South African Revenue Service

SARS REQUEST FOR INFORMATION

SARS RFI 06/2025

DOCUMENT AUTHENTICATION

AS PART OF THE INTELLIGENT AI VERIFICATION SOLUTION

BUSINESS REQUIREMENTS SPECIFICATION

TABLE OF CONTENTS

1. PREAMBLE

This RFI aims to assess supplier capabilities in document fraud detection, including detection accuracy, supported document types, advanced technologies (e.g. AI, machine learning, computer vision, etc.), solution architecture, deployment options, integration, scalability, security, and regulatory compliance. Suppliers are also invited to share their experiences, implementation challenges, best practices, and commercial models to guide future solution design, procurement and operational alignment.

These guidelines are flexible; broader or more comprehensive responses are welcome.

2. REFERENCES AND DEFINITIONS

ACRONYM OR TERM	DESCRIPTION
AI	Artificial Intelligence
API	Application Programming Interface
CIPC	Companies and Intellectual Property Commission
DPI	Dots Per Inch
EXIF	Exchangeable Image File Format
GPU	Graphics Processing Unit
ID	Identification Document
ISO	International Standards Organization
JPEG	Joint Photographic Experts Group
MRZ	Machine Readable Zone
NIST	National Institute of Standards and Technology
OCR	Optical Character Recognition
PDF	Portable Document Format
PNG	Portable Network Graphics
POC	Proof of Concept
POPIA	Protection of Personal Information Act
POV	Proof of Value
RFI	Request for Information
SARS	South African Revenue Service
SIEM	Security Information and Event Management
SMS	Short Messaging Solution
SOC	Security Operations Center

3. BACKGROUND

SARS currently struggles to manually verify large volumes of supporting documents, resulting in slow, error-prone checks and delayed taxpayer case finalisation.

To solve this, SARS needs an automated document authentication solution (or modular AI service) using machine learning and image forensics to detect document fraud, support real-time and batch processing, and integrate with existing systems. Scalability is essential to handle increasing volumes. Automating verification will help prevent fraud, enhance accuracy, and improve efficiency and service delivery.one specific supplier.

3.1 BUSINESS CONTEXT

SARS receives supporting documents from taxpayers, legal representatives, and other relevant parties as part of its operations. These documents may include financial statements, correspondence, legal contracts, audit reports, and additional case-related materials.

All submitted documents are converted into PDF format to standardize document handling, facilitate secure digital storage, and enable efficient access within SARS's internal systems. Examples of such documents include:

- Retirement Annuity Certificate
- Additional Voluntary Contributions (for example letters and investment statements)
- Bank Statements and Account Confirmation Letters
- Donations certificate/ receipts
- Medical Certificates
- Medical Claims Statements
- Invoices
- Proof of payment (for example bank statements, receipts, supplier account statements)
- Schedule/ List of medical expenses claimed.
- Personal Identification Documents such as IDs, passports and Asylum Seeker Temporary Visas).
- Pictures of an individual holding their ID
- Proof of address
- Point of Sale Slips

Note: SARS verifies the identity of individuals accessing its solutions to prevent impersonation, unauthorized access, and data breaches. While this control confirms legitimate identities, the focus of this RFI is not on biometric verification or solution access controls. Instead, SARS seeks solutions that detect and flag fraudulent supporting documents—specifically falsified copies of identity documents.

3.2 STATISTICS AND SIZING CONSIDERATIONS

User counts and processing volumes will be finalized during the subsequent phases. However, during the PoC phase, the solution is expected to be used by a small, focused group, primarily comprising technical project team members and administrative officers and a defined number of cases will be processed.

For indicative estimates, solution providers may use their own indicative pricing based on their pricing/licensing model. (Infrastructure, users and processing volumes).

4. BUSINESS REQUIREMENTS

SARS is exploring the landscape of document fraud detection solutions available in the market, with interest in technologies capable of ingesting diverse document types, identifying manipulation with high accuracy, and producing results that are both actionable and explainable. The organization is keen to learn how different approaches handle real-time and batch processing, support multilingual inputs, and incorporate pre-validation features.

In particular, SARS seeks to understand how suppliers detect forgeries, alterations, and synthetic documents across various formats and conditions. It is also interested in the kinds of metrics used—such as confidence scores and error rates—and how solutions demonstrate adaptability to evolving fraud techniques

Suppliers are invited to share how their solutions deliver outputs in structured and human-readable formats, manage secure integration, and support auditability and operational decision-making through alerting, prioritization, and feedback mechanisms. SARS welcomes insights into how these capabilities are implemented and how they contribute to continuous learning and improvement.

4.1 DOCUMENT INGESTION

Describe how the supplier's solutions ingest, interpret, and process various document inputs. Responses can include supported file formats, document types, and input methods such as real-time capture and batch processing. SARS is also interested in how solutions handle multilingual content and perform pre-processing, or validation checks to support real-world submission scenarios.

Table 1: Questions on Document Ingestion

REQUEST #	REQUEST DESCRIPTION
SUPPORTED INPUTS AND EXTENSIBILITY	
4.1.1	Which file formats can your solution ingest and analyse? Please include all supported formats – for example, multi-page documents like PDFs and TIFFs, as well as compressed archives like ZIP files. Also, how does your solution detect and handle any format inconsistencies or corrupted files? Describe the methods it uses to identify unsupported formats or file corruption and how it responds when such issues are detected.
4.1.2	How does your solution respond when it encounters a corrupted or unreadable file? Describe the error handling process, including any internal logging and how the solution notifies users or administrators (e.g. through error messages or alerts).
4.1.3	What document types does your solution support out of the box? Provide the range of document categories it can validate without additional configuration (e.g. national IDs, passports, municipal accounts, certificates bank statements). Clarify whether the solution supports variations of these documents and how it distinguishes between similar document types. For instance, explain if it can differentiate among various national ID formats or different types of certificates.
4.1.4	How does your solution accommodate adding new document types? Explain the process for onboarding a new document type into the solution. Specify if introducing a new document type requires vendor involvement or custom development, or if it can be achieved through configuration or training by your team. Describe any tools or interfaces provided to streamline this process and indicate how long it typically takes to operationalize a new document type from start to finish.
4.1.5	Can your solution adapt to different templates of the same document type without significant reconfiguration? For example, if a document type (like a retirement annuity certificate) comes in various layouts or branding from different issuers, describe how your solution handles these variations. How does the solution maintain accuracy across multiple templates for one document type? Explain whether it uses template-agnostic detection methods, machine learning models, or rule-based logic to generalize across different formats, and how it ensures consistent accuracy despite variations in layout or design.
BUSINESS RULE CONFIGURATIONS	
4.1.6	Business Rule Configuration and Data Validation Capabilities Does the solution enable users to define and configure custom business rules that can be applied to key entities or data fields? Please describe how the solution supports user-driven rule creation,

REQUEST #	REQUEST DESCRIPTION
	particularly for validating extracted data. For example, can users implement a rule to verify the validity of a South African ID number using checksum logic?
INPUT HANDLING	
4.1.7	How does your solution detect and respond to fraudulent documents in real time during live transactions and workflows? Describe whether the solution instantly flags suspicious inputs, provides immediate user feedback, and integrates with transactional solutions to trigger alerts or actions without delay.
4.1.8	How does your solution process and analyse multi-page documents within a single file? Clarify whether it maintains contextual continuity across pages, supports page-level fraud detection, and identifies inconsistencies or tampering across sections. Can it detect multiple document types within a single file (e.g. a 20-page file containing both bank statements and certificates) and analyse it according to the document type? Does it have the functionality to analyse preselected pages?
4.1.9	How does your solution validate a small batch of supporting documents submitted simultaneously in real time? Describe whether it can detect inconsistencies across multiple documents (e.g. matching data between an ID, proof of address, and income certificate).
4.1.10	How does your solution handle large batches of documents for bulk verification scenarios? Provide details on the solution's throughput capacity and parallel processing capabilities when verifying hundreds or thousands of files. Describe the error handling mechanisms and indicate whether the solution generates consolidated reports or efficiently flags anomalies across the entire batch.
INPUT VALIDATION	
4.1.11	What minimum image quality does your solution require to reliably validate a document? Specify the required resolution, focus, and clarity levels (for example, any recommended DPI or megapixel values).
4.1.12	What pre-processing does your solution perform on document images before analysis? Describe how it adjusts brightness and contrast, crops out irrelevant borders, and rotates pages for proper orientation prior to analysis.
4.1.13	How does your solution detect and handle files that are encrypted or password-protected? Explain how the solution identifies these files during ingestion. Describe whether it flags them for manual review, prompts for credentials, or excludes them from analysis, and how it communicates the event to users or administrators.

4.2 CORE DETECTION CAPABILITIES

Effective document fraud detection depends on the solution's ability to identify forgeries, alterations, and synthetic identities across various document types and formats. Suppliers are invited to describe their use of advanced techniques such as image forensics and metadata analysis. Inclusion of accuracy metrics will assist in demonstrating consistent performance with minimal false positives or negatives

Table 2: Questions on Core Detection Capabilities

REQUEST #	REQUEST DESCRIPTION
IMAGE ANALYSIS	
4.2.1	How does the solution detect manipulation in scanned or digital documents, including both native and scanned PDFs? Please describe the image forensics capabilities used to identify signs of tampering, such as duplicated regions, lighting or shadow inconsistencies, copied signatures, altered fields, splicing, copy-paste edits, overlays, and pixel-level anomalies (e.g., edge detection or pixel distribution analysis). Can the solution detect proprietary logos and branding specific to a time period, and identify design changes before or after a given date? Does it support identification of similar or identical images and elements across multiple documents, whether embedded in native PDFs or scanned formats?
4.2.2	How does the solution validate visual security features in document images to detect counterfeiting or manipulation? Please describe its ability to verify barcodes (including various types), watermarks, holograms, embossed seals, and MRZ patterns. Can it detect forged or distorted elements, such as incorrect Visa Card Application Identifiers on POS slips or repeated watermarks across documents? What techniques are used to identify missing or improperly replicated features, and does the system support time-based recognition of proprietary logos and branding?
TEXT-BASED FRAUD DETECTION	
4.2.3	How does the solution extract and verify text from documents, including scanned and native PDFs? Please describe its ability to flag anomalies in font style, size, spacing, alignment, spelling, and proprietary fonts. Does it use NLP, NER, and Regex to extract key entities such as names, contact details, bank account numbers, domains, URLs, and email addresses? Can it detect OCR processing and artefacts in native PDFs, and identify reuse of objects like email addresses, phone numbers, and logos across documents?
4.2.4	How does the solution validate machine-readable data against printed text to detect tampering or partial forgery?

REQUEST #	REQUEST DESCRIPTION
	Please describe its ability to compare embedded data—such as Unique Document Identifiers, MRZs, barcodes, QR codes, and proprietary features like SkyQR—with visible information, and flag mismatches. How does it handle detection of forged or distorted machine-readable elements?
4.2.5	How does the solution extract and analyse dates from both document content and metadata? Please describe its ability to identify creation, modification, statement period, and transaction dates, and detect logical inconsistencies (e.g., invalid or out-of-range dates). Can it normalize varying date formats across banks, for example, and flag anomalies? Does the system perform financial consistency checks on documents such as bank statements, including transaction counts, totals, and sequence integrity?
METADATA & FILE FORENSICS	
4.2.6	How does the solution validate machine-readable data against printed text to detect tampering or partial forgery? Please describe its ability to compare embedded data—such as Unique Document Identifiers, MRZs, barcodes, QR codes, and proprietary features like SkyQR—with visible information, and flag mismatches.
4.2.7	How does your solution detect and manage duplicate or reused documents across taxpayer submissions? Describe whether it uses fingerprinting or cryptographic hashing to generate unique identifiers, confirm document integrity, distinguish legitimate reuse from suspicious duplication, and link reused documents to specific taxpayer profiles or submission timelines
APPLICATION OF DETECTION TECHNIQUES	
4.2.8	How does your solution apply fraud detection techniques and decide which ones to use? Describe how it combines or selects methods based on document type, input quality, or risk profile, and explain how it balances performance, accuracy, and false alarm reduction across different scenarios.
MULTIPLE LANGUAGE SUPPORT	
4.2.9	How does your solution handle different languages, alphabets, and document layouts? Describe its ability to extract and validate text from multilingual documents, including those with mixed languages on the same page, and explain any limitations or performance impacts when processing non-English content.
EXTRACTED DATA	
4.2.10	What specific data does the solution extract from each document, and can it make this extracted data available for integration with SARS systems?

REQUEST #	REQUEST DESCRIPTION
	Describe what data the solution extracts from each document and how it exposes this data for integration with SARS systems.

4.3 PERFORMANCE METRICS AND MODEL OPTIMIZATION

Suppliers are requested to provide details on accuracy, error, and confidence metrics, demonstrate robustness across formats, and describe how the solution adapts through retraining and feedback. Include benchmarking methods, where available and evaluation methods used to validate performance and alignment with SARS's operational and compliance requirements.

Table 3: Questions on Performance Metrics

REQUEST #	REQUEST DESCRIPTION
ACCURACY, ERROR RATES & CONFIDENCE	
4.3.1	What is your solution's documented fraud detection accuracy across different document types and fraud scenarios? Include overall accuracy across all document types and fraud scenarios Quantify accuracy differences by document category, quality, file format, and fraud type, when deviations exceeded e.g. ±5% from the overall detection accuracy of the solution. What thresholds would you propose for determining when full automation is appropriate versus when manual review should be applied.
4.3.2	What is your solution's false-positive and false-negative rates, and how do you minimize these errors? Describe how you balance the trade-off between missed fraud and false alarms. Indicate whether the solution allows clients to calibrate this trade-off to align with operational risk tolerance. How does this customization affect detection performance and manual review workload.
4.3.3	How do you calculate and interpret the fraud detection confidence score? Explain how users should use scores (for example, recommended thresholds for manual review) and whether it correlates with the probability of fraud.
4.3.4	Do you report metrics like precision, recall, or F1-score for your detection algorithm? If so, provide those metrics, describe how you compute them, and clarify how you define a "fraudulent" (positive) vs. "genuine" (negative) case to ensure correct interpretation.

REQUEST #	REQUEST DESCRIPTION
MODEL DRIFT AND CONTINUOUS IMPROVEMENT	
4.3.5	How does your solution monitor and manage model drift over time? Describe how it tracks accuracy and error rates in production, detects performance degradation, and triggers retraining—either on a fixed schedule or dynamically based on drift indicators or emerging fraud patterns.
4.3.6	How does your solution support continuous improvement and major model retraining? Explain how it incorporates new fraud patterns, user feedback, or operational insights through incremental updates or rule adjustments. Describe how it manages full retraining cycles and delivers updates to maintain accuracy and responsiveness.
BENCHMARKING & EVALUATION TRANSPARENCY	
4.3.7	Has any third-party organization or industry benchmark evaluated your solution? If so, provide details of the evaluation, including scores and testing bodies. If not, describe any internal or comparative benchmarking you've conducted using public datasets or competitor solutions.
4.3.8	How do you calculate your reported performance metrics? Describe the testing methodology, including dataset size, diversity (e.g. document types, countries, image conditions), and whether tests included real-world scenarios. Clarify how you define and calculate false positives and false negatives. Describe the label split in your dataset (e.g. proportion of fraudulent vs legitimate documents). What modelling techniques are used to address dataset imbalance?
4.3.9	What level of transparency do you offer with performance results? Explain whether you provide detailed evaluation reports, such as confusion matrices, per-document-type success rates, or error breakdowns to support your claims.

4.4 FRAUD DETECTION OUTPUT AND DELIVERY

Effective fraud detection relies on robust output capabilities. SARS is interested in understanding how solutions deliver results in both human-readable and machine-readable formats, and how they integrate securely with existing systems. Suppliers are encouraged to describe the structure and content of outputs—such as data points, risk classifications, and explainability features—and to share how their solutions handle alerting, fraud case prioritization, and feedback mechanisms that support continuous learning and auditability.

Table 4: Questions on Detection Output and Delivery

REQUEST #	REQUEST DESCRIPTION
DETECTION OUTPUT CONTENT	
4.4.1	What data points does your solution include in its fraud detection output? List elements such as fraud score, confidence level, anomaly type (e.g. signature forgery, image manipulation), detection granularity (e.g. logo replaced in top-left corner, font mismatch in line 3 of invoice), and document ID, and explain how it classifies fraud types or risk levels.
4.4.2	How does your solution explain why it flagged a document? Describe how it uses explainable AI, rule-based logic, or annotated outputs to justify decisions and help users understand and trust the results. Include an example of fraud detection results from your solution showing how the solution communicates its findings and rationale.
4.4.3	How does your solution ensure transparency in its decision-making? Clarify whether users can view or audit the logic, rules, or models used, and explain how the solution maintains traceability and allows users to review or challenge its conclusions.
4.4.4	How is the fraud detection outcome delivered and stored? Describe how the solution provides and stores the data points e.g. annotated documents, scores, anomalies? Can the outcomes and supporting artefacts be stored on SARS infrastructure
ALERTING AND ACTIONABILITY	
4.4.5	How does your solution alert users when it detects fraud? Describe how it sends notifications (e.g. email, SMS, in-app, dashboard), whether alerts are configurable by severity, user role, or workflow stage, and whether they include actionable details for follow-up.
4.4.6	How does your solution prioritize fraud cases by severity or risk? Explain how it ranks cases using scoring models, rule-based thresholds, or AI-driven categorization, and how he ranking supports triage, escalation, or resource allocation.
4.4.7	How does your solution incorporate user feedback on flagged fraud cases? Describe how users confirm or reject flags, whether the solution learns from the feedback to improve accuracy, and whether it maintains audit trails or revision history for transparency.

4.5 TECHNICAL SOLUTION CAPABILITIES AND COMPLIANCE

SARS invites suppliers to outline deployment flexibility, including support for real-time and batch processing, latency and bandwidth considerations, and scalability under peak conditions. While integration with SARS platforms (e.g. eFiling and case management) will be defined later, suppliers are encouraged to describe interoperability via APIs and integration tools.

Responses should also explore security and compliance capabilities, including data protection, access controls, audit logging, and alignment with POPIA and ISO 27001. SARS is interested in how AI components remain explainable and traceable to support audit and dispute resolution. Final security and governance requirements will be shared in subsequent phases.

Table 5: Questions on Technical Solution Capabilities and Compliance

REQUEST #	REQUEST DESCRIPTION
DEPLOYMENT OPTIONS	
4.5.1	What deployment models does your solution support? Describe whether it runs in cloud environments (specifically within the South African region), on-premises, or hybrid environments, and explain how each option affects infrastructure, scalability, data residency, integration, and compliance.
4.5.2	How much latency does your solution introduce during document authentication? Explain how it performs in real-time and batch processing scenarios, whether latency varies by document type or solution load, and how it maintains accuracy and responsiveness.
4.5.3	What bandwidth does your solution require for optimal performance? Describe how it handles bulk uploads or frequent validations, and whether it includes features to optimize performance in low-bandwidth environments.
4.5.4	How is your solution architected? Provide a high-level overview, including whether it uses modular microservices or a monolithic platform. Note any dependencies on third-party tools (e.g. OCR engines, databases, GPUs), and explain how the architecture supports high availability and fault tolerance.
INTEGRATION AND INTEROPERABILITY	
4.5.5	How does your solution support integration with third-party platforms and internal applications? Provide documentation for APIs and tools that enable technical connectivity and workflow automation. Describe how your solution reliably exchanges data without requiring extensive customization.

REQUEST #	REQUEST DESCRIPTION
SCALABILITY	
4.5.6	How does your solution scale to handle high volumes of document submissions? Describe how it maintains performance and accuracy under load, and whether it supports horizontal or vertical scaling, load balancing, or cloud-native deployment.
4.5.7	What is your solution's performance limits under peak usage? Provide metrics such as maximum documents processed per minute/hour, concurrent user capacity, and queue behaviour. Explain how the solution handles overload and whether it degrades gracefully or triggers alerts.
DATA PROTECTION	
4.5.8	How does your solution secure data at rest and in transit? Describe the encryption standards and protocols used, whether data is encrypted end-to-end, and how you manage encryption keys to meet security and privacy requirements.
4.5.9	How does your solution ensure full compatibility and compliance with SARS enterprise security controls and infrastructure? For example: Can your application operate entirely through an enterprise web proxy without requiring direct internet access? Does your solution support integration with SIEM platforms and which platforms? How does the solution integrate with enterprise identity and access management solutions, e.g. via Microsoft Active Directory (including Single Sign-On, group-based permissions, and role mapping)?
AUDITABILITY	
4.5.10	How does your solution log events and user actions? Describe how it records document submissions, validation results, errors, alerts, and user interactions (e.g. uploads, overrides, feedback), including timestamps, user identifiers, and outcomes for traceability and compliance.
4.5.11	How can authorized users access and review audit logs? Explain available interfaces, export formats, filtering options, and whether access is role-restricted or integrated with external audit solutions.

REQUEST #	REQUEST DESCRIPTION
4.5.12	How does your solution protect logs and outputs from tampering? Describe safeguards such as immutability, cryptographic protections, and tamper-evident mechanisms like hash verification or blockchain anchoring
COMPLIANCE	
4.5.13	How does your solution comply with data protection and security regulations such as POPIA and ISO 27001? Describe your certification status, audit readiness, and how you embed regulatory requirements into solution design, data handling, and access controls.

5. IMPLEMENTATION REFERENCES

To assess solution relevance, provide case studies from similar industries and operational contexts, highlighting measurable outcomes and fraud detection success. Share key implementation challenges and the best practices you developed to support effective deployment, user adoption, and sustained performance.

Table 6: Questions on Implementation

REQUEST #	REQUEST DESCRIPTION
5.1	Provide references or case studies of previous deployments for clients with similar requirements, scale, and environments. Include the client's industry, use case, measurable outcomes, and how your solution addressed specific fraud detection needs.
5.2	Do you object if SARS contacts the reference organizations to learn more about their implementation and use of your solution?
5.3	What major challenges have you encountered during implementation and ongoing support of document fraud detection solutions?
5.4	What best practices have you developed to ensure successful deployment, user adoption, and long-term solution effectiveness?

6. CONTRACTING MODELS

Describe your typical contracting models, pricing structures, and ongoing technical support. Outline any Proof of Value (PoV) or implementation support you offer to assist with solution evaluation, deployment, and long-term success in meeting decision-making and fraud detection needs.

Table 7: RFI Questions on Contracting Models

REQUEST #	REQUEST DESCRIPTION
6.1	Describe the type of contracting models that your organisation typically uses for full implementation of your decision-making solution?
6.2	How do you determine pricing for your solution? Is it based on usage volume, number of API calls, or another model? Describe pricing structures, such as tiered options and licensing models.
6.3	How would the technical solution be supported and maintained? Provide details on service levels, response times, system uptime guarantees and maintenance schedules.
6.4	What Proof of Value (POV)/ implementation support would your organisation offer?