

The Human Factor in Cybersecurity: Can Machine Learning Detect Phishing Attempts More Effectively than Traditional Rule-Based Methods?

Manvigas Singh Aulakh¹, Raghu Raja Mehra²

¹Department of Information Technology, iNViCTUS International School, Amritsar, Punjab, India

²Mentor, Department of Information Technology, iNViCTUS International School, Amritsar, Punjab, India

Abstract: *Phishing emails succeed not because computers are weak but because people can be persuaded. A message that looks like it comes from a bank, a school or a delivery company, and that creates a sense of urgency, can lead even careful users to click a harmful link or reveal a password. The first line of defence is the email filter. Traditionally, filters have relied on fixed rules such as lists of suspicious words and known bad websites. Machine-learning (ML) filters instead learn the patterns of phishing from labelled examples. This paper asks how effectively ML models classify phishing emails compared with rule-based methods. It combines a review of published studies with a small classroom-scale experiment in which a keyword rule filter was compared with three ML models (Naive Bayes, Logistic Regression and Random Forest) on a constructed set of 1,212 emails. On phishing written in a familiar style, all methods performed well, with the rule filter reaching 96.0% accuracy and the ML models 99.6% to 100%. On phishing written in a new, calmer style, the rule filter caught only 8 of 100 phishing emails, while the Random Forest model caught 87 with only one false alarm. The findings agree with the wider literature: ML is clearly more effective and adapts better to new tricks, but it is not perfect, and it does not remove the human factor. A layered defence that combines ML filtering with user awareness is recommended.*

Keywords: phishing, social engineering, machine learning, rule-based filtering, email security, Random Forest, human factor

1. Introduction

Most people imagine cyber-attacks as hackers breaking through firewalls with clever code. In reality, a large share of attacks begins with something much simpler: a message that tricks a person into doing what the attacker wants. This is known as social engineering, and its most common form is phishing [13], where an attacker pretends to be a trusted organisation in order to steal passwords, money or personal details.

The scale of the problem is large. Verizon's 2024 Data Breach Investigations Report found that a human element, such as falling for a scam or making an error, was involved in about 68 percent of the breaches it studied [1]. The Anti-Phishing Working Group recorded close to one million phishing attacks in a single quarter of 2024 [2]. In India, messages about pending KYC updates, blocked bank accounts, unpaid electricity bills and undelivered parcels have become familiar examples, and they often target people who are new to digital payments. Because people are the weak point, the email filter has an important job: it should stop as many phishing messages as possible before a person ever has to make a decision. For many years, filters worked mainly with hand-written rules. Today, many email services also use machine learning (ML), which allows a computer to learn the difference between genuine and phishing emails from examples. This paper examines whether, and by how much, the ML approach is actually better.

2. Research Question and Scope

The central research question is:

How effectively can machine-learning models classify phishing emails compared with traditional rule-based detection methods?

To answer it, the paper (a) explains why phishing works on people, (b) describes how rule-based and ML filters operate, (c) reviews published findings, and (d) reports a small demonstration experiment designed and run as part of this study. The paper focuses on email phishing. SMS phishing (smishing), voice phishing (vishing) and fake websites are mentioned where relevant but are not tested.

3. The Human Factor: Why Phishing Works

Phishing is less a technical attack than a psychological one. Cialdini's well-known principles of persuasion [3] explain why people respond to these messages. Attackers borrow the authority of banks, government departments or a school principal. They create scarcity and urgency with phrases such as "within 24 hours" or "final notice". They offer rewards like prize money, or they use fear, such as the threat of an account being closed. Each of these pushes the reader to act quickly rather than think carefully.

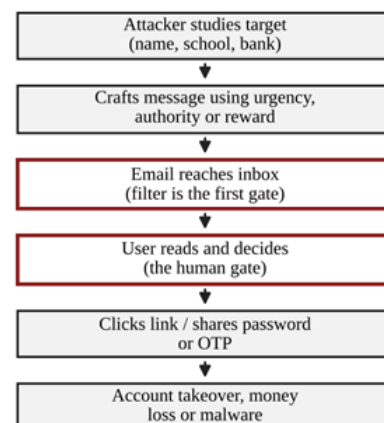


Figure 1: The phishing attack chain. The email filter and the user's own judgement (outlined in red) are the two points where the attack can be stopped.

Figure 1 shows the typical chain of a phishing attack. There are two main points where it can be broken: when the email filter decides whether the message reaches the inbox, and when the user decides whether to trust it. A strong filter reduces how often the human has to make a difficult judgement. This is why the quality of automatic detection matters so much for the human factor.

Table I: Common Persuasion Tactics in Phishing Emails

Tactic	Typical wording	Why it works
Authority	"Your bank", "Income Tax Department"	People tend to obey trusted institutions
Urgency	"Act within 24 hours"	Leaves no time to check
Fear	"Account will be blocked"	Panic reduces careful thinking
Reward	"You have won Rs 50,000"	Greed overrides doubt
Familiarity	"Hi Aman, see the shared file"	Personal tone lowers suspicion

Recent advances make the problem harder. Attackers can now use generative AI tools to write fluent, grammatically correct messages that copy the tone of a real organisation. Older advice such as "look for spelling mistakes" is becoming less useful. This raises the importance of filters that can detect phishing from deeper patterns rather than surface clues.

4. Anatomy of a Phishing Email in the Indian Context

Figure 2 shows a typical phishing email, using a made-up bank name, with its warning signs marked. None of these signs is proof on its own. A real bank may also write "Dear Customer", and a real alert may be urgent. What gives the message away is the combination: an unfamiliar sender domain, pressure to act at once, a threat, and a link that leads somewhere other than the bank's official website.

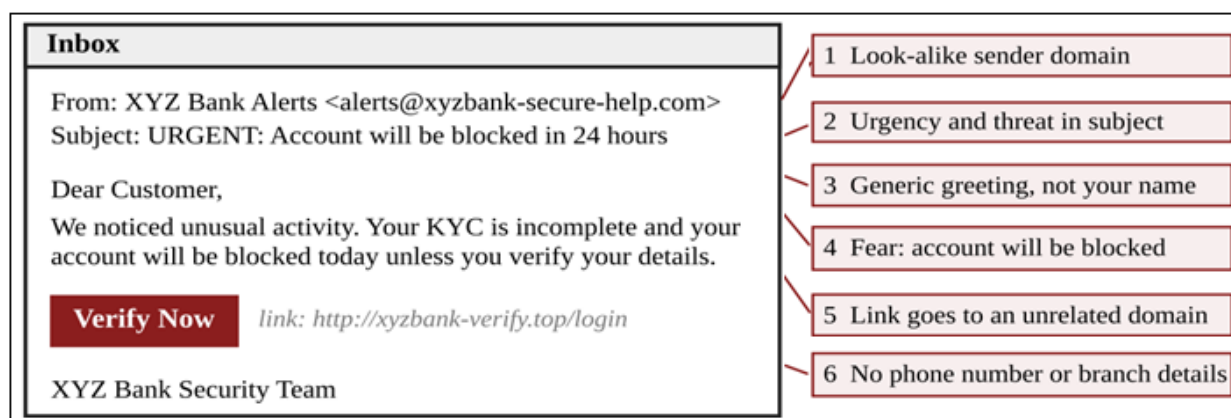


Figure 2: Anatomy of a phishing email (fictional bank). Each numbered clue is weak by itself; together they form a strong pattern, which is exactly the kind of pattern a machine-learning model can learn

In India, phishing messages often follow local events and services. Table II lists some of the most common types reported in news and awareness campaigns, along with the main red flag a careful reader should notice. Many of these arrive by SMS or WhatsApp as well as email, but the tricks are the same.

Table II: Common Phishing Themes Seen in India

Theme	Typical claim	Main red flag
KYC update	Account blocked until KYC is done	Banks do not ask for KYC through a link
Electricity bill	Power cut tonight, pay now	Personal mobile number or odd link
Parcel delivery	Address incomplete, pay small fee	Unexpected parcel, unknown site
Job offer	Easy work-from-home income	Asks for a joining fee
Scholarship / fee	Claim scholarship or pay exam fee	Not on the official school portal
UPI refund	Approve request to get refund	Approving a request sends money out

These examples show why the human factor is so important. Each message is designed for a moment when the reader is busy, worried or hopeful. They also show why a filter needs to recognise patterns rather than fixed words: the wording changes with each new campaign, but the structure of the trick stays similar.

5. Traditional Rule-Based Detection

Rule-based filters follow instructions written by people. The simplest form is a *blacklist*, a list of web addresses and senders already known to be harmful. Another form is a *keyword or scoring rule*. For example, the open-source filter SpamAssassin gives a message point for each suspicious feature, such as certain phrases or a mismatched sender address, and flags it when the total crosses a threshold [14]. Technical checks such as SPF and DMARC, which verify whether an email really came from the domain it claims, are also rule-based [15].

Rule-based methods have real strengths. They are fast, easy to understand and easy to explain: if an email is blocked, the system can say exactly which rule caused it. They also need no training data. Their weakness is that they can only catch what someone has already thought of. A study of phishing blacklists found that most of them caught fewer than 20 percent of new phishing websites at "hour zero", when the attack first started and was most dangerous [7]. Attackers know this and change their wording and web addresses frequently. Keyword rules can also produce false alarms, for

example when a genuine bank security message contains words like "verify" and "password".

6. Machine-Learning Detection

An ML filter does not follow a fixed list of rules. Instead, it is trained on a large set of emails that have already been labelled as "phishing" or "legitimate". During training it learns which features, and which combinations of features, separate the two groups. Figure 3 compares the two pipelines.

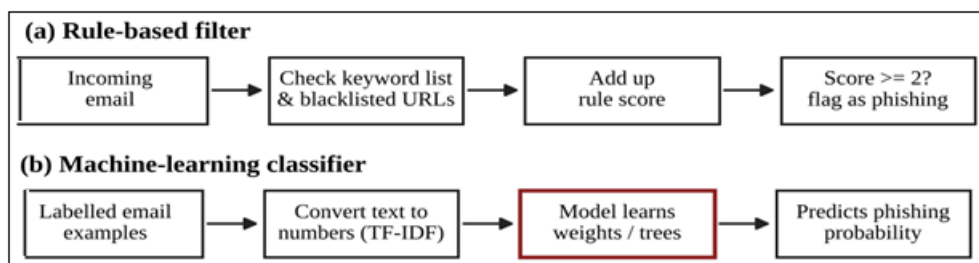


Figure 3: Simplified pipelines of (a) a rule-based filter, where people write the decision rules, and (b) a machine-learning classifier, where the model learns its decision rules from labelled examples

Before learning can happen, each email must be turned into numbers. A common method is TF-IDF (term frequency–inverse document frequency), which gives each word a weight based on how often it appears in this email and how rare it is across all emails. Other useful features describe the link in the message, such as whether it uses secure HTTPS, contains a hyphenated look-alike domain, or uses an unusual domain ending. Three widely used models were considered in this study:

- **Naive Bayes** calculates the probability that an email is phishing from the words it contains. It was one of the first successful spam filters [6].
- **Logistic Regression** gives each feature a positive or negative weight and adds them up to estimate the chance of phishing. Its weights can be inspected, which makes it fairly easy to explain.
- **Random Forest** builds many decision trees, each asking a series of yes-or-no questions, and lets them vote [12]. It can capture combinations of clues that a single rule would miss.

Published studies report strong results. Fette, Sadeh and Tomasic built one of the earliest ML phishing filters using only ten features, and it detected about 96 percent of phishing emails while wrongly flagging only a tiny fraction of genuine ones [4]. Abu-Nimeh et al. compared six ML methods and found that each had different strengths, with Random Forest among the most accurate [5]. Ma et al. showed that ML on web-address features alone could detect malicious websites far more completely than blacklists [8]. More recent reviews report that modern methods, including deep learning and natural language processing, often exceed 95 to 99 percent accuracy on standard datasets [9], [10].

7. Methodology

The study used two methods. First, published research on rule-based and ML phishing detection was reviewed and compared. Second, a small demonstration experiment was

carried out to see the difference directly. Because real phishing datasets often contain personal information, the experiment used a constructed dataset of short emails generated from templates based on common Indian phishing and genuine messages. This makes the experiment a classroom-scale demonstration rather than a full scientific benchmark, and its results should be read in that light.

A. Dataset

Genuine emails included meeting notes, order updates, bank statements, school reminders, one-time password messages and security alerts. Phishing emails were of two kinds. *Familiar-style* phishing used classic signs such as "URGENT", "click here" and "verify your password". *New-style* phishing was calm and business-like, for example a shared salary sheet, an invoice to review or a full mailbox, with very few classic warning words. Random greetings and sign-offs were added, some genuine emails used plain HTTP links, and some phishing used realistic HTTPS look-alike domains, so that no single clue could solve the task. About 3 percent of training labels were deliberately flipped to imitate human labelling errors.

Table III: Composition of the Constructed Email Dataset

Set	Genuine	Phishing	Total
Training	400	312 *	712
Test A (familiar style)	150	100	250
Test B (new style)	150	100	250

* 300 familiar-style and only 12 new-style phishing examples.

B. Methods compared

The rule-based filter gave one point for each suspicious phrase from a list of 15 (for example "urgent", "verify", "password", "click here", "prize", "KYC") and two points for a link pointing to a raw IP address. An email scoring two or more was flagged. The three ML models were trained with the scikit-learn library [11] on TF-IDF word features plus simple link features. All models were trained once and then tested on both test sets without any changes.

C. Measures

Four standard measures were used. **Accuracy** is the share of all emails classified correctly. **Precision** is the share of flagged emails that really were phishing, so low precision means many false alarms. **Recall** is the share of phishing emails that were caught, so low recall means attacks get through. **F1-score** combines precision and recall into one number. For security, recall is especially important because each missed phishing email is a chance for an attack to succeed.

8. Results

Table IV summarises the results, and Figure 4 shows them as a chart.

Table IV: Performance of each method on the Two Test Sets (%)

Method	Acc.	Prec.	Recall	F1
Test A: familiar style				
Rule-based	96.0	90.9	100.0	95.2
Naive Bayes	100.0	100.0	100.0	100.0
Logistic Reg.	100.0	100.0	100.0	100.0
Random Forest	99.6	99.0	100.0	99.5
Test B: new style				
Rule-based	56.4	32.0	8.0	12.8
Naive Bayes	86.0	100.0	65.0	78.8
Logistic Reg.	84.8	100.0	62.0	76.5
Random Forest	94.4	98.9	87.0	92.6

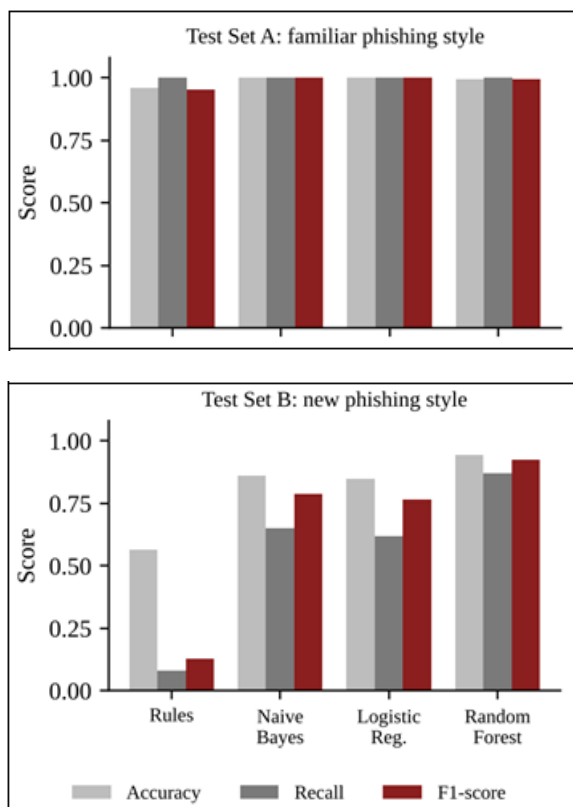


Figure 4: Accuracy, recall and F1-score of each method.

On familiar phishing all methods do well; on new-style phishing the rule-based filter almost collapses while the ML models, especially Random Forest, hold up much better.

A. Familiar-style phishing

On Test A, every method performed well. The rule filter caught all 100 phishing emails, but it also wrongly flagged

10 genuine emails, mostly real bank security alerts and quiz-selection notices that happened to contain words like "click here", "immediately" and "verify". The ML models made almost no mistakes. Such near-perfect scores are partly a result of the template-based data, since the test emails closely resembled the training emails. In real use, scores would be lower.

B. New-style phishing

Test B shows the key difference. The rule filter caught only 8 of 100 phishing emails, because the new messages simply did not contain the words on its list. At the same time, it still produced 17 false alarms on genuine emails. Its F1-score fell from 95.2 to 12.8 percent.

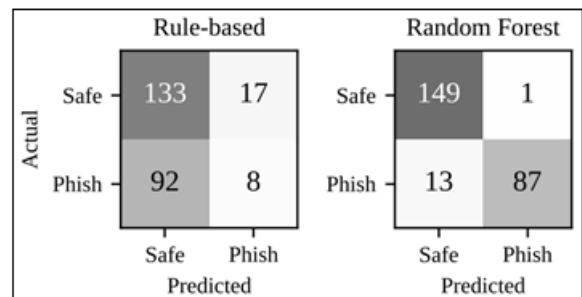


Figure 5: Confusion matrices on Test B. The rule filter let 92 of 100 phishing emails through; Random Forest missed 13 and raised one false alarm.

The ML models also found Test B harder, but they did far better. Naive Bayes and Logistic Regression caught about two-thirds of the new phishing emails with no false alarms. Random Forest caught 87 percent, again with only one false alarm. Its advantage seems to come from combining several weaker clues, such as an unusual link together with a request to "sign in" or "review", rather than depending on any single word.

C. What the model learned

Figure 6 shows the features with the largest weights in the Logistic Regression model. The strongest phishing signals were link features, such as a hyphen in the domain name or a link without HTTPS, rather than scary words. Words common in genuine messages, such as "net banking" in real statement emails, pulled the score towards "legitimate". This suggests the model learned something closer to how an expert checks an email: look carefully at the link, not only at the wording.

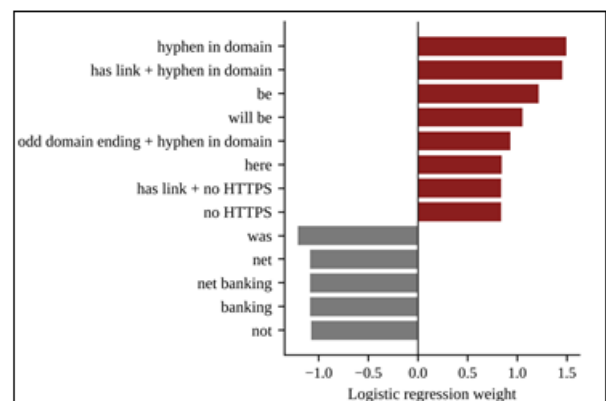


Figure 6: Largest Logistic Regression weights. Red bars push an email towards "phishing"; grey bars push it towards "legitimate"

9. Discussion

Both the experiment and the published studies point in the same direction. When phishing looks like what the filter has seen before, both approaches can work, although rules tend to create more false alarms. When attackers change their style, fixed rules fail quickly, while ML models continue to detect a large share of attacks because they rely on many clues at once. This matches the finding that blacklists catch very little at hour zero [7] and that ML-based detection consistently reports higher accuracy in reviews [9], [10].

Table V: Rule-Based and ML Detection Compared

Aspect	Rule-based	Machine learning
Known phishing	Good	Very good
New phishing styles	Poor	Fairly good
False alarms	Higher	Lower
Needs training data	No	Yes, large and labelled
Easy to explain	Very easy	Harder
Upkeep	Rules written by hand	Regular retraining
Can be fooled	By avoiding keywords	By carefully crafted inputs

A. Why did Random Forest do best?

Naive Bayes and Logistic Regression mostly add up the evidence from each word and link feature separately. Random Forest, in contrast, can learn combinations, for example "a request to sign in" together with "a hyphenated domain". New-style phishing avoids the obvious words, so single clues become weak, but combinations of weak clues still reveal it. This is the same reasoning a careful human uses, and it explains why tree-based and deep-learning models often perform best in published comparisons [5], [9].

B. The cost of a mistake

The two kinds of error are not equally serious. A false alarm, where a genuine email is blocked, is annoying and may cause a student to miss a notice. A missed phishing email can lead to a stolen account or lost money. For this reason, security teams usually prefer high recall even if a few extra genuine emails are sent to the spam folder. In the experiment, the rule filter was weak on both counts in Test B, while Random Forest kept false alarms almost at zero and still caught most attacks.

C. Limits of machine learning

However, ML is not a complete answer. First, it needs large amounts of well-labelled data, and its quality depends on that data. Second, determined attackers can study a model and design emails specifically to avoid it, a problem known as adversarial evasion. Third, ML decisions can be harder to explain, which matters when a genuine message is blocked. Finally, even a filter that catches 95 percent of phishing will let some messages through when millions are sent. The human factor therefore never disappears.

A further challenge is phishing written with the help of AI. Such messages can be fluent, polite and personalised, which

removes many of the word-level clues that both rules and simple ML models depend on. The gift-card message in Appendix A, which contained no link at all and was missed by every method, shows the kind of attack that only an alert human, or a far more advanced model, is likely to stop.

For these reasons, security experts recommend layered defence, shown in Figure 7. Rules and blacklists remove the obvious attacks cheaply. An ML classifier catches many of the new ones. Doubtful messages can be shown with a warning banner. A trained, alert user is the next layer, and two-factor authentication limits the damage if a password is stolen.

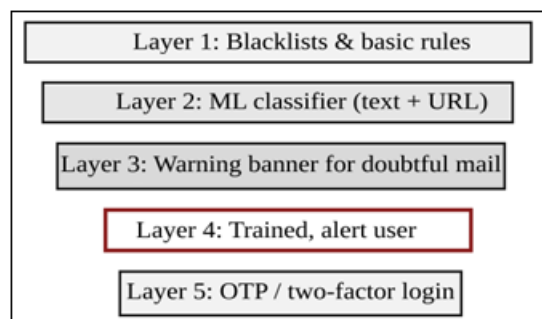


Figure 7: Layered defence against phishing. The human layer (outlined in red) remains essential even with strong automatic filtering.

10. Recommendations

Based on the findings, the following steps are suggested for schools and families:

- Use email services that combine ML filtering with rule checks such as SPF and DMARC, rather than relying only on keyword lists.
- Teach students to check the actual link and sender address, since the experiment shows these are stronger signs than scary words.
- Turn on two-factor authentication for school and banking accounts.
- Run short, regular phishing awareness sessions, using real but safe examples, so that the human layer stays alert.
- Encourage reporting of suspicious emails, which also helps filters learn new patterns.

11. Conclusion

This paper asked how effectively machine-learning models classify phishing emails compared with traditional rule-based methods. The answer is that ML models are clearly more effective, especially against phishing written in new ways. In the demonstration experiment, the rule filter's ability to catch phishing fell from 100 percent to 8 percent when the attackers changed their style, while the Random Forest model still caught 87 percent with almost no false alarms. Published research reports the same pattern on much larger datasets.

Yet the title of this paper also points to the human factor, and that remains central. Phishing works by exploiting trust, urgency and fear, and no filter can be perfect. Machine learning greatly reduces the number of dangerous decisions a person has to make, but it does not remove them. The

most effective protection is therefore a partnership between intelligent filters and informed users.

12. Limitations and Future Work

The main limitation is the dataset. It was constructed from templates and was small, so its results show a pattern rather than real-world accuracy. The rule filter was also simple, and commercial rule systems are much more developed. Future work could test the same methods on a larger public research dataset, include email header features, and measure how well models trained today perform on phishing collected several months later, which would test their ability to adapt over time.

References

- [1] Verizon, "2024 Data Breach Investigations Report," Verizon Business, 2024.
- [2] Anti-Phishing Working Group, "Phishing Activity Trends Report, 1st Quarter 2024," APWG, 2024.
- [3] R. B. Cialdini, *Influence: The Psychology of Persuasion*, rev. ed. New York, NY, USA: Harper Business, 2006.
- [4] I. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," in Proc. 16th Int. Conf. World Wide Web (WWW), 2007, pp. 649–656.
- [5] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in Proc. Anti-Phishing Working Groups eCrime Researchers Summit, 2007, pp. 60–69.
- [6] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," in AAAI Workshop on Learning for Text Categorization, 1998.
- [7] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in Proc. 6th Conf. Email and Anti-Spam (CEAS), 2009.
- [8] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious URLs," in Proc. 15th ACM SIGKDD, 2009, pp. 1245–1254.
- [9] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A systematic literature review on phishing email detection using natural language processing techniques," *IEEE Access*, vol. 10, pp. 65703–65727, 2022.
- [10] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," *Telecommunication Systems*, vol. 76, pp. 139–154, 2021.
- [11] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [13] C. Hadnagy, *Social Engineering: The Science of Human Hacking*, 2nd ed. Indianapolis, IN, USA: Wiley, 2018.
- [14] Apache Software Foundation, "Apache SpamAssassin Project." [Online]. Available: <https://spamassassin.apache.org>
- [15] M. Kucherawy and E. Zwicky, "Domain-based Message Authentication, Reporting, and Conformance (DMARC)," RFC 7489, IETF, 2015.

Appendix A: Sample Emails from the Experiment

The table shows a selection of test emails with the decision of the rule filter (flagged or not) and the phishing probability given by the Random Forest model. An email is flagged by the model when this probability is 0.5 or more. Links are shortened.

Table VI: Selected Test Emails and Model Decisions

Email (shortened)	True label	Rule filter	RF prob.
"URGENT: Your SBI account has been suspended. Click here immediately to verify your password..." (link: amaz0n-rewards.co)	Phishing	Flagged	0.99
"Dear customer, your KYC is incomplete. Update your details now or your account will be blocked today..." (link: raw IP address)	Phishing	Flagged	0.99
"Your ICICI net banking password was changed successfully. If this was not you, click here immediately..." (link: icicibank.com)	Legitimate	Flagged (false alarm)	0.01
"Your Amazon order has been shipped and will arrive by Thursday. Track it in the app."	Legitimate	Not flagged	0.15
"Hi Harpreet, Rohan has shared the revised salary sheet with you. Open the document from the shared drive..." (link: sbionline-help.in)	Phishing (new style)	Missed	0.88
"Hello, attached is the invoice for last month's supplies. Kindly review the payment record through the portal..." (link: icici.netbank-help.ru)	Phishing (new style)	Missed	0.93
"Dear Ishita, your mailbox storage is almost full. Sign in to the mail portal to keep receiving messages." (link: account-services.net)	Phishing (new style)	Missed	0.52
"Hi Ishita, the principal asked me to buy gift cards for the staff event. Can you reply with your mobile number..." (no link)	Phishing (new style)	Missed	0.30 (missed)

Appendix B. Experimental Settings

For transparency, the settings used in the demonstration experiment are listed below so that the experiment can be repeated by other students.

Table VII: Settings of the Rule Filter and the Machine-Learning Models

Item	Setting used
Rule filter word list (1 point each)	urgent, suspended, click here, verify, password, winner, lucky, claim, prize, KYC, blocked, immediately, final notice, dear customer, disconnected
Extra rule (2 points)	Link pointing to a raw IP address instead of a domain name
Rule threshold	Email flagged as phishing when the total score is 2 or more
Text features	TF-IDF of single words and word pairs appearing in at least two training emails
Link features	Link present, no HTTPS, raw IP address, unusual domain ending, hyphen in domain
Naive Bayes	Multinomial Naive Bayes with default smoothing
Logistic Regression	Regularisation strength $C = 2$, up to 2,000 training iterations
Random Forest	200 decision trees, decision threshold 0.5
Software	Python with the scikit-learn library [11]
Label noise	About 3% of training labels flipped at random to imitate human error

Appendix C. Glossary**Table VIII:** Key Terms

Term	Meaning
Phishing	A fake message that pretends to be from a trusted source in order to steal information or money
Social engineering	Tricking people, rather than computers, into giving access or information
Blacklist	A list of senders or websites already known to be harmful
Machine learning	A way for computers to learn patterns from examples instead of fixed instructions
Feature	A measurable clue used by a model, such as a word or a property of a link
Recall	The share of real phishing emails that the filter caught
Precision	The share of flagged emails that were really phishing
False positive	A genuine email wrongly flagged as phishing (a false alarm)
False negative	A phishing email wrongly allowed into the inbox
Two-factor authentication	A login that needs a second proof, such as an OTP, as well as the password