HOW-TO GUIDE SEVEN STEPS

How to automate KYC document verification

A practitioner's sequence for taking identity, address and income documents out of the review queue without losing the audit trail.

WHO THIS IS FOR	WHAT YOU NEED FIRST	WHAT YOU GET
Onboarding, credit ops and compliance leads at banks and NBFCs	A sample of real files, including the rejected ones	A sequenced plan, a rule set and a way to measure it

Most KYC automation projects fail in the same place. The extraction works, the demo goes well, and then the exception queue turns out to be as large as the review queue it replaced, because nobody defined what "verified" means field by field before switching it on. This guide runs the sequence in the order that avoids that, with the decision to make at each step.

01 Map the document set to the decision it supports

Start with the list, not the technology. For each document a customer submits, write down the fields you extract, the check those fields feed, and who currently performs it. Teams routinely discover they are capturing fields nobody uses, and manually verifying fields nobody recorded.

DOCUMENT	FIELDS THAT MATTER	THE CHECK IT FEEDS
Officially valid document	Name, date of birth, document number, validity, issuing authority	Identity establishment and name consistency
Address proof	Full address, issue date, name as printed	Current address, recency window
PAN	PAN, name, father's name, date of birth	Tax identifier validation and dedupe
Bank statement	Account number, IFSC, period, credits, closing balance	Income assessment and banking behaviour
Income proof	Employer, gross and net pay, deductions, period	Eligibility and obligation ratios
Entity documents	Constitution, GSTIN, authorised signatories, beneficial owners	Non-individual KYC and ownership

02 Baseline the current process honestly

You cannot show improvement against a number you did not record. Before any pilot, capture four measurements on a real week of volume. Most teams have the first and none of the rest.

- → **Files received** per day, and the share that arrive incomplete on first submission.
- → **Elapsed time** from submission to decision, including the waiting, not just the working.
- → **Touch time** per file, separating reading and keying from judgment.
- → **Rework rate**: files reopened after a data error was found downstream or at audit.

> **Measure elapsed time, not touch time, as the headline.** Touch time improves the cost case. Elapsed time is what the customer experiences and what determines whether they complete onboarding with you or with someone else.

03 Choose extraction that survives the worst submission

Test on the files your team dreads, not on clean specimens. A KYC document arrives as a phone photograph taken at an angle in poor light, a fourth-generation photocopy, or a scan with a thumb across one corner. Any approach requiring a template per document type will need one per state, per issuing authority and per layout revision, and coverage will plateau.

TEST SET TO INSIST ON

- Photographs, not scans
- Regional-language and bilingual documents
- Documents from smaller issuing authorities
- Files that were rejected last quarter

ASK FOR ACCURACY THIS WAY

- Per field, not per document
- On your files, not a benchmark set
- With the confidence score shown
- Including what it got wrong, and how

04 Write the validation rules before you switch anything on

This is the step that determines whether automation holds. Extraction tells you what the document says. Validation tells you whether to believe it. Write these rules as an explicit list, in four layers, and have compliance sign the list rather than the vendor.

Format and structure
Does each identifier match its expected pattern and checksum? Is the date valid, and does the document fall inside its recency window?

Cross-document consistency
Does the name match across every document, allowing for initials, order and transliteration? Do the date of birth and address agree?

Internal reference
Is this applicant already on file under a different spelling? Does anything match an existing record, a negative list or a prior rejection?

Document integrity
Does the document show signs of alteration, splicing or reuse? Is the submission consistent with what that authority actually issues?

> **The rule most teams miss.** Decide what happens when two documents disagree, before it happens. An unresolved name mismatch that silently picks the first value is worse than one that stops the file, because nobody finds out until an audit.

05 Set confidence thresholds per field, and route by consequence

A single global threshold is the most common design error. The cost of getting a PAN wrong is not the cost of getting a middle initial wrong, so they should not share a rule. Set the threshold field by field, according to what a mistake costs, and make it something the business can move without a release.

FIELD CLASS	IF A MISTAKE GETS THROUGH	ROUTING
Regulatory identifiers	Reportable error, remediation, possible penalty	Highest threshold; review anything below it
Financial values	Wrong eligibility, wrong limit, credit loss	High threshold; verify against a second source
Name and address	Failed dedupe, undeliverable correspondence	Fuzzy match with a tolerance the policy defines
Supporting detail	Minor record inaccuracy, no decision impact	Accept at lower confidence; flag, do not stop

Route exceptions to a reviewer with the source page open at the field in question. A reviewer who has to hunt for context is slower than one who never had the automation.

06 Integrate into the system of record, and keep the evidence

Extracted data that lands in a spreadsheet has not been automated, it has been relocated. Structured output should reach the LOS and core banking system directly, and the audit trail has to travel with it. The question to design against is the one an auditor will ask two years later: why did this field hold this value?

- Every field traceable to its source document and page coordinates.
- Every reviewer override logged with who, when and the prior value.
- The confidence score recorded, not just the accepted value.
- Retention and access aligned to your data protection obligations.

07 Run it in shadow, then tune on real exceptions

Run the automated path alongside the manual one for a defined period on live volume. Compare outcomes file by file rather than in aggregate: where the two disagree is your entire tuning backlog, and it is far more informative than an overall accuracy percentage.

Then review the exception queue weekly for the first quarter. Most queues are dominated by a handful of recurring causes: one issuing authority's layout, one field with a threshold set too high, one validation rule written more strictly than the policy requires. Fixing three of those typically moves the touchless rate more than any model change.

> **Track four numbers monthly:** straight-through rate per document type, exception rate by cause, elapsed time from submission to decision, and reviewer time per exception. The second one tells you what to fix next.

FIVE WAYS THIS GOES WRONG

Piloting on clean files
The pilot succeeds and production does not, because the sample excluded exactly the submissions that cause the work.

One threshold for every field
Set high, the queue swells. Set low, errors reach the record. Neither is a threshold problem; it is a design problem.

Automating capture, not the decision
Extraction feeds a review screen someone still opens for every file. Elapsed time barely moves.

Compliance brought in at UAT
Validation rules get rewritten a week before go-live, and the timeline absorbs it.

No baseline recorded
The programme works and cannot prove it, so the second phase competes for budget without evidence.

A template per format
Coverage plateaus at the formats someone configured, and every layout change reopens the project.

BEFORE GO-LIVE

- [ ] Baseline recorded on a real week, all four measures
- [ ] Validation rules written down and signed by compliance
- [ ] Thresholds set per field class, changeable without a release
- [ ] Disagreement behaviour defined for every cross-document check
- [ ] Audit trail tested by answering an auditor's question end to end
- [ ] Shadow-run period agreed, with a decision rule for going live

How DocXtract supports this

An AI-powered IDP platform built for steps 3 to 6: context-aware extraction with no templates and no per-format training, a validation engine covering identifier checks, cross-document consistency and duplicate detection, field-level confidence scoring, and connectors into the LOS, CBS and wider stack.

- 98%+ accuracy, any document format
- 5,000 pages/hour, real-time JSON
- ₹1.5 per page, live in days

START AT STEP THREE

Test it on your files

Send a sample of real KYC submissions,We report field-level accuracy with confidence scores, and where the current process would have stopped.

Book a Demo

This guide describes an operating approach, not a regulatory interpretation. Your KYC document set, acceptable verification methods, record-keeping obligations and data protection requirements are set by your regulator and internal policy, and should be confirmed with your compliance function before any change to the onboarding process. No specific outcome is implied; measurements described are ones you record in your own environment.