi.org/10.1080/01431160600589179>
- Yeh, A. G.-O., & Li, X. (2001). Measurement and monitoring of urban sprawl in a rapidly growing region using entropy. *Photogrammetric Engineering & Remote Sensing*, 67(1), 83–90.
- Zha, Y., Gao, J., & Ni, S. (2003). Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal of Remote Sensing*, 24(3), 583–594. <https://doi.org/10.1080/01431160304987>

REVIEW ARTICLE

Phishing Detection Techniques: A Review of Rule-Based, Machine Learning and Deep Learning Approaches

Mr. B. S. Bukka*

Assistant Professor, Department of Computer Science, Sangameshwar College (Autonomous), Solapur, Maharashtra, India

* Corresponding author

OPEN ACCESS

This article is licensed under CC BY-NC 4.0.

**ARTICLE HISTORY**

Received: 23 Sep 2026

Accepted: 05 Oct 2026

Published: 14 Oct 2026

CORRESPONDING AUTHOR

Mr. B. S. Bukka, Department of Computer Science, Sangameshwar College (Autonomous), Solapur, Maharashtra, India.

HOW TO CITEBukka, B. S. (2026). Phishing Detection Techniques: A Review of Rule-Based, Machine Learning and Deep Learning Approaches. *Global Journal of Advanced Science and Technology*, 1(1), 17–21.

© 2026 The Author(s).

ABSTRACT

Phishing, the use of fake websites and messages to trick people into revealing passwords, bank details or one-time passwords, remains one of the most common ways in which online accounts are compromised. In India, the spread of mobile banking and UPI payments has made SMS and messaging-app phishing a daily risk for ordinary users. This review examines the main techniques developed to detect phishing websites and links. It covers list-based methods such as blacklists, heuristic and rule-based methods, content and visual-similarity methods, machine learning on features extracted from URLs and web pages, and deep learning models that learn directly from the characters of a URL. For each family it discusses how the method works, what it requires and where it fails. The literature shows that blacklists are fast and precise but miss newly created sites, that machine learning on lexical URL features detects most phishing links with little computation, and that deep learning models perform best when large training sets are available. All approaches face the same problems: attackers adapt quickly, public datasets age, and reported accuracies are hard to compare. The review concludes that layered defences, combining lists, lightweight classifiers on the user's device and user awareness, are most effective, and it identifies the need for datasets and models suited to Indian languages and SMS-based phishing.

Keywords: Phishing; Cybersecurity; URL classification; Machine learning; Deep learning; Blacklists

1. Introduction

Most people in India have received a message like this: your electricity will be disconnected tonight, your bank account has been blocked, or your parcel is waiting at the post office, followed by a link. The link leads to a page that looks like the real one and asks for a login, a card number or a one-time password. This is phishing, and it works because it attacks people rather than computers.

Phishing exploits the gap between what a website looks like and what it is. Dhamija et al. (2006) showed participants a set of genuine and fake websites and found that the best fake site fooled 90 per cent

of them; many ignored the browser's security indicators altogether. Because user training alone cannot close this gap, a large body of research has developed automatic methods to detect phishing links and pages.

This review summarises that research. It follows the classification used in the survey by Khonji et al. (2013), updated to include the deep learning methods that have appeared since, and discusses which approaches suit the conditions of Indian users, who encounter phishing mostly through SMS and messaging applications on mobile phones.

2. How phishing works

A phishing attack has three parts: a lure, usually an email, SMS or social media message that creates urgency; a link to a fake site; and the fake site itself, which imitates a trusted brand and collects the victim's information. Attackers register domains that resemble real ones, use URL shorteners to hide the destination, host pages on compromised legitimate websites, and increasingly obtain valid HTTPS certificates so that the padlock icon gives no warning. Each detection method targets one or more of these parts.

3. Detection approaches

Figure 1 groups the detection methods into five families.

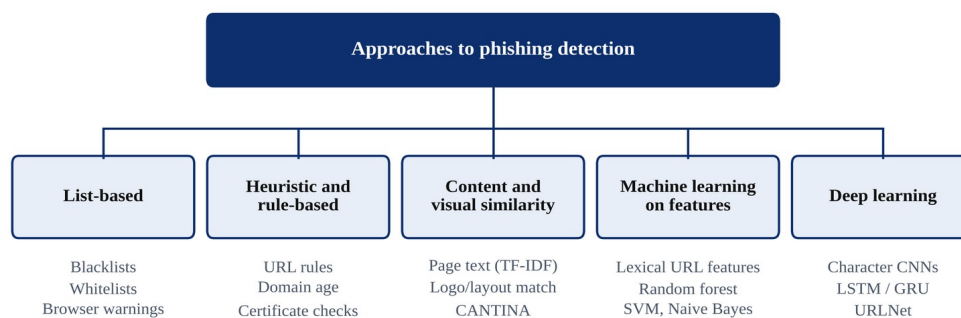


Figure 1. Families of phishing detection approaches.

3.1 List-based methods

Blacklists record URLs and domains already known to be phishing sites; browsers and email services check every link against them. They are fast and almost never wrong, but they can only block what someone has already reported. Sheng et al. (2009) tested several blacklists against newly reported phishing sites and found that most blacklists caught only a small share of them at the moment they first appeared, although coverage improved over the following hours. Since most phishing sites are active for a short time, this delay matters. Whitelists take the opposite approach, trusting only known sites, but are too restrictive for general browsing.

3.2 Heuristic and rule-based methods

Heuristic methods check for features that phishing sites often share: an IP address in place of a domain name, an unusually long URL, an @ symbol, many sub-domains, a recently registered domain, or a login form that sends data to a different domain. Rules can detect new sites that are not yet on any list. Their weakness is that legitimate sites sometimes break the same rules, and attackers soon learn which features to avoid.

3.3 Content and visual-similarity methods

Content-based methods examine the page itself. CANTINA, proposed by Zhang et al. (2007), extracts the most characteristic words on a page using TF-IDF weighting and searches for them on a search engine; if the page's own domain does not appear among the top results, the page is probably imitating another site. Visual-similarity methods compare the appearance of a page, its logo or layout, with those of frequently targeted brands. These methods catch convincing copies but require the page to be downloaded and analysed, which is slower than checking a URL.

3.4 Machine learning on extracted features

Machine learning methods learn the combination of features that separates phishing from legitimate sites. Ma et al. (2009) showed that lexical features of the URL, together with host-based features such as registration and hosting details, allowed a classifier to identify malicious URLs with high accuracy without visiting the page. Mohammad et al. (2014) used a self-structuring neural network on website features, and Abdelhamid et al. (2014) applied associative classification, which produces readable rules as well as predictions. Sahingoz et al. (2019) built a classifier using natural language processing features of the URL alone and reported accuracy close to 98 per cent with a random forest. Table 1 lists the URL features most often used.

Table 1. URL features commonly used by phishing classifiers.

Feature	Example	Why it signals risk
IP address as host	http://203.0.113.5/login	Genuine services rarely use raw IPs
URL length	Very long paths and queries	Used to hide the real domain
@ symbol or “//” in path	bank.com@evil.example	Browser ignores text before @
Number of sub-domains	secure.login.bank.example.xyz	Imitates a trusted brand name
Hyphens and look-alike words	sbi-kyc-update.example	Brand name combined with urgency
Shortened URL	Links via URL shorteners	Hides the destination
Domain age	Registered days earlier	Phishing domains are short-lived

Examples are illustrative. Sources: Ma et al. (2009); Mohammad et al. (2014); Sahingoz et al. (2019).

3.5 Deep learning

Feature-based classifiers depend on the features their designers choose. Deep learning models instead learn their own representation from raw input. URLNet, proposed by Le et al. (2018), applies convolutional neural networks to both the characters and the words of a URL and outperformed models based on hand-crafted lexical features on a large dataset. Recurrent networks such as LSTMs have been used in the same way. These models capture subtle patterns, such as unusual character sequences, but they need large training sets and more computing power, and their decisions are harder to explain to users.

4. Comparison and common problems

Table 2 compares the five families. No single approach is sufficient. Lists are precise but late; heuristics and lightweight classifiers are fast and catch new sites but produce some false alarms; content methods are accurate but slow; deep learning is powerful but demanding.

Table 2. Comparison of phishing detection approaches.

Approach	Detects new sites?	Speed and cost	Main weakness
List-based	No	Very fast; cheap	Misses sites not yet reported
Heuristic rules	Partly	Fast	False alarms; easy to evade
Content / visual	Yes	Slow; page must load	Costly; fails on new brands
Machine learning (features)	Yes	Fast on URL features	Needs updating as attacks change
Deep learning	Yes	Slower; needs GPU to train	Large data needs; low transparency

Three problems affect all of them. First, attackers adapt: once a feature becomes widely used for detection, phishing kits change to avoid it. Second, datasets age quickly, and many studies still use collections that are several years old, so high reported accuracies may not hold on today's attacks. Third, studies use different data, features and measures, which makes their results difficult to compare (Khonji et al., 2013). A model that reports 99 per cent accuracy on an old, balanced dataset may do much worse on real traffic, where legitimate links far outnumber phishing ones.

5. The Indian context

Most research has focused on email phishing in English. In India, the main channels are SMS and messaging applications, and the messages are often written in Hindi, Marathi or other Indian languages, or in a mix of English and an Indian language written in Roman script. Lures commonly imitate banks, UPI applications, electricity boards, courier services and government schemes. Detection on the phone itself, at the moment a link is opened, suits this setting better than server-side filtering, which favours lightweight URL-based classifiers over heavy content analysis.

Research in India would benefit from three things: an openly available dataset of phishing SMS and URLs targeting Indian users; classifiers for multilingual and code-mixed message text; and evaluations that measure performance on realistic, imbalanced streams of links rather than on balanced benchmark datasets.

6. Conclusion

Phishing detection has moved from blacklists to machine learning and deep learning, and each step has improved the ability to catch new attacks. Yet the best protection still comes from layers: lists to block known sites instantly, lightweight classifiers to judge new links on the user's device, content analysis where time allows, and users who know what a phishing message looks like. For Indian users, whose exposure comes mostly through mobile messages in several languages, the most useful next steps are local datasets and models designed for SMS and code-mixed text.

Conflict of interest: The author declares no conflict of interest.

Funding: This research received no external funding.

Declaration of generative AI use: During the preparation of this work the author used an AI writing assistant (Claude, Anthropic) to help structure and draft the text. The author reviewed, edited and verified the content and takes full responsibility for this article.

References

- Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). Phishing detection based associative classification data mining. *Expert Systems with Applications*, 41(13), 5948–5959. <https://doi.org/10.1016/j.eswa.2014.03.019>
- Dhamija, R., Tygar, J. D., & Hearst, M. (2006). Why phishing works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 581–590). <https://doi.org/10.1145/1124772.1124861>
- Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing detection: A literature survey. *IEEE Communications Surveys & Tutorials*, 15(4), 2091–2121. <https://doi.org/10.1109/SURV.2013.032213.00009>
- Le, H., Pham, Q., Sahoo, D., & Hoi, S. C. H. (2018). URLNet: Learning a URL representation with deep learning for malicious URL detection. *arXiv*. <https://doi.org/10.48550/arXiv.1802.03162>
- Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Beyond blacklists: Learning to detect malicious web sites from suspicious URLs. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1245–1254). <https://doi.org/10.1145/1557019.1557153>
- Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications*, 25(2), 443–458. <https://doi.org/10.1007/s00521-013-1490-z>
- Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, 117, 345–357. <https://doi.org/10.1016/j.eswa.2018.09.029>
- Sheng, S., Wardman, B., Warner, G., Cranor, L. F., Hong, J., & Zhang, C. (2009). An empirical analysis of phishing blacklists. In *Proceedings of the Sixth Conference on Email and Anti-Spam (CEAS)*.
- Zhang, Y., Hong, J. I., & Cranor, L. F. (2007). CANTINA: A content-based approach to detecting phishing web sites. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 639–648). <https://doi.org/10.1145/1242572.1242659>

REVIEW ARTICLE

Sample Size Determination in Applied Research: A Critical Review of Cochran's, Yamane's and Power-Analysis Approaches

Sarika B. Patil*

Research Scholar, Punyashlok Ahilyadevi Holkar Solapur University, Solapur, Maharashtra, India

* Corresponding author

OPEN ACCESS

This article is licensed under CC BY-NC 4.0.

**ARTICLE HISTORY**

Received: 21 Sep 2026

Accepted: 06 Oct 2026

Published: 14 Oct 2026

CORRESPONDING AUTHORMs. Sarika B. Patil, Punyashlok
Ahilyadevi Holkar Solapur University,
Solapur, Maharashtra, India.**HOW TO CITE**Patil, S. B. (2026). Sample Size
Determination in Applied Research: A
Critical Review of Cochran's, Yamane's
and Power-Analysis Approaches. *Global
Journal of Advanced Science and
Technology*, 1(1), 22–26.

© 2026 The Author(s).

ABSTRACT

Few questions are asked more often by research students than how large a sample should be, and few are answered more mechanically. Many theses and articles in India justify their sample size with a single formula, most often Yamane's or the Krejcie and Morgan table, without explaining whether the formula suits the study's purpose. This review examines the three approaches most commonly used in applied research: Cochran's formula for estimating a proportion, Yamane's simplified formula, and statistical power analysis for testing hypotheses. It sets out the assumptions behind each, shows with worked calculations how they compare across population sizes and effect sizes, and discusses common errors. The comparison shows that Cochran's formula and the Krejcie and Morgan table give practically identical results, that Yamane's formula is equivalent to Cochran's at a confidence level of about 95.5 per cent and therefore gives slightly larger samples, and that none of these formulas is appropriate when the aim is to compare groups or test a relationship, where power analysis is needed. A study comparing two groups for a small effect needs nearly 400 participants per group at 80 per cent power, far more than survey formulas suggest. The review recommends that researchers state the purpose of the sample, the assumptions used and adjustments for non-response and design effects, and that they justify sample size in the way the analysis will actually be carried out.

Keywords: Sample size; Cochran's formula; Yamane's formula; Power analysis; Research methodology; Effect size

1. Introduction

At some point in every research project, usually when the questionnaire is ready, the researcher must decide how many people to survey or how many units to observe. Too small a sample gives estimates so uncertain that they mean little, or tests that cannot detect real effects. Too large a sample wastes time, money and the goodwill of respondents. Choosing the size well is therefore an ethical as well as a statistical matter.

In practice, the choice is often made by habit. A formula is copied from an earlier thesis, a population size is entered, and the result is reported as the sample size, sometimes rounded up for safety. Whether the formula was designed for the kind of question the study asks is rarely discussed. Button et al. (2013) estimated that the median statistical power of studies in neuroscience was only about 21 per cent, a reminder that inadequate samples are a problem across disciplines, not only in student research.

This review examines the three approaches that applied researchers in India use most: Cochran's formula, Yamane's formula and power analysis. It explains what each is designed to do, compares their results with worked examples, and offers practical recommendations.

2. Two different purposes

The most important distinction is between two purposes of a study. Some studies aim to estimate a quantity, such as the proportion of farmers who use crop insurance or the average household income in a ward. Others aim to test a hypothesis, such as whether a training programme improves test scores or whether income differs between two groups. The first purpose calls for a sample large enough to give an estimate within a chosen margin of error. The second calls for a sample large enough to detect an effect of a given size with a chosen probability. Cochran's and Yamane's formulas serve the first purpose; power analysis serves the second (Lenth, 2001; Lakens, 2022).

3. Cochran's formula

For estimating a proportion p with margin of error e at a confidence level corresponding to the standard normal value z , Cochran (1977) gives the sample size for a large population as $n_0 = z^2 p(1 - p) / e^2$. When the proportion is unknown, $p = 0.5$ is used because it gives the largest sample. At 95 per cent confidence ($z = 1.96$) and a 5 per cent margin of error this gives $n_0 = 384.16$, which is rounded up to 385. For a finite population of size N , the finite population correction gives $n = n_0 / [1 + (n_0 - 1) / N]$.

The table published by Krejcie and Morgan (1970), widely reproduced in Indian methodology textbooks, is based on essentially the same calculation, using the chi-square value 3.841 (which equals 1.96^2) with $p = 0.5$ and $e = 0.05$. The two give the same sample sizes, as Table 1 shows. Bartlett et al. (2001) extended the approach to continuous variables, for which the formula uses the expected standard deviation of the variable in place of $p(1 - p)$.

4. Yamane's formula

Yamane (1967) gave a simpler expression, $n = N / (1 + Ne^2)$, which needs only the population size and the margin of error. Its popularity comes from this simplicity. It is not a separate method, however. Substituting $p = 0.5$ and $z = 2$ into Cochran's formula with the finite population correction gives almost exactly Yamane's result, so Yamane's formula assumes maximum variability and a confidence level of about 95.5 per cent (Israel, 1992). This is why it gives slightly larger samples than Cochran's formula at 95 per cent confidence, and why its results approach 400 rather than 385 for large populations.

Table 1. Required sample size by population size (95% confidence for Cochran and Krejcie–Morgan; $e = 0.05$; $p = 0.5$).

Population (N)	Cochran (with correction)	Yamane	Krejcie and Morgan
100	80	80	80
500	218	223	218
1,000	278	286	278
5,000	357	371	357
10,000	370	385	370
1,00,000	383	399	383
10,00,000	385	400	384

Source: calculated by the author using the formulas in Sections 3 and 4; values rounded up to the next whole number.

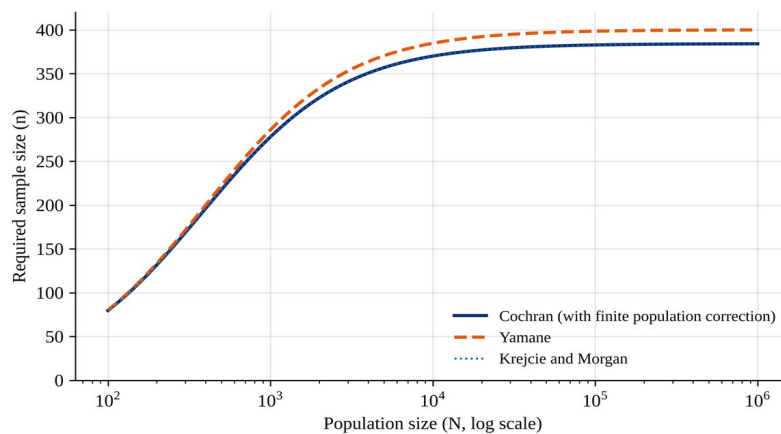
**Figure 1.** Required sample size as population size increases, for the three formulas.

Figure 1 brings out a point that is often misunderstood. Beyond a population of about 20,000, the required sample barely changes. A survey of a district with ten lakh households needs hardly more respondents than one of a town with twenty thousand. The precision of an estimate depends mainly on the size of the sample, not on the fraction of the population it represents.

5. Power analysis

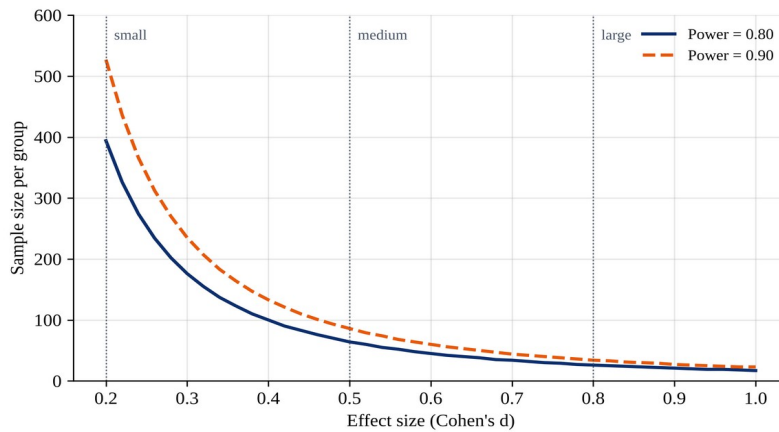
When a study aims to test a hypothesis, the relevant question is different: how large must the sample be to have a good chance of detecting an effect, if a real effect of a given size exists? Power analysis answers this by combining four quantities: the significance level (usually 0.05), the desired power (usually 0.80 or 0.90), the expected effect size and the sample size. Fixing any three determines the fourth (Cohen, 1988).

Cohen (1992) proposed conventional values for effect sizes, such as $d = 0.2$, 0.5 and 0.8 for small, medium and large differences between two means. Table 2 and Figure 2 show the resulting sample sizes for comparing two independent groups with a t-test. Free software such as G*Power performs these calculations for most common tests (Faul et al., 2007).

Table 2. Sample size per group for a two-sample t-test (two-sided, $\alpha = 0.05$).

Effect size (Cohen's d)	Description	Power = 0.80	Power = 0.90
0.2	Small	394	527
0.5	Medium	64	86
0.8	Large	26	34

Source: calculated by the author using the non-central *t* distribution; values agree with those given by G*Power.

**Figure 2.** Sample size per group needed to detect a given effect size in a two-sample t-test ($\alpha = 0.05$).

The contrast with Table 1 is striking. A survey formula might suggest that 385 respondents are enough, yet if the aim is to compare men and women and the true difference is small, nearly 400 respondents are needed in each group. Conversely, for a large effect, 26 per group may be sufficient. Lenth (2001) warns against relying on the conventional small, medium and large values: the expected effect should come from previous studies, pilot data or the smallest difference that would matter in practice.

6. Common errors

- Using Yamane's or Cochran's formula to justify the sample for a study whose main analysis is a test of difference or association.
- Applying the formula to the whole population when results will be reported separately for subgroups; each subgroup then needs an adequate sample.
- Ignoring non-response. If 70 per cent of those approached are expected to respond, the number contacted should be the required sample divided by 0.70 (Bartlett et al., 2001).
- Ignoring the design effect in cluster or multistage sampling, which usually requires a larger sample than simple random sampling.
- Reporting the formula but not the assumptions used: the confidence level, margin of error, expected proportion or effect size.

7. Recommendations

Lakens (2022) describes several legitimate ways to justify a sample size, including measuring the whole population, accepting the limits set by available resources, aiming for a given precision of estimates, and conducting an a priori power analysis. Whatever the approach, the justification should be transparent. Drawing on the literature reviewed here, we recommend that researchers:

1. state whether the main purpose of the study is estimation or hypothesis testing, and choose the method accordingly;

2. report every assumption: confidence level, margin of error, expected proportion or standard deviation, significance level, power and effect size;
3. base the expected effect size on earlier studies or pilot data rather than on conventions alone;
4. adjust the sample for expected non-response and for the design effect where cluster sampling is used;
5. ensure that every subgroup to be analysed separately has an adequate sample.

8. Conclusion

Cochran's formula, the Krejcie and Morgan table and Yamane's formula are closely related tools for one task, estimating a proportion with a chosen precision, and they give very similar answers. They are not suitable for justifying the sample of a study designed to compare groups or test relationships, for which power analysis is the appropriate method and often calls for considerably larger samples. A sample size is not a number to be produced by a formula and forgotten; it is a design decision that should follow from what the study is trying to find out.

Conflict of interest: The author declares no conflict of interest.

Funding: This research received no external funding.

Declaration of generative AI use: During the preparation of this work the author used an AI writing assistant (Claude, Anthropic) to help structure and draft the text. The author reviewed, edited and verified the content and takes full responsibility for this article.

References

- Bartlett, J. E., Kotrlik, J. W., & Higgins, C. C. (2001). Organizational research: Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*, 19(1), 43–50.
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376. <https://doi.org/10.1038/nrn3475>
- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). John Wiley & Sons.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. <https://doi.org/10.1037/0033-2909.112.1.155>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Israel, G. D. (1992). *Determining sample size* (Fact Sheet PEOD-6). Institute of Food and Agricultural Sciences, University of Florida.
- Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement*, 30(3), 607–610. <https://doi.org/10.1177/001316447003000308>
- Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, 8(1), Article 33267. <https://doi.org/10.1525/collabra.33267>
- Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3), 187–193. <https://doi.org/10.1198/000313001317098149>
- Yamane, T. (1967). *Statistics: An introductory analysis* (2nd ed.). Harper and Row.