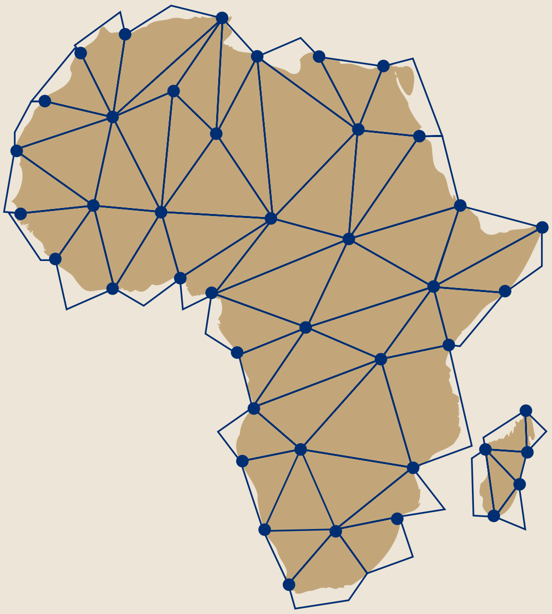
AJIC *The African Journal of Information and Communication*

Issue 35
2025



THEMATIC SECTION: BRICS COUNTRIES AND AI SOVEREIGNTY

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

LINK
centre

The African Journal of Information and Communication

Issue 35, 2025

Published by the LINK Centre, School of Literature, Language and Media (SLLM), Faculty of Humanities, University of the Witwatersrand (Wits), Johannesburg, South Africa

https://ajic.wits.ac.za

Editors

Managing Editor: Ntsibane Ntlatlapa
Associate Professor, LINK Centre, University of the Witwatersrand, Johannesburg
ntsibane.ntlatlapa@wits.ac.za

Corresponding Editor: Lucienne Abrahams
Director, LINK Centre, University of the Witwatersrand, Johannesburg
witslinkcentre20@gmail.com

Publishing Editor: Chris Armstrong
Research Associate, LINK Centre, University of the Witwatersrand, Johannesburg
chris.armstrong@wits.ac.za

The African Journal of Information and Communication (AJIC) is a peer-reviewed, interdisciplinary, open access academic journal focused on the myriad dimensions of electronic and digital ecosystems that facilitate information, communication, innovation and transformation in African economies and in the broader Global South.

Accredited by the South African Department of Higher Education and Training (DHET), *AJIC* publishes online, free to the user, under a Creative Commons licence, and does not impose article processing charges. *AJIC* is indexed in Scientific Electronic Library Online (SciELO) SA, the Directory of Open Access Journals (DOAJ), Sabinet African Journals, and Wits University WIReDSpace, and is hosted on the Academy of Science of South Africa (ASSAf) Khulisa Journals platform.

Peer-reviewing

AJIC acknowledges with gratitude the following peer reviewers of submissions to this issue: Anthony Faustine, Arthur Gwagwa, Maria Luciano, Ekaterina Martynova, Machdel Matthee, Thipe I. Modipa, Anastasia Nedayvoda, Olaperi Okuboyejo, Rosemary Quilling, Felipe Medon and Smriti Parsheera.

Production

Design: Farm Design

Sub-editing: LINK Centre

Proofreading: Linda Van de Vijver and Philippa Van Aardt

Desktop-publishing: LINK Centre

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

LINK centre

SLLM SCHOOL OF LITERATURE LANGUAGE AND MEDIA

TSHIMOLOGONG WITS UNIVERSITY

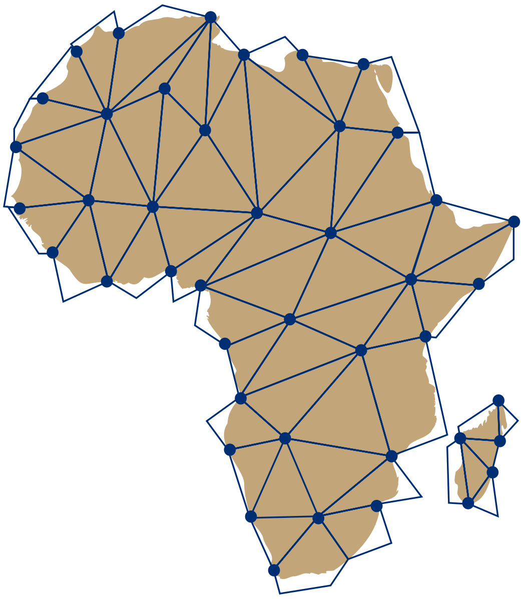
AJIC is published by the LINK Centre, School of Literature, Language and Media (SLLM), Faculty of Humanities, University of the Witwatersrand (Wits), PO Box 601, Wits 2050, Johannesburg, South Africa. The LINK Centre is based at the Wits Tshimologong Digital Innovation Precinct, Johannesburg, https://www.tshimologong.joburg

ISSN 2077-7213 (online version)

ISSN 2077-7205 (print version)

Past issues of *AJIC,* and its precursor *The Southern African Journal of Information and Communication (SAJIC)*, are available at https://ajic.wits.ac.za/issue/archive and https://www.wits.ac.za/linkcentre/sajic

CONTENTS



THEMATIC SECTION: BRICS COUNTRIES AND AI SOVEREIGNTY

Detection of GenAI-produced and student-written C# code: A comparative study of classifier algorithms and code stylometry features

Adewuyi Adetayo Adegbite
Postdoctoral Research Fellow, Department of Computer Science and Informatics, University of the Free State, Bloemfontein, South Africa; and Lecturer, Adekunle Ajasin University, Akunba Akoko, Ondo State, Nigeria
https://orcid.org/0000-0001-8195-1382

Eduan Kotzé
Associate Professor and Head of Department, Department of Computer Science and Informatics, University of the Free State, Bloemfontein, South Africa
https://orcid.org/0000-0002-5572-4319

Abstract

The prevalence of students using generative artificial intelligence (GenAI) to produce program code is such that certain courses are rendered ineffective because students can avoid learning the required skills. Meanwhile, detecting GenAI code and differentiating between GenAI-produced and human-written code are becoming increasingly challenging. This study tested the ability of six classifier algorithms to detect GenAI C# code and to distinguish it from C# code written by students at a South African university. A large dataset of verified student-written code was collated from first-year students at South Africa's University of the Free State, and corresponding GenAI code produced by Blackbox.AI, ChatGPT and Microsoft Copilot was generated and collated. Code metric features were extracted using modified Roslyn APIs. The data was organised into four sets with an equal number of student-written and AI-generated code, and a machine-learning model was deployed with the four sets using six classifiers: extreme gradient boosting (XGBoost), k-nearest neighbors (KNN), support vector machine (SVM), AdaBoost, random forest, and soft voting (with XGBoost, KNN and SVM as inputs). It was found that the GenAI C# code produced by Blackbox.AI, ChatGPT, and Copilot could, with a high degree of accuracy, be identified and distinguished from student-written C# code through use of the classifier algorithms, with XGBoost performing strongest in detecting GenAI code and random forest performing best in identification of student-written code.

Keywords
C# code, generative AI (GenAI) code, student-written code, machine-learning, code classification, code stylometry features

DOI: https://doi.org/10.23962/ajic.i35.21309

Recommended citation

Adegbite, A.A., & Kotzé, E. (2025). Detection of GenAI-produced and student-written C# code: A comparative study of classifier algorithms and code stylometry features. *The African Journal of Information and Communication (AJIC)*, *35*, 1–20. https://doi.org/10.23962/ajic.i35.21309

1. Introduction

Artificial intelligence (AI) has recently advanced significantly in several domains, transforming businesses, industries, and academia with its power, especially with the advent of large language models (LLMs) (Makridakis, 2017). Software development is one field where AI is having a strong influence (Kuhail et al., 2024). Generative artificial intelligence (GenAI) code is rapidly progressing due to advancements in natural language processing (NLP) and deep neural language models. GenAI code is autonomously generated source code (such as Python, C++, or C#) based on high-level requirements, specifications, or samples using machine-learning methods, especially deep-learning models (Odeh et al., 2024). The algorithms learn to produce new code that satisfies predetermined standards through the use of enormous repositories of pre-existing code, computer languages and patterns (Song et al., 2019).

The proliferation of GenAI code is fuelled by a few key technologies. LLMs, such as OpenAI's GPT (generative pre-trained transformer) series, have shown impressive capacities for comprehending and producing text that resembles human-written text (Cao et al., 2023). When used in code creation, these models can transform plain-language descriptions of desired functionality into executable code. Neural architectures are a crucial element in the production of code that satisfies predetermined requirements, and architectures are created specifically for code-creation activities (Dehaerne et al., 2022). The models learn to map input–output pairings, such as code snippets and their accompanying functionality. Code semantics, syntax, and patterns can be analysed and understood by such AI models when trained on extensive code repositories (Wan et al., 2023). In this field, transformers (Vaswani et al., 2017), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) are model variants that are often used.

Thus, AI models can generate code that respects coding standards and complies with best practices (White et al., 2023) in the form of patterns, to solve common problems when using LLMs. Integrated development environments (IDEs) and code editors with AI capabilities such as Microsoft Copilot (Nghiem et al., 2024) offer intelligent and instantaneous code completion, recommendations and corrections (Cao et al., 2023). Through context, user behaviour, and pre-existing code analysis, these technologies improve developer efficiency and decrease mistakes. In codebases, AI algorithms can recognise common errors, anti-patterns, and code smells (potentially problematic code), and automatically recommend optimisations, refactorings, or repairs (Zhang et al., 2022).

While GenAI code is showing great promise, several issues and concerns need to be considered. Retaining good quality, correctness, and semantic meaning in produced code is still a challenge (Krasniqi & Do, 2023). AI models trained on biased or incomplete datasets may produce unfair or undesirable results (Varona & Suárez, 2022). Also, in the educational setting, there is the problem of students presenting GenAI code as their own when submitting computer-programming assignments, and this undermines the development of efficient and effective programmers. Accordingly, for educational institutions to maintain educational standards in their computer-programming courses, it is necessary to have tools that can assist educators in detection of possible student submission of assignments comprising AI-generated code instead of code written by the student. In line with this need for detection tools, the study presented in this article tested the ability of classifier algorithms to distinguish between AI-generated C# code and C# code written by first-year students at the University of the Free State, South Africa.

2. Literature review

A branch of software engineering called "code stylometry" examines programmers' writing styles and habits by analysing their source code (ShaukatTamboli & Prasad, 2013; Zafar et al., 2020). Code stylometry assigns authorship to sections of code based on their stylistic characteristics, much like the use of text stylometry in NLP, which examines writing styles to identify authors of texts (Benzebouchi et al., 2019; Ding et al., 2019). Code stylometry uses a variety of linguistic and structural elements taken from source code to describe programmers' writing styles (Odeh et al., 2024; Tereszkowski-Kaminski et al., 2022). These code stylometry features include lexical, structural, statistical, and syntactic features. Lexical features comprise vocabulary choices, comments, and programming construct usage. Structural features describe the arrangement of control structures, loops, and function definitions. Statistical features explain token distributional properties, language construct frequencies, and code metrics. Syntactic features, which are patterns in code structure, include indentation, naming conventions, and code organisation.

Code stylometry methodologies include elements of machine-learning, statistical analysis and NLP. To find patterns in code and determine authorship, researchers use methods including authorship attribution models, clustering algorithms, and classification techniques (Kalgutkar et al., 2019). A wide range of applications for code stylometry can be found in: authorship attribution; software evolution (which analyses the evolution of code-writing styles over time to understand developer behaviour, project dynamics, and software quality); code reuse and plagiarism detection (which compares writing styles and code patterns); and security analysis, forensics, malware analysis, and cyber-attack attribution (Caliskan et al., 2018; Czibula et al., 2022). Code stylometry tools are, thus useful for determining programmers' writing styles and code authorship (Tereszkowski-Kaminski et al., 2022). Source-code plagiarism is a critical issue in programming, and several studies have been conducted to explore detection methods. Table 1 lists key successful studies of code-plagiarism detection and the detection methods used.

Table 1: Studies on detection of code plagiarism (i.e., detection of plagiarised non-GenAI code)

Study	Programming language(s)	Detection method
Ebrahim and Joy (2023)	Java and C++	Binary classification via pretrained models: UnixCoder, PLBART, and CodeBERTa
Cheers et al. (2023)	Java	Combination of three classifiers: JPlag (structural), Graph ED (semantic), and BPlag (behavioural)
Eliwa et al. (2023)	C, C++, and Java	Similarity detection strategies using JPlag embedded with LMS
Cheers et al. (2021)	Java	Analysis of program-execution behaviour
Lalitha et al. (2021)	Java and Python	Combination of three classifiers: naïve Bayes, KNN, and AdaBoost
Srivastava et al. (2021)	Java	Levenshtein algorithm using edit distance between original code and perceived plagiarised code (the difference between the two codes, and the estimated plagiarism percentage)
Maryono et al. (2019)	Pascal	Euclidean distance on data for similarity measurement (by determining term-document matrices using keywords and programming characters, and then applying hierarchical clustering)
Zheng et al. (2018)	Python and Java	Abstract syntax trees
Portillo-Dominguez et al. (2017)	C++	Combination of three plagiarism tools (JPlag, Sherlock, and SIM)

GenAI production of programming code originated in early work on symbolic AI and automated programming. Early efforts focused on rule-based systems, expert systems, and genetic programming techniques. Notable progress was then achieved with the introduction of LLMs and deep-learning architectures. Today's GenAI-coding uses a variety of techniques and methods (Odeh et al., 2024; Raiaan et al., 2024). Natural language descriptions of functionality can be interpreted by NLP models, such as OpenAI's GPT series, and converted into executable code. Symbolic AI approaches produce code by combining statistical techniques with rule-based systems (Kotsiantis et al., 2024; Raiaan et al., 2024).

In the education context, according to Idialu et al. (2024), even without AI use, programming courses already suffer from high levels of plagiarism and contract-cheating (a situation where students give their tasks and assignments to an expert to solve the given problems). The use of AI tools for code generation (e.g., GitHub Copilot, Tabnine, Gemini, ChatGPT, Blackbox.AI, Mistral, Microsoft Copilot) is now further undermining academic integrity in such courses. The ease with which GenAI tools can generate code has produced a new form of academic dishonesty, with students submitting GenAI code as their own work (Kazemitabaar et al., 2024). Thus, it has now become necessary for academic instructor to find ways to detect possible AI-based plagiarism of code. Table 2 lists studies that have succeeded in detecting GenAI-produced Python, Java and C code as produced by LLMs including ChatGPT models, GitHub Copilot and others.

Table 2: Studies on detection of GenAI code

Study	Programming language	GenAI model(s) used	Detection method(s)
Corso et al. (2024)	Java	GitHub Copilot, Tabnine, ChatGPT, Google Bard	CodeBLEU and Levenshtein similarity analysis on both the generated code and developer code
Idialu et al. (2024)	Python	ChatGPT-4	Machine-learning classifier: XGBoost
Pan et al. (2024)	Python	ChatGPT (version not indicated)	Existing AI text detectors: GPTZero, GPT-2 Detector, DetectGPT, Sapling, and giant language model test room (GLTR)
Bukhari et al. (2023)	C	Code-cushman-001, code-davinci-001, code-davinci-002 (OpenAI code model variants)	Machine-learning classifiers: random forest, SVM, KNN, XGBoost

The study set out in this article focused on detection of GenAI C# code, and on distinguishing between GenAI and student-written C# code, because, to our knowledge, no such studies had previously been carried out in the South African educational context.

3. Study design

The GenAI C# code used in the study was produced by the Blackbox.AI, ChatGPT, and Microsoft Copilot LLMs, and the student-written code was produced by university students. Table 3 lists the versions used for each of the GenAI models.

Table 3: GenAI models used

Model	Version	Data freshness
Blackbox.AI	Blackbox.AI 1.0	Up to September 2024
ChatGPT-4o-mini	Gpt-4o-mini-2024-09-31	Up to October 2023
Microsoft Copilot	Copilot 1.1	Up to February 2023

The study tested the ability of six classifiers—extreme gradient boosting (XGBoost), k-nearest neighbors (KNN), support vector machine (SVM), AdaBoost, random forest, and soft voting (with XGBoost, KNN and SVM as inputs)—to distinguish between GenAI C# code and student-written C# code. These six classifiers were selected based on their successful application in existing studies of code stylometry.

XGBoost constructs decision trees iteratively optimising an objective function to strike a balance between prediction accuracy and model simplicity (Bukhari et al., 2023; Idialu et al., 2024). KNN is an instance-based learning algorithm that classifies data points based on the majority class of their nearest neighbours, identifying the k closest points in the feature space to a new example and assigning the most common class label among them. This method relies on the assumption that similar data points exist in proximity (Bukhari et al., 2024). SVM constructs a hyperplane, or set of hyperplanes, in a high-dimensional space to separate different classes, optimising the distance between the hyperplane and the nearest points of each class, called support vectors. By maximising this margin, SVM ensures robust classification focusing on generalisation to unseen data (Bukhari et al., 2024).

AdaBoost combines multiple weak classifiers, typically decision trees, to form a strong classifier focusing on misclassified examples by adjusting their weights, thereby forcing subsequent classifiers to pay more attention to complex cases. Each classifier contributes to the final prediction with a weight proportional to its accuracy. Random forest builds an ensemble of decision trees by training multiple trees on random subsets of the training data and features. This ensemble approach reduces the risk of overfitting and enhances the model's generalisation capabilities (Bukhari et al., 2024). Soft voting is an ensemble method that combines predictions from three base models, namely XGBoost, KNN, and SVM, to improve overall performance. Prediction is based on the average probabilities assigned by the models (Lalitha et al., 2021). All classifiers were trained using their default hyperparameters without additional tuning. The study also used SHAP (SHapley Additive exPlanations) to help explain the performance of the classifiers (Lundberg & Lee, 2017).

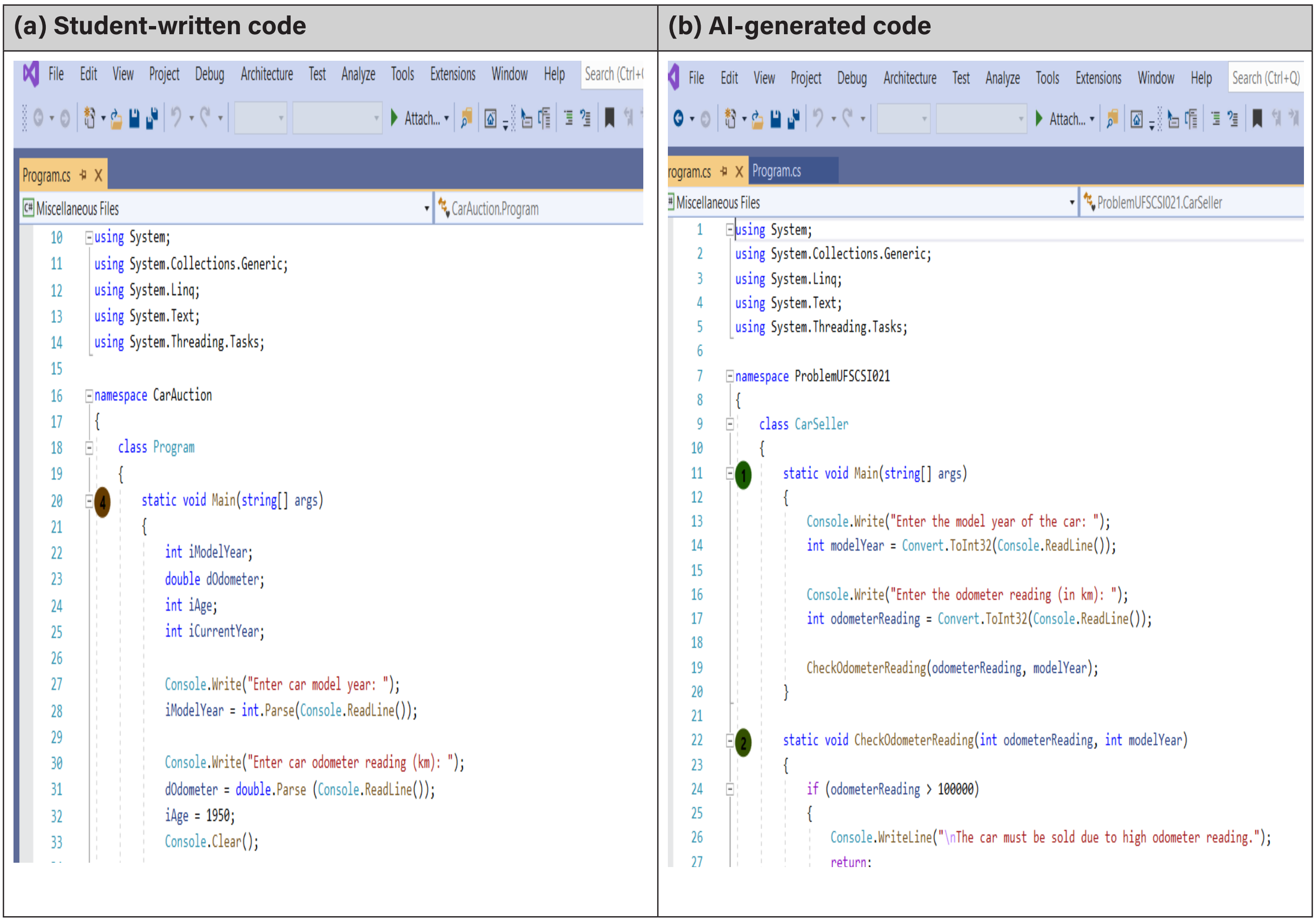
Ethical clearance

Ethical clearance for this study was granted by the General/Human Research Ethics Committee of the University of the Free State, South Africa, and the ethical clearance number is UFS-HSD2024/0601.

Data collection

During data collection, C# code files (*.cs) were extracted from Visual Studio C# solution files submitted in response to nine problems by 314 first-year Computer Science students at the Bloemfontein campus of the University of the Free State. The solution files, a sample of which is shown in Figure 1(a), were written in a controlled environment under the supervision of lecturers and student assistants, thus ensuring that the code produced was purely student-written. The same nine problems were presented (by a separate group of 219 students) to Blackbox.AI, ChatGPT, and Microsoft Copilot to solve, and this allowed for the collection of GenAI-generated C# code data, a sample of which is shown in Figure 1(b).

Figure 1: Sample C# code (for Problem 2)



(a) Student-written code

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;

namespace CarAuction
{
    class Program
    {
        static void Main(string[] args)
        {
            int iModelYear;
            double dOdometer;
            int iAge;
            int iCurrentYear;

            Console.Write("Enter car model year: ");
            iModelYear = int.Parse(Console.ReadLine());

            Console.Write("Enter car odometer reading (km): ");
            dOdometer = double.Parse (Console.ReadLine());
            iAge = 1950;
            Console.Clear();
```

(b) AI-generated code

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;

namespace ProblemUFSCSI021
{
    class CarSeller
    {
        static void Main(string[] args)
        {
            Console.Write("Enter the model year of the car: ");
            int modelYear = Convert.ToInt32(Console.ReadLine());

            Console.Write("Enter the odometer reading (in km): ");
            int odometerReading = Convert.ToInt32(Console.ReadLine());

            CheckOdometerReading(odometerReading, modelYear);
        }

        static void CheckOdometerReading(int odometerReading, int modelYear)
        {
            if (odometerReading > 100000)
            {
                Console.WriteLine("\nThe car must be sold due to high odometer reading.");
                return;
```

The nine problems were labelled Problem 1 through 9 (Appendix C), with the ordering in ascending order of difficulty, i.e., Problem 9 was the most difficult. In the student-written C# code dataset, there were 1,043 solutions across the nine problems (Table 4). In the GenAI C# code dataset (Table 5), there were 1,120 GenAI C# categorised code solutions across the nine problems, and 195 uncategorised code solutions in total. (The uncategorised code solutions were those for which the GenAI model could not be clearly identified.)

Table 4: Student-written C# code

Problem no.	No. of student C# code solutions
Problem 1	108
Problem 2	105
Problem 3	113
Problem 4	202
Problem 5	104
Problem 6	102
Problem 7	104
Problem 8	105
Problem 9	100
Total	1,043

Table 5: GenAI C# code

Problem no.	No. of Blackbox.AI C# code solutions	No. of ChatGPT C# code solutions	No. of Copilot C# code solutions	Total categorised GenAI C# code solutions	Total uncategorised GenAI C# code solutions	Grand totals of C# code solutions
Problem 1	60	42	44	**146**	14	160
Problem 2	47	48	49	**144**	4	148
Problem 3	31	34	42	**107**	48	155
Problem 4	37	40	39	**116**	17	133
Problem 5	47	44	43	**134**	14	148
Problem 6	40	45	40	**125**	13	138
Problem 7	44	45	44	**133**	17	150
Problem 8	32	30	33	**95**	40	135
Problem 9	40	41	39	**120**	28	148
Total	378	369	373	**1,120**	195	1,315

Feature extraction

Modified Roslyn API[1] was used to extract the code metrics. For use of modified Roslyn API, the code must be written in visual code by creating a .NET console app with the addition of Microsoft.CodeAnalysis, Microsoft.CodeAnalysis.CSharp, and Microsoft.CodeAnalysis.CSharp.Syntax. The modified Roslyn code read the contents of the C# files into a string and then parsed the C# code into a syntax tree by using its syntax walk to analyse and extract features such as InterpolatedStringCount, StatementCount, MethodCount, ClassCount, VariableDeclarationCount, which are significant to code analysis as regards structure, syntax and semantics. The modified Roslyn API checked through the syntax tree to extract the code metric features, which are embedded into the modified Roslyn code. The extracted metrics were aggregated into a data structure for analysis, with the extracted features saved into an Excel file.

[1] https://learn.microsoft.com/en-us/dotnet/csharp/roslyn-sdk

Eighty-three code stylometry features (Appendix A) extracted from the modified Roslyn API were used to train and evaluate the classifiers. The metrics fell into four categories, namely lexical, syntactic, layout, and semantic. Lexical features focus on individual tokens in the code such as keywords, identifiers, operators, and literals, reflecting the vocabulary and basic elements used. Extracted examples included UniqueIdentifiers, AverageIdentifierLength, and InterpolatedStringCount. Syntactic features capture the structural organisation of the code, including arrangement of statements, control flow constructs, and the syntax tree. The syntactic features extracted included IfStatementCount, MethodCount, NestedBlockDepth, and NamespaceCount. Layout features differentiate code based on formatting consistency, and the features extracted included NonWhitespaceLines, TotalLines, LineCount, and AverageLineLength. Semantic features capture the meaning or behaviour of the code, such as data flow, control flow, or implemented logic, and extracted features included MethodInvocationCount, CyclomaticComplexity, and ExpressionStatementCount.

CsvHelper and CsvHelper.Configuration were used to extract these code stylometry features into an Excel file for easy training and testing on the six classifier models. The command prompt was used to run the extraction command, with the directory set to the location of the modified Roslyn Visual Studio file. The "dotnet restore" command was run to check and read the .csproj in the project folder, and then the needed package from NuGet was downloaded into the solution package, and this command addressed any version conflicts. After this, the "dotnet build" command was used to compile the source code into a code that could be executed by .NET runtime and also checked for errors. Finally, the "dotnet run" command was used at the command prompt template, followed by a double quotation of the folder directory housing and saving the C# code files to extract the code stylometric. A confirmation message appeared in the command prompt, indicating the creation of the Excel file and successful writing of the code metrics.

Creation of four datasets

The experiment used 80% of the collected data for training and 20% for testing. All training was performed using group five-fold cross-validation with five splits: in each split, one fold was used for testing and the other four for training. Splitting ensured that no data point from any group appeared in both the training and testing sets. For all models, the number of estimators (n_estimators) was set manually to 100, and no hyperparameter search was performed. This default value provides a good balance between performance and computational cost.

Data was arranged into four sets:

- Set 1 comprised student-written code and a combination of Blackbox.AI, ChatGPT, and Copilot code.
- Set 2 comprised student-written code and Blackbox.AI code.
- Set 3 comprised student-written code and ChatGPT code.
- Set 4 comprised student-written code and Microsoft Copilot code.

After feature extraction, the data used ensured a balance between the student-written code and the GenAI code across each problem. The data used for the training and testing of each set included 1882, 756, 738, and 746 code solutions for Set 1, Set 2, Set 3, and Set 4, respectively. The training data for Set 1 is outlined in Table 6. The training data for Sets 2–4 is given in Table 7.

Table 6: Data in Set 1

Problem no.	No. of student code solutions	No. of Blackbox.AI code solutions	No. of ChatGPT code solutions	No. of Copilot code solutions	Total no. of AI code solutions	Grand total
Problem 1	108	**36**	36	36	**108**	**216**
Problem 2	105	**35**	35	35	**105**	**210**
Problem 3	107	**31**	34	42	**107**	**214**
Problem 4	116	**37**	40	39	**116**	**232**
Problem 5	104	**34**	35	35	**104**	**208**
Problem 6	102	**34**	34	34	**102**	**204**
Problem 7	104	**34**	35	35	**104**	**208**
Problem 8	95	**32**	30	33	**95**	**190**
Problem 9	100	**33**	33	34	**100**	**200**
Totals	**941**	306	312	323	**941**	**1,882**

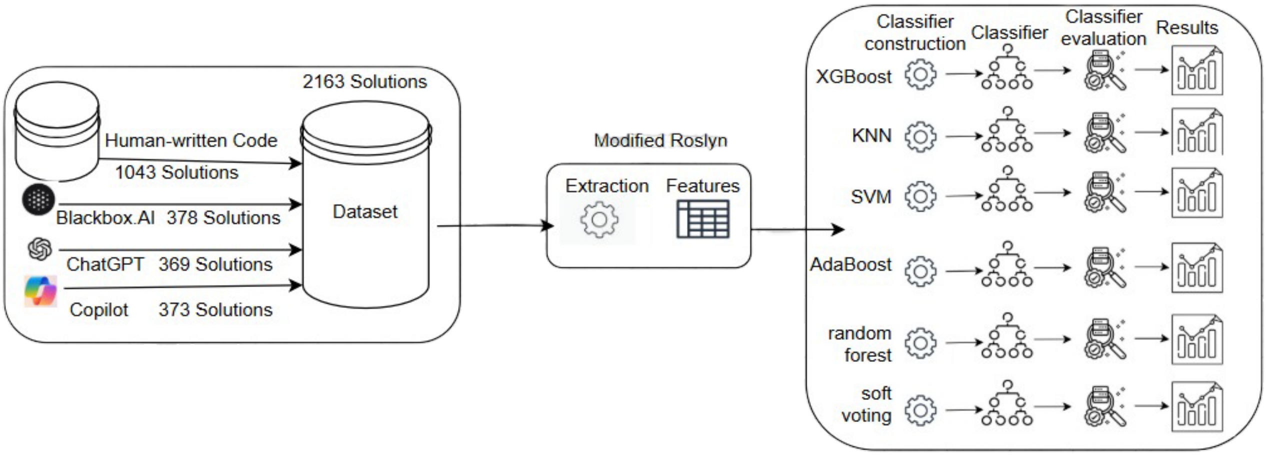
Table 7: Data in Sets 2–4

Problem no.	Set 2 Student-written and Blackbox.AI code			Set 3 Student-written and ChatGPT code			Set 4 Student-written and Copilot code		
	Student code	Blackbox.AI code solutions	Total	Student code	ChatGPT code solutions	Total	Student code	Copilot code solutions	Total
Problem 1	60	60	120	42	42	84	44	44	88
Problem 2	47	47	94	48	48	96	49	49	98
Problem 3	31	31	62	34	34	68	42	42	84
Problem 4	37	37	74	40	40	80	39	39	78
Problem 5	47	47	94	44	44	88	43	43	86
Problem 6	40	40	80	45	45	90	40	40	80
Problem 7	44	44	88	45	45	90	44	44	88
Problem 8	32	32	64	30	30	60	33	33	66
Problem 9	40	40	80	41	41	82	39	39	78
Totals	378	378	**756**	369	369	**738**	373	373	**746**

Testing of the classifier algorithms

In this study, no preprocessing or encoding was applied to the datasets prior to training the classifiers, because the 83 code stylometry features extracted using modified Roslyn were inherently numerical and continuous, thus representing quantitative properties of the code with no categorical variables (Appendix B). There were no missing values in the dataset, as all 83 features were successfully extracted. A machine-learning model (Figure 2) that sought to distinguish between GenAI C# code and student-written C# code was constructed, using the six aforementioned classifiers: XGBoost, KNN, SVM, AdaBoost, random forest, and soft voting (with XGBoost, KNN and SVM as inputs).

Figure 2: Model pipeline



Performance metrics

The performance of each classifier was measured using five metrics: accuracy, recall, precision, F1 score, and AUC-ROC (area under the curve-receiver operating characteristic).

Accuracy gives an overall measure of correctness and, to avoid giving misleading information, the datasets in this study were balanced with equal amounts of GenAI code and student-written code.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Where TP represents correctly identified GenAI code; TN represents correctly predicted human-written code; FP represents incorrectly classified GenAI code; and FN stands for incorrectly classified human-written code.

Recall measures the actual Gen AI code that is correctly identified, and a high recall indicates that GenAI code is rarely missed.

$$\text{Recall} = \frac{TP}{TP+FN}$$

Precision measures the instances predicted as GenAI code that are actually GenAI. High precision indicates the low possibility of human-written code being classified and flagged as GenAI code.

$$\text{Precision} = \frac{TP}{TP+FP}$$

F1 score is the harmonic mean of precision and recall, which checks the balance between detecting GenAI code and reducing the possibility of human-written code classified as GenAI code.

$$\text{F1 score} = 2 \times \frac{Precision \times Recall}{Precision+Recall}$$

AUC-ROC valuates the model’s ability to differentiate between GenAI and human-written code across different classification thresholds.

$$\text{AUC-ROC} = \int_0^1 TPR(FPR)d(FPR)$$

Where TPR is the true positive rate and the same as recall, and FPR is the false positive rate,

$$\text{FPR} = \frac{FP}{FP+TN}$$

4. Results and discussion

Set 1

In the test results for Set 1, which combined student-written code and code produced by all three LLMs (Blackbox.AI, ChatGPT, and Microsoft Copilot), it was found that random forest performed best for student-code detection with accuracy of 0.97, supported by recall of 0.97, precision of 0.88, and F1 score of 0.92 (Table 8). This strong performance indicates that random forest effectively identified nearly all the student-written code, with minimal false negatives (i.e., student code misclassified as AI-generated). For the AI-generated code, XGBoost performed best with accuracy of 0.93, recall of 0.89, precision of 0.96, and F1 score of 0.92. Both random forest and XGBoost achieved an AUC-ROC of 0.98, indicating strong discrimination between student and AI code across various thresholds.

Table 8: Classifier performance with Set 1 (student-written and AI code (from three LLMs))

Classifier	**Accuracy**		**Recall**		**Precision**		**F1 score**		**AUC-ROC**
	Student code	AI code	Student code	AI code	Student code	AI code	Student code	AI code	
XGBoost	**0.93**	**0.93**	0.89	0.89	**0.96**	**0.96**	**0.92**	**0.92**	**0.98**
KNN	0.78	0.70	0.78	0.70	0.75	0.76	0.75	0.70	0.81
SVM	0.79	0.75	0.79	0.75	0.81	0.83	0.77	0.75	0.92
AdaBoost	0.87	0.87	0.94	0.80	0.84	0.94	0.88	0.85	0.95
random forest	**0.97**	0.87	**0.97**	0.87	0.88	**0.97**	**0.92**	0.91	**0.98**
soft voting	0.90	0.90	0.87	**0.94**	0.93	0.87	0.90	0.90	0.97

Set 2

In the test results for Set 2, which combined student-written and Blackbox.AI-produced code, it was found that XGBoost and random forest performed best for student-code detection with accuracy of 0.92 (Table 9). Also, for the student code, XGBoost had recall of 0.88, precision of 0.95, and F1 score of 0.91, while random forest had recall of 0.92, precision of 0.88, and F1 score of 0.89. XGBoost's higher precision indicated fewer false positives, while random forest's higher recall suggested that it was slightly better at capturing all student code. For the Blackbox.AI-generated code, XGBoost was the best classifier, with accuracy of 0.92, recall of 0.88, precision of 0.95, and F1 score of 0.91. These metrics suggest that Blackbox.AI-generated code has distinct features that XGBoost effectively leverages. Both XGBoost and random forest achieved an AUC-ROC of 0.98, reinforcing their strong performance on this set.

Table 9: Classifier performance with Set 2 (student-written and Blackbox.AI code)

Model	**Accuracy**		**Recall**		**Precision**		**F1 score**		**AUC-ROC**
	Student code	AI code	Student code	AI code	Student code	AI code	Student code	AI code	
XGBoost	**0.92**	**0.92**	0.88	0.88	**0.95**	**0.95**	**0.91**	**0.91**	**0.98**
KNN	0.79	0.70	0.79	0.74	0.75	0.81	0.76	0.76	0.85
SVM	0.71	0.89	0.71	0.89	0.88	0.77	0.78	0.82	0.92
AdaBoost	0.89	0.89	**0.93**	0.86	0.88	0.93	0.90	0.89	0.96
random forest	**0.92**	0.86	**0.92**	0.86	0.88	0.92	0.89	0.88	**0.98**
soft voting	0.90	**0.90**	0.86	**0.94**	0.93	0.87	0.89	0.90	0.97

Set 3

In the test results for Set 3, which combined student-written and ChatGPT-produced code, it was found that random forest performed best for student-code detection with accuracy of 0.97, recall of 0.97, precision of 0.82, and an F1 score of 0.88 (Table 10). For detection of ChatGPT-generated code, AdaBoost was the best classifier with accuracy of 0.88, recall of 0.85, precision of 0.91, and F1 score of 0.87. The lower recall compared to other sets suggests that ChatGPT code is harder to detect, potentially due to student-like characteristics. XGBoost, AdaBoost, random forest, and soft voting all achieved an AUC-ROC of 0.97, indicating robust class separation despite the challenges posed by ChatGPT code.

Table 10: Classifier performance with Set 3 (student-written and ChatGPT code)

Model	Accuracy		Recall		Precision		F1 score		AUC-ROC
	Student code	AI code	Student code	AI code	Student code	AI code	Student code	AI code	
XGBoost	0.86	0.86	0.79	0.79	**0.93**	0.93	0.85	0.85	**0.97**
KNN	0.82	0.68	0.82	0.68	0.75	0.78	0.77	0.71	0.81
SVM	0.62	0.91	0.62	0.91	0.89	0.72	0.71	0.80	0.87
AdaBoost	0.88	**0.88**	0.90	0.85	0.88	0.91	**0.88**	**0.87**	**0.97**
random forest	**0.97**	0.74	**0.97**	0.74	0.82	**0.97**	**0.88**	0.82	**0.97**
soft voting	0.87	0.87	0.80	**0.93**	0.92	0.83	0.86	**0.87**	**0.97**

Set 4

In the test results for Set 4, which combined student-written and Microsoft Copilot-produced code, it was found that random forest performed best for student code detection with accuracy of 0.97, recall of 0.97, precision of 0.84, and F1 score of 0.90 (Table 11). For the Copilot-generated code, the soft voting classifier performed best with accuracy of 0.90, recall of 0.96, precision of 0.86, and F1 score of 0.90. Random forest achieved the highest AUC-ROC of 0.99, followed by XGBoost with 0.98 and AdaBoost and soft voting with 0.96.

Table 11: Classifier performance with Set 4 (student-written and Copilot code)

Model	Accuracy		Recall		Precision		F1 score		AUC-ROC
	Student code	AI code	Student code	AI code	Student code	AI code	Student code	AI code	
XGBoost	0.89	**0.89**	0.80	0.80	**0.97**	**0.97**	0.88	0.88	0.98
KNN	0.86	0.70	0.86	0.70	0.77	0.82	0.80	0.72	0.82
SVM	0.72	0.86	0.72	0.86	0.88	0.78	0.75	0.80	0.91
AdaBoost	0.88	0.88	**0.97**	0.79	0.85	0.96	**0.90**	0.85	0.96
random forest	**0.97**	0.79	**0.97**	0.79	0.84	**0.97**	**0.90**	0.86	**0.99**
soft voting	0.90	**0.90**	0.84	**0.96**	0.95	0.86	0.89	**0.90**	0.96

Results across the four sets

The results across the four sets indicated that with respect to the AI-generated code, the Blackbox.AI code was the easiest to detect, as demonstrated by the high accuracies in Set 2. ChatGPT (Set 3) and Copilot (Set 4) were more challenging, with lower detection accuracies for AI-generated code. This suggests that Blackbox.AI produces code with stylistic or structural features that are more distinct than

ChatGPT and Copilot when compared to student code. ChatGPT and Copilot apparently generate code that more closely mimics human patterns, possibly due to their advanced language-modelling capabilities. XGBoost dominated AI code detection in Sets 1 and 2 (accuracies of 0.93 and 0.92), while AdaBoost and voting classifier were better suited to Sets 3 and 4 (accuracies of 0.88 and 0.90), respectively. XGBoost's optimisation of decision trees appears to make it more attuned to optimising patterns in AI-generated code, as suggested by the results from Sets 1 and 2. Also notable was the fact that for detection of AI-generated code, the use of the soft voting classifier, which integrates inputs from three classifiers, markedly improved the recall rate.

The results also indicated that the student-written C# code was generally easier to detect than the GenAI C# code, as most classifiers demonstrated superior or equivalent accuracy in identifying the student-written code. This finding may reflect greater variability in student coding styles, making them easier to distinguish from the more uniform AI-generated code. Random forest consistently excelled in student-code detection across Sets 1, 3, and 4, and tied with XGBoost in Set 2, showing its robustness for identifying student code. Random forest's superior performance can be attributed to its algorithmic strength, which reduces overfitting and enhances generalisation, making it well suited to capturing diverse patterns in student-written code.

Feature analysis

SHAP was used to further explain the outputs of the machine-learning models, through assigning an importance value to each feature for prediction, i.e., showing the features that were most influential in identifying and classifying the AI-generated code and the student-written code.

For Set 1, the five most important features were InterpolatedStringCount, StatementCount, NamespaceCount, LineCount, and TotalLines. For Set 2, the five most important features are InterpolatedStringCount, LineCount, TotalLines, StatementCount, and ExpressionStatementCount. For Set 3, the five most important features were InterpolatedStringCount, NamespaceCount, StatementCount, NonWhitespaceLines, and ExpressionStatementCount. For Set 4, the five most important features were InterpolatedStringCount, NamespaceCount, StatementCount, NonWhitespaceLines, and ExpressionStatementCount.

InterpolatedStringCount is the number of interpolated strings such as $"Hello, {name}" in the C# code. GenAI code tends to use more of these strings for dynamic values, while human-written code tends to have more concatenation methods. StatementCount is the number of statements in the C# code. GenAI code uses more dense lines of code than the corresponding human-written code, showing structural differences. NamespaceCount is the number of namespaces used in the C# code. GenAI code tends to use a limited number of namespaces and to use an optimised relevant one, while human-written code tends to include unnecessary and even unused namespaces, thus having more namespaces than GenAI code.

NonWhitespaceLines are the lines in the C# code that contain actual code or comment. GenAI code tends to have more NonWhitespaceLines than human-written code. TotalLines measures the overall length of C# code. GenAI code tends to have fewer lines of code than its human-written counterpart. ExpressionStatementCount is the number of expression statements in the C# code, e.g., expressions such as assignments, compound, type casting, function return, and conditional statements. GenAI code tends to have more such statements than human-written code because GenAI code seeks to explicitly state each operation for clarity. LineCount is the total number of lines in the C# code, and GenAI code tends to use fewer lines than human-written code.

As seen in the SHAP feature-importance graphs below for Sets 1 and 2 (Figure 3) and Sets 3 and 4 (Figure 4), the most important features across the four sets were InterpolatedStringCount and StatementCount, which both appear in the top five features for each set.

Figure 3: SHAP feature importance for Sets 1 and 2

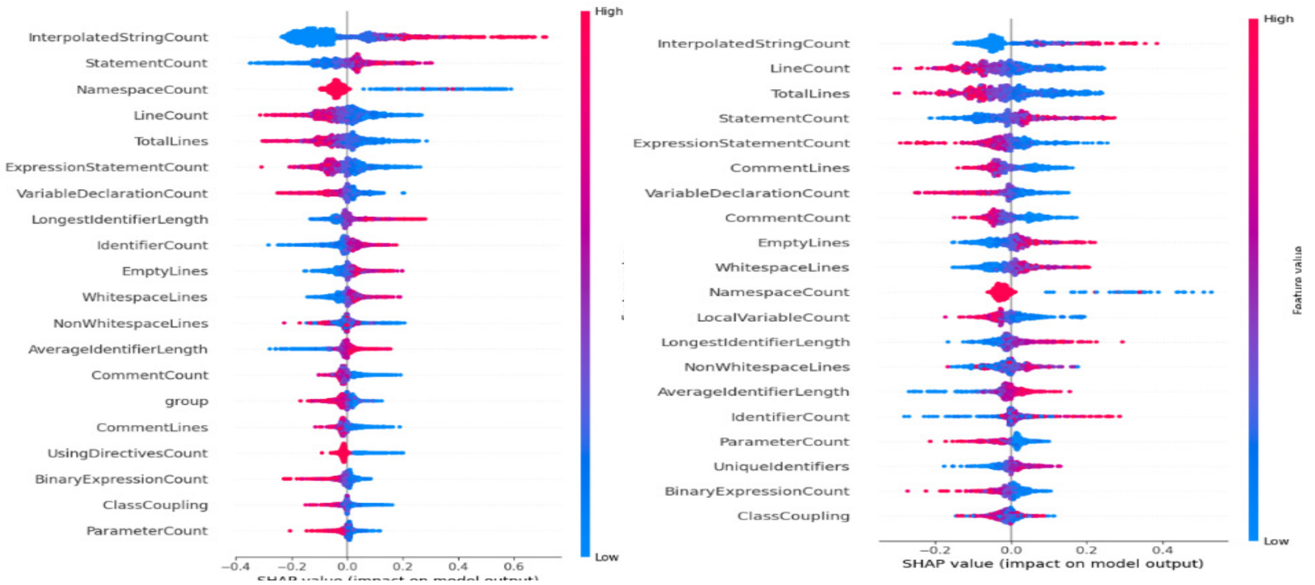
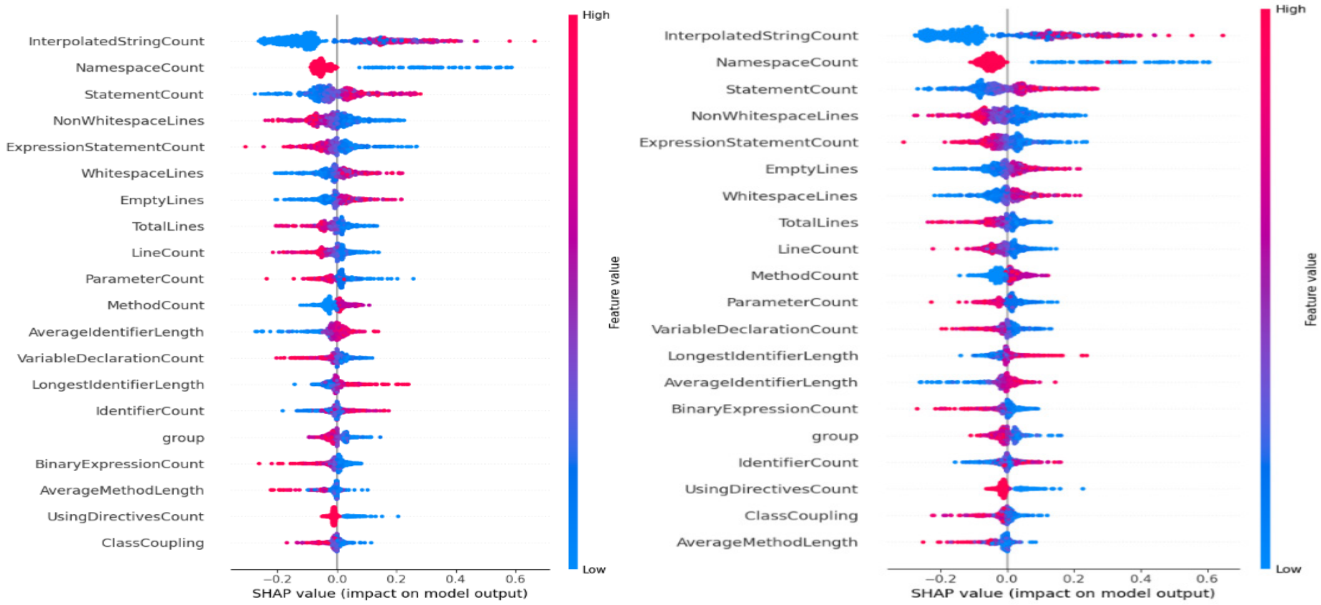


Figure 4: SHAP feature importance for Sets 3 and 4



Comparison with other similar studies

Table 12 compares the classifier accuracy, for detection of GenAI code, that was found in this C#-based study with accuracies detected in similar studies focused on different programming languages. As seen in the table, the approach in this study, which leveraged a comprehensive feature set and advanced classifiers, achieved higher accuracies than the models used in the Bukhari et al. (2023), Idialu et al. (2024), and Pan et al. (2024) studies of GenAI code detection in the Python and C programming languages.

Table 12: Comparison of GenAI code detection accuracy

Study	Programming language	GenAI model(s) used	Classifier(s) used	Classifier accuracy
This study: Adegbite and Kotzé (2025)	C#	Blackbox.AI ChatGPT Microsoft Copilot	XGBoost KNN SVM AdaBoost random forest soft voting	0.97 0.82 0.89 0.96 0.97 0.95 (highest accuracy among the accuracy figures for the 4 sets)
Bukhari et al. (2023)	C	OpenAI code-cushman-001, code-davinci-001, and code-davinci-002	XGBoost KNN SVM random forest	0.92 0.73 0.85 0.90
Idialu et al. (2024)	Python	ChatGPT-4	XGBoost	0.89
Pan et al. (2024)	Python	ChatGPT	GPT Zero GPT-2 Detector DetectGPT GLTR Sapling	0.49 0.50 0.48 0.50 0.60

The feature extraction in the Bukhari et al. (2023) study includes only lexical and syntactic features, while Idialu et al. (2024) add the layout features to the two features considered by Bukhari et al. (2023). Our study focused on a larger list of code stylometry features (comprising 83 lexical, syntactic, layout, and semantic features) than those included by Idialu et al. (2024) and Bukhari et al. (2023), and we can conclude that this wider range of features was integral to the higher classifier accuracy achieved in our study.

5. Conclusions

This study has demonstrated that GenAI C# code produced by Blackbox.AI, ChatGPT, and Copilot can, to a great extent, be identified, and distinguished from student-written C# code, through use of classifier algorithms. The random forest and XGBoost classifiers performed best, with Blackbox.AI C# code being the easiest to detect. This study's focus on the C# programming language helps to fill a research gap, as GenAI code detection in C# is a relatively unexplored area in the education sector in South Africa and globally. This study is also significant in several other respects.

Implications for educators

The study findings also have the potential to assist educational institutions and educators in developing tools for detection of potential use of GenAI code in student assignments. Student use of AI tools for programming and software course assignments can be expected to decrease if detection systems are in place, which in turn will help maintain adherence to academic standards. SHAP identification of features in student assignments can also help to reveal patterns in students' coding behaviours, enabling targeted interventions to improve foundational programming skills. Students can also be encouraged to critically evaluate the strengths and limitations of GenAI code, which will improve their critical thinking skills.

Implications for researchers

Through its use of a large list of code stylometry features (comprising 83 lexical, syntactic, layout, and semantic features), this study has highlighted certain features that were particularly important to the detection of GenAI C# code and to distinguishing between AI-generated and student-written code. Researcher identification of more important features can increase the optimisation of GenAI detection algorithms for programming tasks across varying coding styles and structures. Interdisciplinary studies can follow from this research, as GenAI code detection is at an intersection of NLP, cybersecurity, software engineering, ethics, and education. Future research could incorporate broader sets of coding problems, and broader sources of human-written C# code. The code could be sourced from different educational levels or institutions, as well as from professional developers, to improve the generalisability of the results.

Implications for software developers

Improved detection of GenAI code can help software developers to understand the structure, style and logic of AI-generated code contributions to software. With improved detection and understanding of GenAI code, developers can more easily collaborate in an environment that allows for contributions from both GenAI tools and human developers. Developers can focus more on refining AI contributions, while preserving the nuances of human creativity. Enhanced detection of GenAI code can also help developers to strengthen application security and cybersecurity, particularly with respect to malicious actors who use GenAI to produce scripts used in attacks. When GenAI code is detected early, pre-emptive measures can be put in place to reduce vulnerabilities and safeguard systems against evolving threats.

Limitations of the study

A limitation of this study was that the human-written code dataset was collated from first-year programming students from only one campus of one university: the University of the Free State, South Africa. Thus, this code does not represent the diversity of human-written C# code, which limits the generalisability of the study findings. A larger, more diverse dataset would have provided a better representation of human-written C# code. Furthermore, the nine problems in terms of which the human-written and GenAI C# code was prepared presented potential limitations. The problems could have imposed biases and are unlikely to have fully captured the nuances and complexities of software development, e.g., matters of performance optimisation, security vulnerabilities, maintainability, and real-world applicability.

Funding declaration
No funding was received for this study or for preparation of the article.

Data availability
Data will be made available upon request to the first-listed author at adewuyi.adegbite@gmail.com.

AI declaration
GenAI tools were used for data collection of GenAI C# code, with the versions stated in the article.

Competing interests declaration
The authors have no competing interests to declare.

Authors' contributions
A.A.A.: Conceptualisation; methodology; data collection; sample analysis; data analysis; validation; data curation; writing – initial draft; writing – revisions; student supervision; project management.
E.K.: Conceptualisation; methodology; data collection; writing – revisions; student supervision; project leadership; project management; funding acquisition.

References

Benzebouchi, N. E., Azizi, N., Hammami, N. E., Schwab, D., Khelaifia, M. C. E., & Aldwairi, M. (2019). Authors' writing styles based authorship identification system using the text representation vector. In *16th International Multi-Conference on Systems, Signals and Devices (SSD 2019)* (pp. 371–376). https://doi.org/10.1109/SSD.2019.8894872

Bukhari, S., Tan, B., & De Carli, L. (2023). Distinguishing AI- and human-generated code: A case study. In *SCORED 2023 – Proceedings of the 2023 Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses* (pp. 17–25). https://doi.org/10.1145/3605770.3625215

Caliskan, A., Yamaguchi, F., Dauber, E., Harang, R., Rieck, K., Greenstadt, R., & Narayanan, A. (2018). When coding style survives compilation: De-anonymizing programmers from executable binaries. In *25th Annual Network and Distributed System Security Symposium (NDSS 2018)*. https://doi.org/10.14722/ndss.2018.23304

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. *Journal of the ACM*, *37*(4). http://arxiv.org/abs/2303.04226

Cheers, H., Lin, Y., & Smith, S. P. (2021). Academic source code plagiarism detection by measuring program behavioral similarity. *IEEE Access*, *9*, 50391–50412. https://doi.org/10.1109/ACCESS.2021.3069367

Cheers, H., Lin, Y., & Yan, W. (2023). Identifying plagiarised programming assignments with detection tool consensus. *Informatics in Education*, *22*(1), 1–19. https://doi.org/10.15388/infedu.2023.05

Corso, V., Mariani, L., Micucci, D., & Riganelli, O. (2024). Generating Java methods: An empirical assessment of four AI-based code assistants. In *Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension (ICPC 2024)*. https://doi.org/10.1145/3643916.3644402

Czibula, G., Lupea, M., & Briciu, A. (2022). Enhancing the performance of software authorship attribution using an ensemble of deep autoencoders. *Mathematics*, *10*(15). https://doi.org/10.3390/math10152572

Dehaerne, E., Dey, B., Halder, S., De Gendt, S., & Meert, W. (2022). Code generation using machine learning: A systematic review. *IEEE Access*, *10*(July), 82434–82455. https://doi.org/10.1109/ACCESS.2022.3196347

Ding, S. H. H., Fung, B. C. M., Iqbal, F., & Cheung, W. K. (2019). Learning stylometric representations for authorship analysis. *IEEE Transactions on Cybernetics*, *49*(1), 107–121. https://doi.org/10.1109/TCYB.2017.2766189

Ebrahim, F., & Joy, M. (2023). Source code plagiarism detection with pre-trained model embeddings and automated machine learning. In *International Conference Recent Advances in Natural Language Processing (RANLP)* (pp. 301–309). https://doi.org/10.26615/978-954-452-092-2_034

Eliwa, E., Essam, S., Ashraf, M., & Sayed, A. (2023). Automated detection approaches for source code plagiarism in students' submissions. *Journal of Computing and Communication*, *2*(2), 8–18. https://doi.org/10.21608/jocc.2023.307054

Ghosal, S. S., Chakraborty, S., Geiping, J., Huang, F., Manocha, D., & Bedi, A. S. (2023). Towards possibilities and impossibilities of AI-generated text detection: A survey. arXiv preprint. https://doi.org/10.48550/arXiv.2310.15264

Idialu, O. J., Mathews, N. S., Maipradit, R., Atlee, J. M., & Nagappan, M. (2024). Whodunit: Classifying code as human authored or GPT-4 generated – A case study on CodeChef problems. https://doi.org/10.1145/3643991.3644926

Kalgutkar, V., Kaur, R., Gonzalez, H., Stakhanova, N., & Matyukhina, A. (2019). Code authorship attribution: Methods and challenges. *ACM Computing Surveys*, *52*(1). https://doi.org/10.1145/3292577

Kazemitabaar, M., Ye, R., Wang, X., Henley, A. Z., Denny, P., Craig, M., & Grossman, T. (2024). CodeAid: Evaluating a classroom deployment of an LLM-based programming assistant that balances student and educator needs. In *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3613904.3642773

Kotsiantis, S., Verykios, V., & Tzagarakis, M. (2024). AI-assisted programming tasks using code embeddings and transformers. *Electronics*, *13*(4), 1–25. https://doi.org/10.3390/electronics13040767

Krasniqi, R., & Do, H. (2023). Towards semantically enhanced detection of emerging quality-related concerns in source code. *Software Quality Journal*, *31*(3), 865–915. https://doi.org/10.1007/s11219-023-09614-8

Kuhail A. M., Mathew, S. S., Khalil, A., Berengueres, J., Jawad, S., & Shah, H. (2024). "Will I be replaced?" Assessing ChatGPT's effect on software development and programmer perceptions of AI tools. *Science of Computer Programming*, *235*, 103111. https://doi.org/10.1016/j.scico.2024.103111

Lalitha, L. V. K., Sree, V., Lekha, R. S., & Kumar, V. N. (2021). Plagiat: A code plagiarism detection tool*. EPRA International Journal of Research and Development (IJRD)*, *7838*, 97–101.

Li, Z., Jiang, Y., Zhang, X. J., & Xu, H. Y. (2020). The metric for automatic code generation. *Procedia Computer Science*, *166*, 279–286. https://doi.org/10.1016/j.procs.2020.02.099

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. https://arxiv.org/abs/1705.07874

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, *90*, 46–60. https://doi.org/10.1016/j.futures.2017.03.006

Maryono, D., Yuana, R. A., & Hatta, P. (2019). The analysis of source code plagiarism in basic programming course. *Journal of Physics: Conference Series*, *1193*(1). https://doi.org/10.1088/1742-6596/1193/1/012027

Nghiem, K., Nguyen, A. M., & Bui, N. D. Q. (2024). Envisioning the next-generation AI coding assistants: Insights and proposals. In *2024 First IDE Workshop (IDE '24)*. https://doi.org/10.1145/3643796.3648467

Odeh, A., Odeh, N., & Mohammed, A. S. (2024). A comparative review of AI techniques for automated code generation in software development: Advancements, challenges, and future directions. *TEM Journal*, *13*(1), 726–739. https://doi.org/10.18421/tem131-76

Pan, W. H., Chok, M. J., Wong, J. L. S., Shin, Y. X., Poon, Y. S., Yang, Z., Chong, C. Y., Lo, D., & Lim, M. K. (2024). Assessing AI detectors in identifying AI-generated code: Implications for education. https://arxiv.org/abs/2401.03676

Portillo-Dominguez, A. O, Ayala-Rivera, V., Murphy, E., & Murphy, J. (2017). A unified approach to automate the usage of plagiarism detection tools in programming courses. In *ICCSE 2017 – 12th International Conference on Computer Science and Education*, *ICCSE*, 18–23. https://doi.org/10.1109/ICCSE.2017.8085456

Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, *12*(February), 26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742

ShaukatTamboli, M., & Prasad, R. (2013). Authorship analysis and identification techniques: A review. *International Journal of Computer Applications*, *77*(16), 11–15. https://doi.org/10.5120/13566-1375

Song, X., Sun, H., Wang, X., & Yan, J. (2019). A survey of automatic generation of source code comments: Algorithms and techniques. *IEEE Access*, *7*, 111411–111428. https://doi.org/10.1109/ACCESS.2019.2931579

Srivastava, S., Rai, A., & Varshney, M. (2021). A tool to detect plagiarism in java source code. *Lecture Notes in Networks and Systems*, *145*, 243–253. https://doi.org/10.1007/978-981-15-7345-3_20

Tereszkowski-Kaminski, M., Pastrana, S., Blasco, J., & Suarez-Tangil, G. (2022). Towards improving code stylometry analysis in underground forums. In *Proceedings on Privacy Enhancing Technologies*, *2022*(1), 126–147. https://doi.org/10.2478/popets-2022-0007

Varona, D., & Suárez, J. L. (2022). Discrimination, bias, fairness, and trustworthy AI. *Applied Sciences*, *12*(12). https://doi.org/10.3390/app12125826

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Illia, P. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems* (pp. 5998–6008). https://doi.org/10.48550/arXiv.1706.03762

Wan, Y., He, Y., Bi, Z., Zhang, J., Zhang, H., Sui, Y., Xu, G., Jin, H., & Yu, P. S. (2023). Deep learning for code intelligence: Survey, benchmark and toolkit. *Arxiv.Org*, *1*(1), 771–783. https://arxiv.org/abs/2401.00288

White, J., Hays, S., Fu, Q., Spencer-Smith, J., & Schmidt, D. C. (2023). ChatGPT prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. https://doi.org/10.1007/978-3-031-55642-5_4

Zafar, S., Sarwar, M. U., Salem, S., & Malik, M. Z. (2020). Language and obfuscation oblivious source code authorship attribution. *IEEE Access*, *8*, 197581–197596. https://doi.org/10.1109/ACCESS.2020.3034932

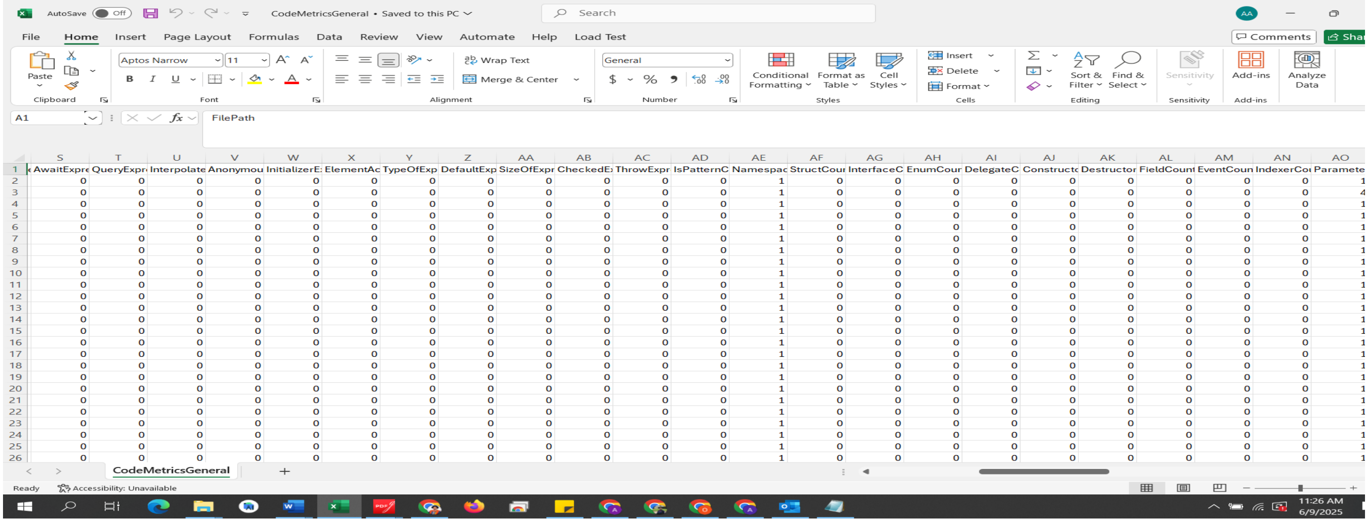
Zhang, H., Cruz, L., & van Deursen, A. (2022). Code smells for machine learning applications. In *Proceedings – 1st International Conference on AI Engineering – Software Engineering for AI (CAIN) 2022* (pp. 217–228). https://doi.org/10.1145/3522664.3528620

Zheng, M., Pan, X., & Lillis, D. (2018). CodEX: Source code plagiarism detection based on abstract syntax trees. *CEUR Workshop Proceedings*, *2259*, 362–373.

Appendix A: The 83 code stylometry features extracted using modified Roslyn

FilePath	UniqueIdentifiers	IfStatementCount	EnumCount
UsingDirectivesCount	StatementCount	AnonymousMethodCount	EventCount
FixedStatementCount	CommentCount	WhileLoopCount	LCOM
UsingStatementCount	MethodInvocationCount	QueryExpressionCount	FieldCount
SwitchStatementCount	ShortestIdentifierLength	AwaitExpressionCount	ClassCount
VariableDeclarationCount	AverageIdentifierLength	DefaultSwitchLabelCount	LineCount
InterpolatedStringCount	AverageMethodLength	LockStatementCount	UsesSpaces
	InitializerExpressionCount	ElementAccessCount	UsesTabs
TypeOfExpressionCount	DefaultExpressionCount	SizeOfExpressionCount	StructCount
CheckedExpressionCount	ThrowExpressionCount	IsPatternCount	TotalLines
NamespaceCount	InterfaceCount	ForEachLoopCount	EmptyLines
DelegateCount	ConstructorCount	YieldBreakCount	ClassCoupling
DestructorCount	ReturnStatementCount	YieldReturnCount	MethodCount
LongestIdentifierLength	ParameterCount	ElseClauseCount	PropertyCount
LocalVariableCount	CyclomaticComplexity	AfferentCoupling	AttributeCount
NestedBlockDepth	DepthOfInheritance	EfferentCoupling	LambdaCount
NonWhitespaceLines	MaxMethodBlockDepth	CaseSwitchLabelCount	TernaryCount
WhitespaceLines	MaxNestedBlockDepth	DoWhileLoopCount	IndexerCount
CommentLines	AverageLineLength	ExpressionStatementCount	ForLoopCount
MinLineLength	MaxLineLength	ObjectCreationCount	IdentifierCount
AssignmentCount	BinaryExpressionCount	LocalFunctionCount	

Appendix B: Snapshot of dataset extracted from modified Roslyn



FilePath

	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO
1	AwaitExpre	QueryExpr	Interpolate	Anonymou	InitializerE	ElementAc	TypeOfExp	DefaultExp	SizeOfExpr	CheckedE	ThrowExpr	IsPatternC	Namespac	StructCour	InterfaceC	EnumCour	DelegateC	Constructo	Destructor	FieldCount	EventCoun	IndexerCo	Paramete
2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	4
4	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
5	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
6	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
8	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
9	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
11	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
12	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
13	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
14	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
15	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
16	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
17	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
18	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
19	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
20	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
21	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
22	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
23	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
24	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
25	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
26	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1

CodeMetricsGeneral

Appendix C: The nine problems used

1.	Develop a C# console application to generate an invoice for the CSI Hoodies company. Collect customer name and full address (street, city, province, postal code). Accept the number of hoodies ordered (whole number). Calculate the total due, including a hardcoded 15% VAT, with a hoodie price of R230. Display a formatted invoice using string.Format() for the address and properly formatted currency values. Clear the console before showing the invoice.
2.	Create a C# program to decide if a car should be sold based on its age and mileage. Input the car's model year and odometer reading (in kilometers) as integers. Sell the car if: Odometer exceeds 100,000 km (regardless of age). Model year is before 2014 (older than 10 years) and after 1950 (not antique). Do not sell if the car is an antique (1950 or earlier) or less than 10 years old (2014 or later). Use one Console.WriteLine() per outcome with newline and tab escape characters, avoiding compound conditions or logical operators.
3.	Build a C# console application named StudentGrades to compute a student's average mark and grade level. Display a title and prompt for three test marks. Calculate the average in one statement, handling integer division. Assign a grade based on the average: A: 80-100 B: 70-79 C: 60-69 D: 50-59 E: Below 50 Use a single Console.WriteLine() to show the result, building a general string and appending the grade dynamically. Add comments to separate code sections.
4.	Write a C# program to find the highest common factor (HCF) of two integers. Accept two positive integers as input. Use a while loop to calculate the HCF by dividing the larger number by the smaller one, updating values with the remainder until it reaches 0; the last non-zero remainder is the HCF. No error checking is required.
5.	Develop a C# console application for a café ordering system. Display a menu of meal items, each with an associated number. Allow the user to select a meal by entering its number and specify the quantity. Display order details: number of meals, price per meal, total price (formatted as currency), and a thank you message. Use a do-while loop to handle multiple orders; exit the program when the user enters -1. Implement three custom static methods: GetInt: Takes a string prompt, displays it, reads user input, and returns it as an integer. TotalPrice: Takes quantity and unit price as parameters, returns the total cost as a decimal. GenerateOrder: Takes a meal price, prompts for quantity using GetInt, calculates the total using TotalPrice, and displays the order details. Use a try-catch block to handle invalid inputs, showing an error message (using an Exception property) and a prompt to retry. Use a switch-case structure to set the price based on the selected meal number and call GenerateOrder.

6.	Create a C# console application named MultiplicationTable to generate multiplication tables. Prompt the user to enter a whole number to specify the multiplication table. Generate and display the table up to the 12th place (e.g., 1 × n to 12 × n) using a while loop. Use string.Format() for aligned output, starting from 1 (not 0). After each table, ask if the user wants to generate another (Y/N), using a do-while loop to repeat the process. Use a char variable for Y/N input and handle both uppercase and lowercase (e.g., with ToUpper() or ToLower()). Exit the program when the user enters 'N.' Use a try-catch block to handle invalid inputs, displaying a custom error message.
7.	Develop a C# console application for a café ordering system. Display a menu of meal items with numbers. Allow the user to select a meal by number and specify the quantity. Display order details: number of meals, price per meal, total price (formatted as currency), and a thank you message. Use a do-while loop to handle multiple orders; exit on -1. Implement custom methods: GetInt, TotalPrice, and GenerateOrder (same as UFSCSI 051). Handle invalid inputs with a try-catch block, showing an error message and retry prompt. Use a switch-case to assign prices and call GenerateOrder.
8.	Develop a C# console application named CompositionOfMoney to break down a monetary amount into the smallest number of coins/notes. Accept and validate a decimal amount using the GetDecimal method (returns a bool and the amount). Convert the amount to cents and use the DisplayUnits method to display the breakdown into units: 1c, 5c, 50c, R1, R10, R100. Loop for multiple conversions using the isAnotherOne method to control repetition. Handle invalid inputs without try-catch.
9.	Create a C# console application named Revision for basic mathematical operations. Show a menu with options: Addition (+), Subtraction (-), Multiplication (*), Division (/). Use a do-while loop to ensure valid operation selection (no if-else for selection). Implement methods for each operation: Addition(): Sum multiple numbers with a while loop. Subtraction(): Subtract two numbers using compound assignment. Multiplication(): Multiply two numbers. Division(): Divide two numbers, handling division by zero with a red error message. Validate numerical inputs with a custom method.

Use of information-fusion deep-learning techniques to detect possible electricity theft: A proposed method

Maria Gabriel Chuwa
PhD candidate, College of Informatics and Virtual Education, University of Dodoma, Tanzania
https://orcid.org/0000-0002-7192-9584

Daniel Ngondya
Lecturer, College of Informatics and Virtual Education, University of Dodoma, Tanzania
https://orcid.org/0000-0003-4267-6351

Rukia Mwifunyi
Lecturer, College of Informatics and Virtual Education, University of Dodoma¸Tanzania
https://orcid.org/0000-0001-7465-3926

Abstract

The performance of electricity utilities in many African countries is undermined by electricity theft. Such non-technical losses (NTLs) pose significant economic challenges to electricity grids, leading to the need for improved detection methods. This study tested an NTL detection method that transformed electricity consumption (EC) profiles into two-dimensional (2D) and one-dimensional (1D) representations, and utilised deep-learning techniques, specifically convolutional neural networks (CNN) and multi-layer perceptron (MLP), to extract features indicating NTLs. This NTL detection method involved three parallel branches: analysing temporal information from application of a Markov transition field (MTF) to EC patterns; analysing spectral information from application of the continuous wavelet transform (CWT) tool; and extracting frequent co-occurrence features from 1D consumption patterns. CNN and MLP were employed within the three branches to capture information from the 2D and 1D inputs, respectively. The features extracted from the three branches were then aggregated through information fusion and applied to EC datasets produced by the State Grid Corporation of China (SGCC) and the Irish Commission for Energy Regulation (CER). This multi-branch approach was found to offer strong NTL detection accuracy. With the SGCC dataset, the method achieved an AUC (area under the curve) of 96.7%, a mAP@100 (mean average precision at 100) of 95.7%, and an FPR (false positive rate) of 8.1%. With the CER dataset, the method achieved an AUC of 96.7%, a mAP@100 of 97.3%, and an FPR of 5.2%.

Keywords

electricity consumption (EC), electricity theft, non-technical losses (NTLs), information fusion, deep-learning, convolutional neural networks (CNN), multi-layer perceptron (MLP), Markov transition field (MTF), continuous wavelet transform (CWT)

DOI: https://doi.org/10.23962/ajic.i35.20652

Recommended citation

Chuwa, M.G., Ngondya, D., & Mwifunyi, R, (2025). Use of information-fusion deep-learning techniques to detect possible electricity theft: A proposed method. *The African Journal of Information and Communication (AJIC), 35*, 1-16. https://doi.org/10.23962/ajic.i35.20652

1. Introduction

In many African countries, the performance and financing of the electricity grid are undermined by high rates of theft. In South Africa, it is estimated that 32% of transmitted electricity is stolen, while in Nigeria, electricity theft is thought to range between 32% to 34% of national transmissions annually (Adongo et al., 2021). Other parts of the world also suffer from high levels of electricity theft. For example, in China, an estimated 16% of generated electricity is stolen, while India and Brazil lose around 25% and 15% of their annual production, respectively (J. Chen et al., 2023). Yan and Wen (2022) point out that electricity theft is one of the main causes of financial difficulties faced by electricity utilities in both the developing and developed world. In addition to the financial impact, electricity theft also leads to risks to public safety, power surges, network damage, and degraded reliability. It is, thus, critical that electricity utilities detect non-technical losses (NTLs) as accurately as possible, so as to be alerted to possible electricity theft (Y. Chen et al., 2023).

In recent years, detection and prevention of electricity theft have received growing attention from researchers and industry practitioners. Conventional machine-learning methods, which rely on feature engineering, have been widely explored and reported to achieve acceptable results in identifying instances of electricity theft (Guarda et al., 2023). These machine-learning-based approaches typically involve extracting a variety of features, such as statistical metrics (e.g., maximum, minimum, mean, and standard deviation), frequency domain characteristics, electricity measurement data (e.g., phase imbalance, power factor), and static information related to geographic location, economic activity, and weather conditions (Chuwa & Wang, 2021). These features are classified using conventional machine-learning algorithms, including support vector machines (SVMs), k-nearest neighbours (KNNs), decision trees, and gradient-boosting methods. For example, Fang et al. (2023) proposed a light gradient-boosting method with 56 statistical features for detection of electricity theft. Zidi et al. (2023) incorporated 10 different electricity features and categorical features to detect theft using five machine-learning techniques: SVMs, KNNs, decision trees, random forest, bagging ensemble, and artificial neural networks (ANNs).

While conventional machine-learning-based methods have demonstrated promising performance in electricity theft detection, they have certain limitations. They primarily depend on human expertise and intervention for crucial feature-extraction and feature-engineering tasks. Handcrafted feature-engineering can lead to important information being missed, thus potentially reducing the effectiveness of conventional machine-learning techniques in accurately detecting electricity theft.

Deep-learning approaches to detecting electrical NTLs

Recently, to address the limitations of conventional machine-learning approaches, researchers have turned to deep-learning methods, which have the ability to extract relevant features automatically. Shi et al. (2023) proposed an approach that uses a transformer neural network (TNN) with a conv-attentional module to extract global and local features. Bai et al. (2023) proposed a hybrid convolutional neural network (CNN)-transformer model to detect electrical NTLs. Their work used a CNN with dual scale and dual branch architecture to extract multi-scale features in a local-to-global fashion. In addition, a transformer model with Gaussian weighting was used to capture the temporal dependence of electricity consumption (EC). In a study by Javaid et al. (2021), CNN, long short-term memory (LSTM), and a deep Siamese network were used to detect electricity theft in smart grids. The study used a CNN model and LSTM to extract features from weekly data and learning sequences from daily data, respectively. At the same time, a deep Siamese network was used to identify similarities between inputs by comparing feature vectors.

All of the aforementioned deep-learning studies extracted features from a 1D representation of the EC patterns. However, it has been found that feature extraction using CNN has a better outcome for 2D data than for 1D data (Nawaz et al., 2023). Extracting patterns directly from 1D time-series data can be difficult due to high variability and the lack of spatial structure (Haq et al., 2023). Encoding energy consumption data into 2D image-like formats enables CNN to capture local and temporal patterns more efficiently through shared convolutional kernels (Massaferro et al., 2022). To enhance the ability of CNN models to capture complex patterns in power consumption data, researchers have explored transforming 1D time-series signals into 2D representations.

For example, Nawaz et al. (2023) proposed a hybrid approach combining CNN with extreme gradient boosting (XGBoost) for electricity theft detection. Their method involved extracting features from energy consumption data in 1D and 2D formats, with weekly consumption data arranged into a matrix for 2D representation. Integrating XGBoost with a wide-and-deep CNN architecture significantly improved detection accuracy for electricity theft. Similarly, studies by Liao et al. (2023) and Xia et al. (2023) have demonstrated strong performance in feature extraction and NTL detection using CNN-based approaches on 2D representations of energy consumption data. Pan et al. (2023) also transformed consumption patterns into 2D image data—using Gramian angular field (GAF), Markov transition field (MTF), and recurrence plot techniques—and combined them into a three-channel image for input into a parallel convolutional neural network (PCNN) model. This architecture enhanced the CNN's capacity to extract robust features from high-dimensional data.

Information fusion

The above studies have demonstrated that representing the EC patterns in 2D formats has significantly enhanced NTL detection accuracy. However, not all characteristics can be adequately captured using one 2D representation alone. Accurately measuring factors such as periodic and recurrent patterns, and transient events, is critical to optimising CNN performance for NTL detection. According to our analysis of data-transformation methods in the existing literature, measuring such factors is most effective through information fusion, specifically by combining MTF and continuous wavelet transform (CWT) with CNN. In addition, our literature analysis found that raw 1D data representations yielded favourable results when processed using a multi-layer perceptron (MLP) model.

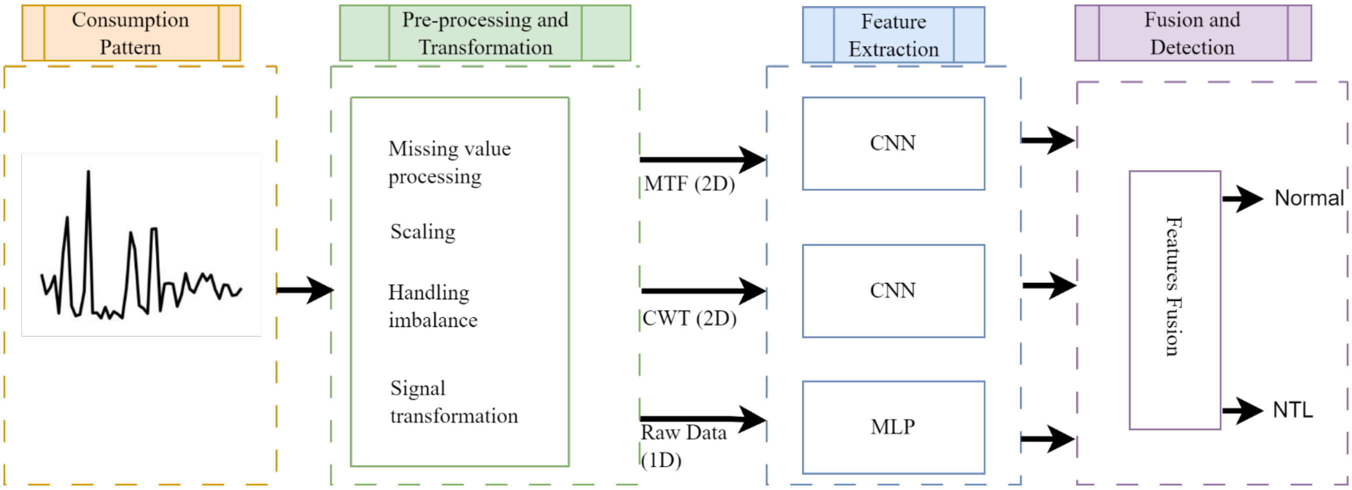
Accordingly, this study tested an information-fusion deep-learning method that extracted features from diverse EC pattern representations (CWT, MTF, and raw data) and fused the obtained features within a classifier for better detection performance. This proposed method was then evaluated using datasets produced by the State Grid Corporation of China (SGCC) and the Irish Commission for Energy Regulation (CER). We found that compared to other existing models, our method achieved superior performance. The main innovations in our proposed method are as follows:

- The method has three input branches: 2D inputs corresponding to CWT and MTF, and a 1D input for the raw EC series. This multi-modal representation captures the temporal, spectral, and periodic information of the adjacent EC pattern segments.
- Not relying on handcrafted features, the method utilises a combination of CNN and MLP deep-learning models to extract information from three input representations of EC. All the features from these three branches were then collated via information fusion using a simple feature concatenation scheme to support final NTL detection.

Framework for the proposed method

The framework for the proposed method consists of the following steps: (1) data pre-processing and transformation; (2) feature extraction and representation from the transformed data samples; (3) feature extraction from different input representations; and (4) fusion of features, by concatenation of features from different modalities, into a single vector for NTL detection. The framework is shown in Figure 1.

Figure 1: Framework



The remainder of this article is organised as follows: Section 2 describes the study's data pre-processing and transformation activities; section 3 sets out the processes for feature extraction and fusion; section 4 presents the findings from evaluation of our proposed method; and section 5 provides conclusions.

2. Data pre-processing and transformation

Data pre-processing and transformation were fundamental to creating a usable data structure—a structure that enabled the proposed model to be trained and to generate reliable predictions.

CER dataset

The CER dataset (CER, 2012) provided by the Irish Social Science Data Archive (ISSDA) comprises EC data from over 5,000 residential and commercial electricity users. The data was recorded at half-hour intervals between July 2009 and December 2010. All customers for this dataset were considered legitimate, with no illegal electricity users. From the CER data samples, we generated attack samples. The six attack scenarios indicated in Table 1, defined by Jokar et al. (2016), were used to create attacks. The attack samples were generated from only 10% of the available load profiles by randomly selecting a subset of users and their load profiles. Each of the six attack-generation methods was applied to the selected users' load profiles.

Table 1: Attacks and their definitions (from Jokar et al., 2016)

	Attack	Definition
1	$m_1(t) = \alpha e_t$ $0 < \alpha < 1$	Report a constant fraction of the energy consumed.
2	$m_2(t) = \beta_t e_t$ $\beta_t = \begin{cases} 0, & t_s < t < t_e \\ 1, & \text{else} \end{cases}$ where, t_s is the time when the attack starts t_e is the time when the attack ends $t_e - t_s$ is randomly defined each day	Report zero consumption at randomly defined times of the day.
3	$m_3(t) = \gamma_t e_t$ $0 < \gamma_t < 1$	Reduce consumption patterns by reporting less from time to time.
4	$m_4(t) = \gamma_t \cdot \text{mean}(e_t)$ $0 < \gamma_t < 1$	Report the consumption with reduced expected mean from time to time.
5	$m_5(t) = \text{mean}(e_t)$	Report constant consumption, which is the mean of day consumption.
6	$m_6(t) = e_{p-t}$ p, is the total number of samples in a period to be reversed	Reverse the order of measured values.

SGCC dataset

The SGCC dataset from China contains the daily recorded EC data of 42,372 electricity customers from January 2014 to October 2016 (SGCC, n.d.). The dataset is labelled and includes 3,615 real-world NTL scenarios.

Handling missing values

The pre-processing of real-world datasets often requires addressing missing or erroneous data. This study used the linear interpolation method described by Zheng et al. (2018) to estimate missing EC samples. This method is useful for time-series data as it captures the relationship between adjacent variables. Equation 1 below presents the mathematical formula of this method, where NaN represents a missing value, and x_i is consumption at time *i*.

$$x_i = \begin{cases} \frac{x_{i-1}+x_{i+1}}{2} & \text{if } x_i = \text{NaN}, x_{i-1}, x_{i+1} \neq \text{NaN} \\ 0 & \text{if } x_i = \text{NaN}, x_{i-1} \text{ or } x_{i+1} = \text{NaN} \\ x_i & \text{if } x_i \neq \text{NaN} \end{cases} \tag{1}$$

This study employed the three-sigma rule to systematically identify and rectify erroneous data samples. Observations were deemed to be outliers if they deviated beyond two standard deviations (±2σ) from the mean of the data vector. Equation 2 presents the mathematical expression for correcting the erroneous data samples (Khan et al., 2020). In this Equation $\bar{x}$ and σ_x represent the mean and standard deviation, respectively, of the consumption vector.

$$g(x_i) = \begin{cases} \overline{x} + 2\sigma_x, & \text{if } x_i > \overline{x} + 2\sigma_x \\ x_i, & \text{otherwise} \end{cases} \tag{2}$$

Data scaling

Since EC differs among customers, a min-max scaler was used to normalise the data, ensuring the data were on the same scale. The scaling process improves model performance and convergence while preventing bias from features with larger values. The min-max scaling is mathematically represented using Equation 3, where x_{min} and x_{max} represent the lowest and highest values in the data, respectively.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

Handling imbalance

One of the significant challenges in machine-learning models is the imbalance of data collected from smart meters. Typically, there are few records for classes with NTL, which leads to difficulties in training robust models. For example, in the SGCC dataset, ≈9% of the data is labelled as theft scenarios. The class imbalance can reduce classification accuracy and create a bias towards the majority class. It is essential to balance the class distributions within the dataset to address the issues associated with class imbalance before training an NTL-detection model. In this study, we employed the synthetic minority over-sampling technique (SMOTE) proposed by Chawla et al. (2002). SMOTE addresses class imbalance by generating new samples and effectively handling imbalance for electricity theft detection (Pereira & Saraiva, 2021). It selects a minority sample and identifies neighbouring samples, and then creates synthetic instances through interpolation between these chosen samples.

Data transformation

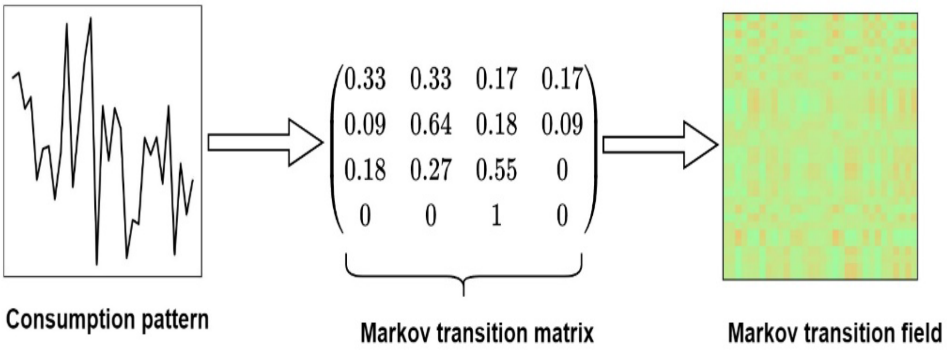
Our model processed EC data in multiple representations in order to capture diverse patterns within the data. A 2D structural representation was chosen to expose hidden temporal and spectral features suitable for CNN analysis. The preliminary experiments indicated that MTF (temporal representation) and CWT (time-frequency representation) yielded superior performance in detecting NTL compared to alternative 2D representations. Simultaneously, a raw 1D representation, which exposed frequent co-occurrence features of the EC data, demonstrated enhanced performance in NTL detection when analysed using MLP. Therefore, this study represented EC patterns using MTF, CWT and 1D raw representations in order to comprehensively capture the diverse characteristics of the EC data.

MTF

The Markov transition field visualisation technique transforms 1D time series data into a 2D image representation while preserving the information in the time domain. This transformation captures the first-order Markov transition probabilities among defined states, enhancing the ability to detect anomalies. For a consumption pattern denoted as $c_t = \{c_1, c_2, c_3, \cdots, c_n\}$, state is identified, and each value c_t is allocated to a corresponding state s_j $(j \in [1,S])$. The Markov transition matrix M is constructed by calculating the frequency of transitions between these states, where p_{ij} of transitioning from state s_i to state s_j. This transition matrix, shown in Equation 4, highlights the relationships between data points in c_t and serves as a foundation for detecting anomalies that indicate electricity theft (Wang & Oates, 2015). The 2D representation facilitates the identification of unusual patterns and fluctuations that may signal fraudulent activity. Figure 2 illustrates the process of transforming the time series into an MTF.

$$M = \begin{vmatrix} p_{ij}|c_1 \in s_i, c_1 \in s_j & p_{ij}|c_1 \in s_i, c_2 \in s_j & \cdots & p_{ij}|c_1 \in s_i, c_n \in s_j \\ p_{ij}|c_2 \in s_i, c_1 \in s_j & p_{ij}|c_2 \in s_i, c_2 \in s_j & \cdots & p_{ij}|c_2 \in s_i, c_n \in s_j \\ \vdots & \vdots & \ddots & \vdots \\ p_{ij}|c_n \in s_i, c_1 \in s_j & p_{ij}|c_n \in s_i, c_2 \in s_j & \cdots & p_{ij}|c_n \in s_i, c_n \in s_j \end{vmatrix} \tag{4}$$

Figure 2: Process of transforming electricity consumption series to MTF



CWT

The continuous wavelet transform tool is a powerful approach to analysing time signals that provides a time–frequency representation. The ability of CWT to analyse signals with time-varying characteristics makes it ideal for detecting inconsistencies in EC patterns that might indicate theft. Electricity theft often manifests as unusual periodicities that are not reflected in the normal usage profile, and traditional time-domain or frequency-domain analyses struggle to capture this variation effectively. However, the CWT is highly effective at pinpointing these variations in time and frequency, enabling accurate identification. Consider a consumption pattern represented by the discrete sequence, $c_t = \{c_1, c_2, c_3, \cdots, c_n\}$ and a wavelet function $\Psi(t)$. Then, the CWT is defined as a convolution between and wavelet , as expressed by Equation 5 (from Boashash, 2009).

$$CW(\tau, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} c_t \Psi^* \left(t - \frac{\tau}{a}\right) dt \tag{5}$$

where * denotes the complex conjugate of Ψ, τ represents the translation parameter controlling the wavelet's position in time, and $a = \omega_0/\omega$ is the scale parameter that controls the stretching of wavelets in time, narrowing it for large frequencies and widening it for small frequencies.

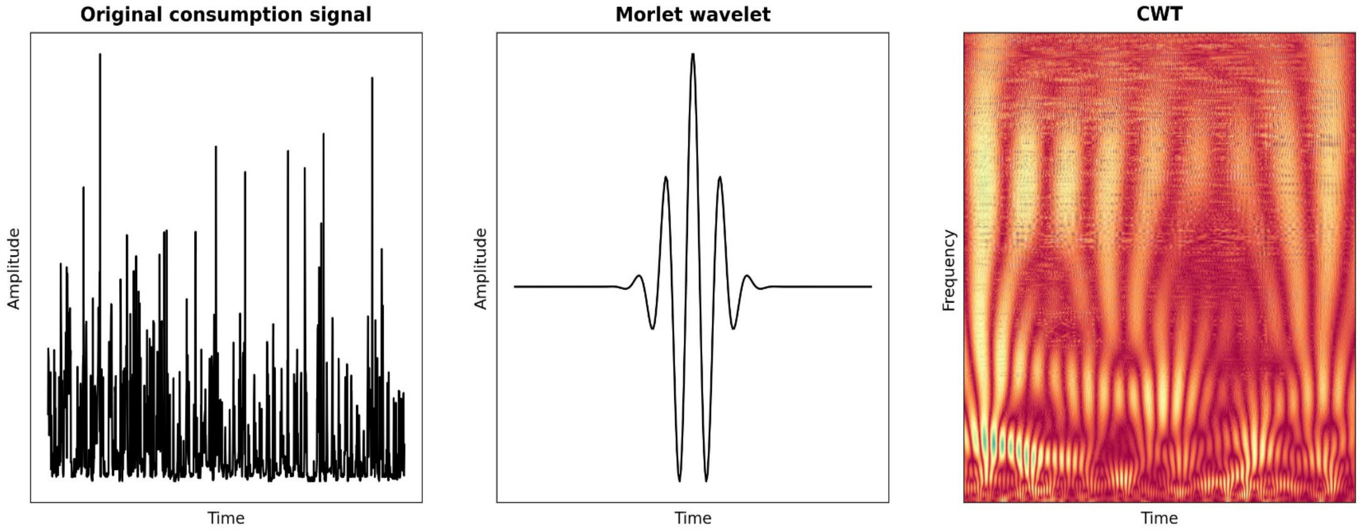
For a wavelet to be valid, it must have zero mean and be concentrated in both the time and frequency domains. A commonly used wavelet for spectral analysis is the Morlet wavelet, which we used in this study. It is defined in Equation 6 (V. C. Chen & Ling, 2002).

$$\Psi(t) = \pi^{-\frac{1}{4}} e^{-\frac{t^2}{2}} e^{-i\omega_0 t} \tag{6}$$

where ω_0 is central frequency.

The choice of ω_0 influences the time-frequency resolution of the analysis. A higher ω_0 provides better frequency resolution at the expense of time resolution. Figure 3 shows analysis of EC patterns using CWT with the Morlet wavelet.

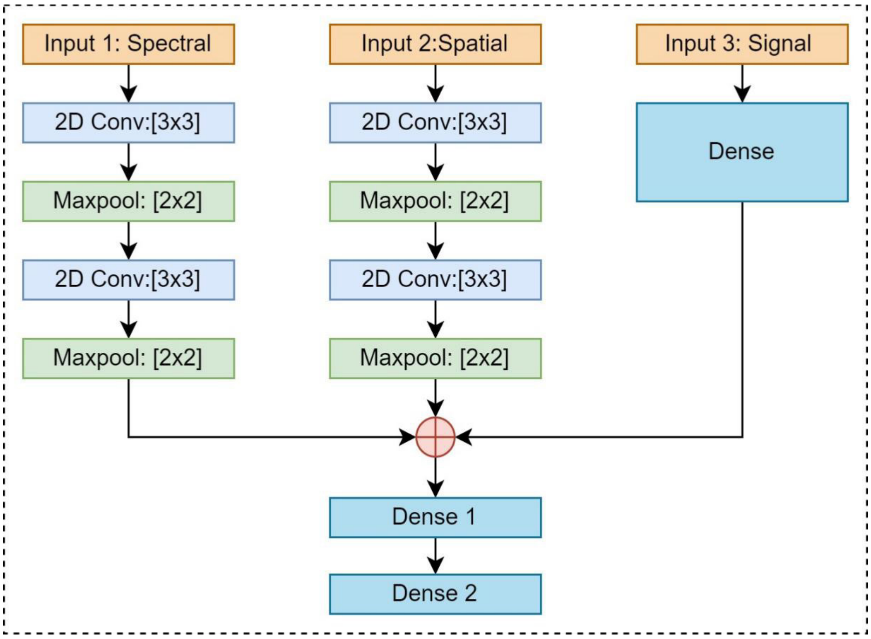
Figure 3: Analysis of EC using CWT with the Morlet wavelet



3. Feature extraction and fusion

We used a deep neural network architecture (CNN and MLP) that allowed the model to learn from different types of data representations of the EC data: temporal (2D), spectral (2D), and raw data (1D), as presented in Figure 4.

Figure 4: Architecture of the joint feature extraction and classification model



The deep CNN feature-extraction component was constructed by two blocks of convolutional layers and two max-pooling layers. The convolutional layers learned to detect patterns and extract meaningful features from the 2D inputs. The convolution layer extracted features from the input by sliding multiple kernels (filters) over the input, generating feature maps that captured important spatial information.

The output of the convolutional operation can be expressed mathematically, as shown in Equation 7, where f_{ReLu} is the rectified linear unit (ReLU) activation function presented by Equation 8.

$$y^{c}_{(a,b)} = f_{\text{ReLU}}\left(b + \sum_{i}\sum_{j} X_{(i,j)} w_{(a-i,b-j)}\right) \tag{7}$$

$$f_{\text{ReLU}}(s_i) = \max(0, s_i) \tag{8}$$

After applying convolution and activation functions, the pooling process was applied to the feature maps. The pooling layers helped to reduce the dimensionality of the feature maps. We used the max-pooling operation, which takes the maximum value within a window in a feature map and is expressed by Equation 9.

Where y_i^p is the maximum value of in a window of size (m,n),

$$y^{p}_{i} = \max\left(y^{c}_{(m,n)}\right) \tag{9}$$

In the CNN feature-extraction component, convolutional and pooling operations worked alternately to capture the features from the two 2D representations of the EC patterns. Equations 10 and 11 express the overall process of feature extraction using convolution and pooling operations.

$$Y_{MTF} = y^{p}\left(y^{c}\left(y^{p}\left(y^{c}\left(X_{MTF}\right)\right)\right)\right) \tag{10}$$

$$Y_{CWT} = y^{p}\left(y^{c}\left(y^{p}\left(y^{c}\left(X_{CWT}\right)\right)\right)\right) \tag{11}$$

Further, the raw 1D input representation was passed through a dense network, as expressed in Equation 12. The dense network identified other information within the raw input that complemented the features extracted from the CWT and MTF inputs.

$$Y_{Raw} = f_{\text{ReLU}}\left(b + \sum_{i} X_{Raw_i} w_i\right) \tag{12}$$

After the feature extraction stage, the feature maps from the CNN of CWT and MTF inputs, and the dense representation for the raw input, were concatenated into a single feature vector, as expressed in Equation 13.

$$Y^{t} = Y^{l}_{MTF} \oplus Y^{m}_{CWT} \oplus Y^{n}_{Raw} \tag{13}$$

This feature-fusion step integrated the information from the different input representations, enabling the model to take advantage of various aspects of the data. The fused-feature vector was then passed through additional dense layers to further capture and learn the interactions between the combined feature representations.

4. Findings from evaluation of the proposed method

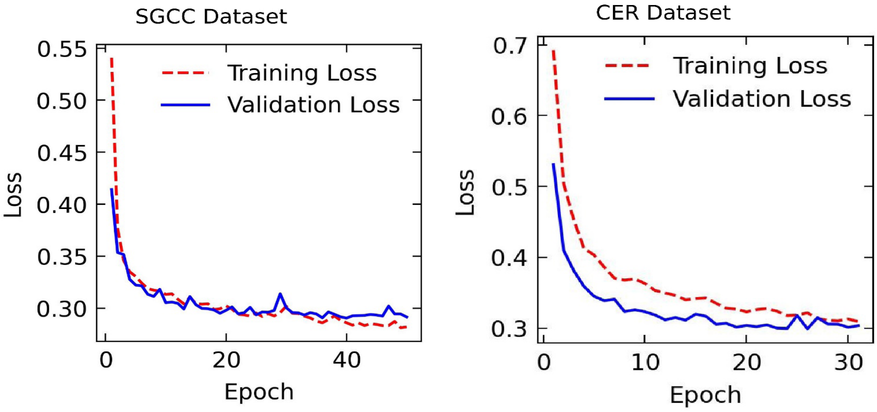
Training and validation

The multi-modal deep-learning architecture took three different types of inputs: a raw signal input, a 32x32 single-channel image input, and another 32x32 single-channel image input. The signal input was passed through a dense layer with 64 ReLUs (rectified linear units), while each image input went through a series of 2D convolutional and max-pooling layers to extract spatial features. The convolutional layers had 32 and 64 filters with 3x3 and 5x5 kernel sizes and ReLU activations. The max-pooling layers reduced the spatial dimensions of the feature maps.

Following the feature extraction for each modality, the outputs were concatenated into a single-feature vector. A dropout layer with a rate of 0.2 was then applied to the combined features to improve generalisation. The fused features vector was then passed through a dense layer of 128 units with ReLU activations. The final output layer utilised a softmax activation function to generate probability estimates for the two classes. The model parameters of all layers were then initialised randomly and trained by a back-propagation algorithm with ADAM (adaptive moment estimation). The ADAM minimises the loss function and updates the parameters during training to achieve effective model convergence.

Figure 5 provides the training loss and validation loss curves for the two datasets: SGCC and CER. Both datasets showed a general downward trend in training loss, indicating successful learning. However, the validation loss curves differed. The SGCC dataset exhibited less overfitting with a smaller gap between training and validation loss, suggesting better generalisation. Meanwhile, the CER dataset showed a more erratic validation loss curve, indicating potential difficulties in generalisation.

Figure 5: Training and validation loss curves for SGCC and CER datasets



Evaluation metrics

To enable a comprehensive performance evaluation, the study deployed widely used performance evaluation metrics, specifically area under the curve (AUC), mean average precision at M (MAP@M), and false positive rate (FPR). AUC measured the effectiveness of the method in distinguishing positive and negative instances. A high AUC would indicate the method's ability to effectively differentiate between classes and correctly identify the NTL cases. The MAP@M was used to assess the quality of the proposed method by evaluating its ability to identify NTLs among the top M electricity consumers. The formula for calculating MAP@M is illustrated in Equation 14, where P_i represents the precision of correctly identified NTL at a position , and denotes the total number of NTL samples among M labels.

$$MAP@M = \frac{1}{m}\sum_{i=1}^{m} P_i \tag{14}$$

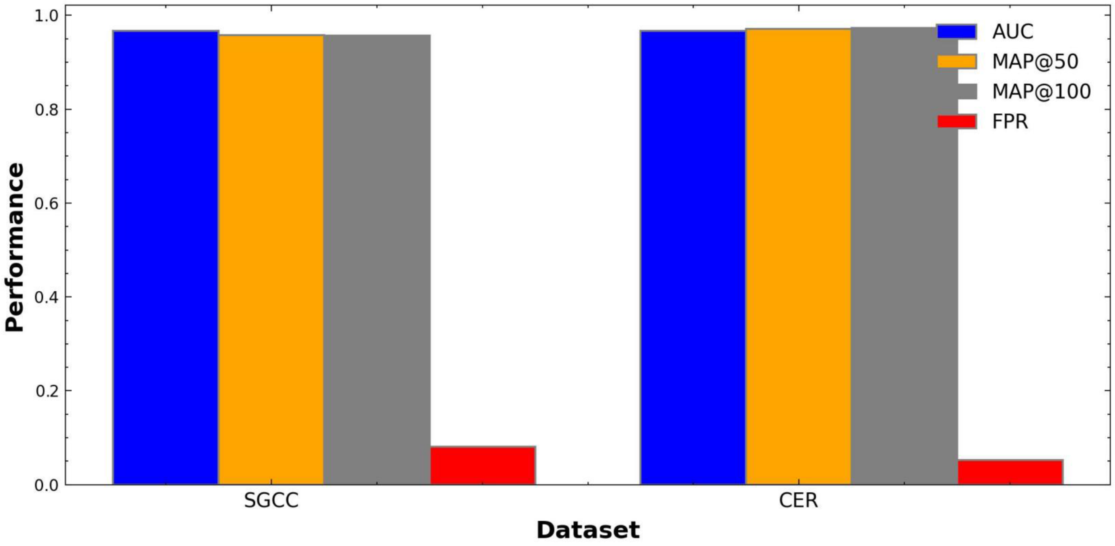
FPR represents the proportion of normal customers that the method incorrectly classified as abnormal, as defined in Equation 15, where FP and TN are the number of false positives and true negatives, respectively. A low FPR indicates good detection performance.

$$FPR = \frac{FP}{FP + TN} \tag{15}$$

Evaluation results

Figure 6 provides the results from the experimental evaluation of the proposed method on the SGCC and CER datasets. Four metrics, AUC, MAP@50, MAP@100, and FPR were used to assess the proposed model. As seen in the Figure 6, the method achieved impressive results for the CER dataset, with a MAP@50 of 97.1%, a MAP@100 of 97.3%, an AUC of 96.7%, and an FPR of 5.2%. These values indicate a strong ability to identify NTLs accurately. The testing of the method with the SGCC dataset yielded slightly weaker (but still strong) metrics, with a MAP@50 of 95.82%, a MAP@100 of 95.65%, an AUC of 96.7%, and an FPR of 8.1%.

Figure 6: The proposed method's AUC, MAP and FPR results



The method's achievement of high MAP and AUC values with both datasets indicated the effectiveness of the proposed method in detecting NTLs. In addition, the low FPR on both datasets indicated the proposed method's capacity to minimise false positives. The consistent performance across both datasets highlighted the robustness of the proposed method. The small variations between the results for the two datasts can be attributed to differences in dataset characteristics.

Performance comparison between information-fusion and stand-alone representations

To further evaluate the effectiveness of the proposed method's NTL information-fusion of features from different representations, we compared the performance of the proposed method with the stand-alone performance of each of the individual representations. The results in Tables 2 and 3 demonstrate that our proposed method's integration of features from several representations significantly improves NTL detection performance when compared with the performance of individual representations.

Table 2 shows the results for the CER dataset, with the results indicating an increase in AUC of approximately 1.4% to 3% when fusing features from CWT, MTF and raw representations compared to using them individually. Also, FPR decreases significantly, from 0.107 with raw features to 0.052 when using fused features. With respect to computational efficiency, the fused model's training time (29.41 secs) was faster than the training times for both MTF and CWT, but slower than than training time (18.48 secs) for the raw 1D representation.

Table 2: NTL-detection performance comparison on CER dataset

Metrics	Raw	CWT	MTF	Fused features
AUC	0.94	0.953	0.937	**0.967**
MAP@100	0.954	0.965	0.969	**0.973**
FPR	0.107	0.073	0.071	**0.052**
Time (sec)	**18.48**	37.81	34.67	29.41

Table 3 presents the results for the SGCC dataset. Again the findings reveal a substantial improvement in detection performance, with AUC values increasing by 7.7% to 14% when transitioning from individual representations to combined features, and with FPR dropping from 0.274 to 0.081. With respect to computational efficiency, the fused model's training time (185.82 secs) was faster than the training times for CWT, but slower than training times for MTF (184.42) and for the raw 1D representation (102.51). It is worth noting that the fused-features approach required longer training times than the raw 1D representation for both datasets. However, these times remained significantly lower than those observed for CWT and MTF individually, suggesting that the fusion process optimised computational efficiency despite its added complexity.

Table 3: NTL detection performance comparison on SGCC dataset

Metrics	Raw	CWT	MTF	Fused features
AUC	0.827	0.89	0.874	**0.967**
MAP@100	0.87	0.894	0.888	**0.957**
FPR	0.274	0.218	0.253	**0.081**
Time (sec)	**102.51**	188.05	184.42	185.82

Performance comparison with existing methods

Furthermore, we compared performance of our proposed method against the performance of other methods applied to the SGCC and CER datasets. Zheng et al. (2018) implemented a wide and deep CNN using consumption patterns represented as 1D and 2D with the SGCC dataset. Also with the SGCC dataset, Shehzad et al. (2022) employed a SVM model that applied 11 features derived from the consumption pattern as the input. With EC represented as 2D matrices derived from monthly consumption data, Massaferro et al. (2022) utilised CNN multi-resolution with the CER dataset; Nawaz et al. (2023) deployed a CNN with XGBoost for the SGCC dataset; and Xia et al. (2023) used CNN with the SGCC dataset. Bastos et al. (2023) proposed an ensemble model combining time series forest, residual network, inception time, time Le-Net, and multi-channel deep CNN, all trained on a 1D EC pattern. Since these studies used datasets similar to ours, we adopted their reported performances for initial comparison.

As shown in Table 4, our method's performance was significantly better than that of even the strongest models discussed in the literature on use of the SGCC and CER datastes, namely the Zheng et al. (2018) wide and deep CNN method, which achieved a MAP@100 of 0.95 with the SGCC dataset, and the Shehzad et al. (2022) SVM method, which achieved an AUC of 0.91 with the CER dataset.

Table 4: Method performance comparison (NTL detection in SGCC and CER datasets)

Method	Dataset	Input(s)	Metrics		
			AUC	FPR	MAP@100
wide and deep CNN (Zheng et al., 2018)	SGCC	1D and 2D	0.782		0.95
SVM (Shehzad et al., 2022)	SGCC	11 features	0.91		
CNN multi-resolution (Massaferro et al., 2022)	CER	2D	0.86		
CNN + XGBoost (Nawaz et al., 2023)	SGCC	1D and 2D	0.54		
CNN (Xia et al., 2023)	SGCC	1D and 2D	0.836		0.951
Ensemble (TSF, ResNet, Inception time, time-Le-Net, MCDCNN) (Bastos et al., 2023)	CER	1D		0.016	
Our proposed method	SGCC	Fused 1D and 2D	**0.967**	0.081	**0.957**
	CER	Fused 1D and 2D	0.967	0.052	**0.971**

Comparison with baseline classifiers using handcrafted features

To further assess the advantages of automatic feature-learning, we compared the detection performance of our proposed method with baseline models trained on handcrafted features. The baseline models we used for the comparison were k-nearest neighbour (KNN), decision tree (DT), random forest (RF), and an SVM model. The input features for these models consisted of five handcrafted attributes per consumption pattern: four statistical measures (mean, standard deviation, variance, skewness) and one frequency-domain feature (spectral centroid) extracted from raw EC time series. Table 5 summarises the configurations and key parameters selected for model training for each baseline classifier.

Table 5: Configuration of baseline classifier models

Model	Key configuration
KNN	Number of neighbours k = 10
DT	Criterion= Gini
RF	Number of trees = 100, criterion = Gini,
SVM	Kernel = RBF, C = 1.0
Common settings	5-fold cross-validation; missing values handled using KNN imputation (k=5)

Table 6 shows the apparent advantages of our deep-learning approach (using CNN and MLP) over the handcrafted feature-engineering used in baseline models. The results show that our proposed method's achievement of an AUC of 0.967 on both datasets markedly surpassed the performance of KNN, decision tree, random forest and SVM, all of which yielded lower AUCs ranging from 0.50 to 0.84. Our proposed method also demonstrated a low FPR and a high MAP@100, indicating its ability to minimise erroneous predictions and to prioritise relevant predictions at the top of the ranked list.

Table 6: Method performance comparison (NTL detection in SGCC and CER datasets)

Baseline Method	Dataset	Inputs	AUC	FPR	MAP@100
KNN	SGCC	Handcrafted features	0.579	0.301	0.723
	CER	Handcrafted features	0.729	0.848	0.714
DT	SGCC	Handcrafted features	0.527	0.381	0.684
	CER	Handcrafted features	0.704	0.607	0.538
RF	SGCC	Handcrafted features	0.624	0.326	0.91
	CER	Handcrafted features	0.842	0.859	0.796
SVM	SGCC	Handcrafted features	0.504	0.529	0.679
	CER	Handcrafted features	0.769	0.859	0.796
Our proposed method	SGCC	Fused 1D and 2D	**0.967**	0.081	**0.957**
	CER	Fused 1D and 2D	**0.967**	0.052	**0.971**

5. Conclusions

This study has proposed and evaluated an information-fusion approach to deep-learning NTL detection in electricity grids. The key innovation of the proposed method is its ability to take advantage of various representations of EC patterns and enhance the feature-extraction capabilities of deep-learning models. The proposed model has three parallel branches that simultaneously analyse: temporal information from the MTF representation; spectral information from the CWT representation; and frequently recurring patterns in the 1D representation of raw EC data. By integrating these diverse representations, the model can sufficiently capture temporal, spectral, and periodicity information without relying on handcrafted features. Moreover, the proposed method employs deep CNN to extract features from 2D representations (using MTF and CWT) while utilising MLP to extract features from the raw 1D representation of EC data. Through our experiments on real-world datasets provided by the SGCC and CER, we found the proposed model demonstrates better NTL performance than that found in similar studies using the the same datasets. The performance and efficiency of our proposed information-fusion deep-learning network suggest a promising response to electrical utilities' need to to improve NTL detection and, in turn, to limit their grids' performance and financial losses.

Acknowledgements

We are grateful to the Irish Social Science Data Archive (ISSDA) for providing access to the CER Smart Metering Project – Electricity Customer Behaviour Trial (2009-2010) dataset, and the Kaggle repository for providing access to the State Grid Corporation of China (SGCC) dataset.

Data availability

The study used two publically available datasets to support the findings. The Irish dataset is available from the ISSDA at https://www.ucd.ie/issda/data/commissionforenergyregulationcer/ with reference number 0012-00. The SGCC dataset is available via the Kaggle repository at https://www.kaggle.com/datasets/bensalem14/sgcc-dataset.

AI declaration

While preparing this work, the authors used Grammarly and online AI-assisted technologies to check grammar and spelling. After using the tools, the authors reviewed and edited the content as required and take full responsibility for the content of the publication.

Competing interests declaration

The authors have no competing interests to declare.

Authors' contributions

MGC: Conceptualisation, methodology, data collection, data analysis, validation, data curation, writing the initial draft.
DN: Student supervision, writing – revisions.
RM: Student supervision, writing – revisions.
All authors read and approved the final manuscript.

References

Adongo, C. A., Taale, F., Bukari, S., Suleman, S., & Amadu, I. (2021). Electricity theft whistleblowing feasibility in commercial accommodation facilities. *Energy Policy*, *155*, 112347. https://doi.org/10.1016/j.enpol.2021.112347

Bai, Y., Sun, H., Zhang, L., & Wu, H. (2023). Hybrid CNN-transformer network for electricity theft detection in smart grids. *Sensors*, *23*(20), 1–21. https://doi.org/10.3390/s23208405

Bastos, L., Pfeiff, G., Oliveira, R., Oliveira, H., Tostes, M. E., Zeadally, S., Cerqueira, E., & Rosário, D. (2023). Data-oriented ensemble predictor based on time series classifiers for fraud detection. *Electric Power Systems Research*, *223*(July), 109547. https://doi.org/10.1016/j.epsr.2023.109547

Boashash, B. (2009). Time-frequency and instantaneous frequency concepts. In *Time-frequency signal analysis and processing: A comprehensive reference* (pp. 31–61). Academic Press.

Chawla, N. V, Bowyer, K. W., & Hall, L. O. (2002). SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, J., Nanehkaran, Y. A., Chen, W., Liu, Y., & Zhang, D. (2023). Data-driven intelligent method for detection of electricity theft. *International Journal of Electrical Power and Energy Systems*, *148*(September), 108948. https://doi.org/10.1016/j.ijepes.2023.108948

Chen, V. C., & Ling, H. (2002). *Time-frequency transforms for radar imaging and signal analysis*. Artech House.

Chen, Y., Li, J., Huang, Q., Li, K., Zhao, Z., & Ren, X. (2023). Non-technical losses detection with Gramian angular field and deep residual network. *The 3rd International Conference on Power and Electrical Engineering, Energy Reports*, *9*, 1392–1401. https://doi.org/10.1016/j.egyr.2023.05.183

Chuwa, M. G., & Wang, F. (2021). A review of non-technical loss attack models and detection methods in the smart grid. *Electric Power Systems Research*, *199*, 107415. https://doi.org/10.1016/J.EPSR.2021.107415

Commission for Energy Regulation (CER). (2012). *CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010* [Dataset]. Irish Social Science Data Archive. SN: 0012-00. https://www.ucd.ie/issda/accessdata/issdadatasets

Fang, H., Xiao, J. W., & Wang, Y. W. (2023). A machine learning-based detection framework against intermittent electricity theft attack. *International Journal of Electrical Power and Energy Systems*, *150*(February), 109075. https://doi.org/10.1016/j.ijepes.2023.109075

Guarda, F. G. K., Hammerschmitt, B. K., Capeletti, M. B., Neto, N. K., dos Santos, L. L. C., Prade, L. R., & Abaide, A. (2023). Non-hardware-based non-technical losses detection methods: A review. *Energies*, *16*(4), 1–27. https://doi.org/10.3390/en16042054

Haq, E. U., Pei, C., Zhang, R., Jianjun, H., & Ahmad, F. (2023). Electricity-theft detection for smart grid security using smart meter data: A deep-CNN based approach. *Energy Reports*, *9*, 634–643. https://doi.org/10.1016/j.egyr.2022.11.072

Javaid, N., Jan, N., & Javed, M. U. (2021). An adaptive synthesis to handle imbalanced big data with deep siamese network for electricity theft detection in smart grids. *Journal of Parallel and Distributed Computing*, *153*, 44–52. https://doi.org/10.1016/j.jpdc.2021.03.002

Jokar, P., Arianpoo, N., & Leung, V. C. M. (2016). Electricity theft detection in AMI using customers' consumption patterns. *IEEE Transactions on Smart Grid*, *7*(1), 216–226. https://doi.org/10.1109/TSG.2015.2425222

Khan, Z. A., Adil, M., Javaid, N., Saqib, M. N., Shafiq, M., & Choi, J. G. (2020). Electricity theft detection using supervised learning techniques on smart meter data. *Sustainability*, *12*(19), 1–25. https://doi.org/10.3390/su12198023

Liao, W., Yang, Z., Liu, K., Zhang, B., Chen, X., & Song, R. (2023). Electricity theft detection using Euclidean and graph convolutional neural networks. *IEEE Transactions on Power Systems*, *38*(4), 3514–3527. https://doi.org/10.1109/TPWRS.2022.3196403

Massaferro, P., Di Martino, J. M., & Fernandez, A. (2022). Fraud detection on power grids while transitioning to smart meters by leveraging multi-resolution consumption data. *IEEE Transactions on Smart Grid*, *3053*(c), 1–10. https://doi.org/10.1109/TSG.2022.3148817

Nawaz, A., Ali, T., Mustafa, G., Rehman, S. U., & Rashid, M. R. (2023). A novel technique for detecting electricity theft in secure smart grids using CNN and XG-boost. *Intelligent Systems with Applications*, *17*, 200168. https://doi.org/10.1016/j.iswa.2022.200168

Pan, H., Feng, X., Na, C., & Yang, H. (2023). A model for detecting false data injection attacks in smart grids based on the method utilized for image coding. *IEEE Systems Journal*, *17*(4), 6181–6191. https://doi.org/10.1109/JSYST.2023.3287924

Pereira, J., & Saraiva, F. (2021). Convolutional neural network applied to detect electricity theft: A comparative study on unbalanced data handling techniques. *International Journal of Electrical Power and Energy Systems*, *131*(March), 107085. https://doi.org/10.1016/j.ijepes.2021.107085

Shehzad, F., Javaid, N., Aslam, S., & Umar Javaid, M. (2022). Electricity theft detection using big data and genetic algorithm in electric power systems. *Electric Power Systems Research*, *209*(October 2021), 107975. https://doi.org/10.1016/j.epsr.2022.107975

Shi, J., Gao, Y., Gu, D., Li, Y., & Chen, K. (2023). A novel approach to detect electricity theft based on conv-attentional Transformer Neural Network. *International Journal of Electrical Power and Energy Systems*, *145*(February), 10862. https://doi.org/10.1016/j.ijepes.2022.108642

Wang, Z., & Oates, T. (2015). *Spatially encoding temporal correlations to classify temporal data using convolutional neural networks*. http://arxiv.org/abs/1509.07481

Xia, R., Gao, Y., Zhu, Y., Gu, D., & Wang, J. (2023). An attention-based wide and deep CNN with dilated convolutions for detecting electricity theft considering imbalanced data. *Electric Power Systems Research*, *214*, 108886. https://doi.org/10.1016/j.epsr.2022.108886

Yan, Z., Member, S., Wen, H., & Member, S. (2022). Performance analysis of electricity theft detection for the smart grid: An overview. *IEEE Transactions on Instrumentation and Measurement*, *71*. https://doi.org/10.1109/TIM.2021.3127649

Zheng, Z., Yang, Y., Niu, X., Dai, H. N., & Zhou, Y. (2018). Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. *IEEE Transactions on Industrial Informatics*, *14*(4), 1606–1615. https://doi.org/10.1109/TII.2017.2785963

Zidi, S., Mihoub, A., Mian Qaisar, S., Krichen, M., & Abu Al-Haija, Q. (2023). Theft detection dataset for benchmarking and machine learning based classification in a smart grid environment. *Journal of King Saud University - Computer and Information Sciences*, *35*(1), 13–25. https://doi.org/10.1016/j.jksuci.2022.05.007

AI-adoption attitudes in Southern Africa's higher education sector: A pilot survey using the capability, opportunity, motivation and behaviour (COM-B) model

Mark E. Patterson
Associate Professor, School of Pharmacy, University of Missouri-Kansas City
https://orcid.org/0000-0002-2600-6887

Johan Breytenbach
Senior Lecturer, Department of Information Systems, University of the Western Cape, Cape Town
https://orcid.org/0000-0001-7883-7140

Ian Coffman
Pharm.D. Candidate, School of Pharmacy, University of Missouri-Kansas City
https://orcid.org/0009-0000-6581-9285

Abstract

Artificial intelligence (AI) drives innovation but faces numerous potential challenges to adoption. This pilot survey applied the capability, opportunity, motivation and behaviour (COM-B) model to examine AI adoption attitudes in the Southern African higher education sector. The study sought to evaluate the extent to which the COM-B framework, rooted in behavioural science, can generate AI-adoption insights that would be complementary to insights generated by established information systems (IS) adoption models, such as the technology acceptance model (TAM) and the unified theory of acceptance and use of technology (UTAUT). Potential facilitators and barriers with respect to adoption of AI tools adoption were mapped against COM-B domains to develop a 10-point Likert-type scale survey that was piloted with 33 individuals working in the Southern African higher education sector. The findings identified key facilitators of AI as adequate technological infrastructure, readiness to address clients' ethical concerns, and beliefs that AI tools benefit clients. The dominant barrier identified was clients' potential ethical concerns regarding AI use in decision-making.

Keywords

artificial intelligence (AI), adoption; higher education sector, Southern Africa, capability, opportunity, motivation and behaviour (COM-B) model

DOI: https://doi.org/10.23962/ajic.i35.21607

Recommended citation

Patterson, M.E., Breytenbach, J., & Coffman, I. (2025). AI-adoption attitudes in Southern Africa's higher education sector: A pilot survey using the capability, opportunity, motivation and behaviour (COM-B) model. *The African Journal of Information and Communication (AJIC), 35*, 1–13.
https://doi.org/10.23962/ajic.i35.21607

1. Introduction

Artificial intelligence (AI) is acknowledged as being a catalyst for socioeconomic development, propelling technological innovation across various sectors and fostering economic growth (Brynjolfsson & McAfee, 2014). In areas such as healthcare, education, and finance, AI applications offer the promise of greater efficiency and enhanced decision-making. However, these positive outcomes are often counterbalanced by challenges, including limited expertise, scarce resources, and unresolved ethical dilemmas (Binns, 2018). Understanding the core factors that drive AI adoption is critical for formulating effective implementation strategies.

Recent studies have highlighted the increasing importance of end-user attitudes and perceptions in shaping AI adoption outcomes, even within environments that possess strong IT infrastructure (Cocosila & Archer, 2017; Dwivedi et al., 2019). For instance, it has been found that sceptical end-user attitudes towards the accuracy or ethical implications of AI in healthcare often outweigh users' technical capabilities and create adoption barriers (Binns, 2018). Other factors found to be linked to AI-adoption resistance include fears of job displacement and concerns over AI's mimicking of human roles in education (Akinwalere & Ivanov, 2022). Recent advances in generative, agentic, and robotic AI are enabling increasingly adaptive human–machine interactions, thus adding complexity to the cognitive and behavioural factors influencing technology adoption (Obrenovic et al., 2024).

In behavioural science, the capability, opportunity, motivation and behaviour (COM-B) model is used to understand factors leading to behaviour change (Cane et al., 2012; Michie et al., 2011). While frequently applied in healthcare settings, including nursing and psychology (de la Fuente Tambo et al., 2024; Luo et al., 2024), COM-B has generally not been used in the context of behaviour change linked to adoption of technology. Technology adoption is typically explored through information systems (IS) models, and, in particular, through the technology acceptance model (TAM), and the unified theory of acceptance and use of technology (UTAUT).

The core aim of this pilot study was to explore the extent to which applying the COM-B model to the analysis of AI-adoption attitudes could generate findings that would complement findings from IS-focused frameworks, such as TAM and UTAUT. Our view is that expansion of the analytic lens for AI adoption to include capability, opportunity, motivation and behaviour domains has the potential to allow IS frameworks to offer more nuanced insights. The emphasis of COM-B on individual capability, environmental opportunities, and motivational factors would appear to make it well-suited for exploring socio-technical attitudes (Michie & West, 2013), and thus applicable to identifying individual end-user attitudes towards AI adoption. Also, although COM-B has been successfully used in South Africa to describe adoption of non-technological elements (see Marsh et al., 2021), the model's applicability to the country's technology-focused settings remains unexplored.

In South Africa, wide and systemic socio-economic inequalities create an environment characterised by a strong tension between AI's innovative potential and wide disparities in end-user capacity to harness this potential. While South Africa is emerging as a regional leader in AI innovation, particularly through startups focused on social impact (Dada & Van Belle, 2023; Opesemowo & Adewuyi, 2024), persistent disparities in education, income, and digital literacy have the potential to hinder broad adoption (Ganapathy et al., 2024). While these challenges are global, they are especially pronounced in South Africa, influencing not only infrastructure access but also adoption attitudes—thus pointing to the need for the application of behavioural science frameworks, such as COM-B, that are sensitive to context-specific behavioural barriers to, and facilitation of, AI adoption.

A strong argument in the existing literature is that to fully capture the complex interplay between infrastructure access and behavioural intent, especially in contexts marked by inequality, existing IS-oriented technology adoption frameworks may need to be expanded to include behavioural intent (Sohn & Kwon, 2020). One such potential expansion is through incorporation of the COM-B behaviour-change framework. Accordingly, the pilot study discussed in this article explored the extent to which the COM-B framework's constructs, rooted in behavioural science, can be effectively applied to understanding AI adoption, thus complementing insights from established IS-focused technology acceptance models.

The context for the pilot study was the Southern African higher education sector**.** Through piloting a COM-B-focused AI adoption survey in this sector as a test case, we explored the extent to which the COM-B framework could capture barriers to and facilitators of AI adoption in a socially complex setting. Rather than seeking to replace established IS models such as TAM and UTAUT, we sought to determine the extent to which COM-B may offer complementary insights by focusing on behavioural drivers.

The COM-B model was chosen for deployment in this study because it provides a framework for understanding how individual capability (e.g., psychological and physical), motivation (e.g., habits, emotions), and environmental opportunities (e.g., infrastructure and social norms) interact to influence behaviour change (Michie et al., 2011). Like TAM (Davis, 1989), COM-B accounts for both individual and environmental factors. However, unlike TAM, which emphasises cognitive beliefs, such as perceived usefulness and ease of use, COM-B prioritises emotional and automatic motivational drivers, including impulses, emotions, and habits that influence behaviour (Michie et al., 2011; West & Michie, 2021). It is our view that drivers of this nature can be particularly relevant in AI-adoption contexts, because these contexts involve controversial and/or unresolved issues such as algorithmic bias, lack of transparency, misinformation, employment and ethical concerns linked to matters of autonomy and surveillance.

2. Study design

Identification of AI-adoption barriers and facilitators

An exploratory literature review was conducted to identify known barriers to, and facilitators of, AI adoption (e.g., infrastructure, training, ethical matters) in South Africa. AI was broadly defined as any digital system or algorithm that supports or automates decision-making in a professional capacity. The healthcare-sector literature was particularly valuable, as it extensively covers behavioural issues that align with the COM-B framework. The literature search was conducted using Google Scholar with keywords such as "artificial intelligence," "machine learning," "implementation science," "barriers to AI implementation/adoption," "healthcare South Africa," and "complication/risk prediction." Boolean operators (AND, OR) were used to refine the searches. Inclusion criteria required articles to be published within the past 10 years and to have a minimum of 10 citations.

A deductive thematic analysis (Braun & Clarke, 2006), guided by the COM-B framework, was applied to the literature-review findings. Using a codebook built from COM-B domain definitions set out by Michie et al. (2011) and supplemented with constructs from the theoretical domains framework (TDF) (Cane et al., 2012), we coded relevant examples of barriers and facilitators according to four COM-B domains:

- psychological capability;
- physical opportunity;
- social opportunity; and
- reflective motivation.

The two remaining domains—automatic motivation, which captures habitual reactions, and physical capability, which captures physical behaviours—were excluded, as they are less relevant to AI adoption in professional contexts where adoption is generally intentional and cognitively mediated. An inductive analysis then grouped these examples into broader themes (see column 3 in Table 1), for example, "accuracy", "data infrastructure", "interpretation", "skills/expertise" and "workflow", which translated sector-specific insights into transferable concepts. These themes informed the first iteration of the survey instrument.

Table 1: COM-B domains (from Michie et al., 2011), barriers and facilitators, and themes

COM-B domains	AI-adoption barriers and facilitators	Themes
Psychological capability	• Lack of systems integration heightens cognitive demands, limiting individuals' capacity to learn and apply new processes (Ahmed et al., 2020; Leeds et al, 2018; Wiens & Shenoy, 2018) • Additional training intensifies the workload, requiring greater psychological stamina to continually acquire, process, and retain new information (Gesulga et al., 2017; Tan et al., 2022) • Insufficient skills and expertise erode confidence, constraining psychological capability and readiness to adopt new behaviours or technologies (Birkhoff et al., 2021; Cai et al., 2019; Guo & Li, 2018; Gesulga et al., 2017; Tan et al., 2022)	• Accuracy • Data infrastructure • Interpretation • Skills/expertise • Workflow
Physical opportunity	• Rural settings limit physical access (Guo & Li, 2018; Owoyemi et al., 2020; Peiffer-Smadja et al., 2020) • Insufficient data availability undermines tool's accuracy (Nelson, 2019; Paiva et al., 2020; Panch et al., 2019; Peiffer-Smadja et al., 2020; Ravi et al., 2016) • Financial demands, from implementation to infrastructure costs, undermine feasibility (Kruse et al., 2016; Owoyemi et al., 2020; Schawalbe et al., 2020) • Privacy, security, and ethical concerns deter uptake (Habehh & Gohel, 2021; Vayena et al., 2018) • Data integration challenges impede seamless operation (Ahmed et al., 2020; Leeds et al., 2018; Wiens & Shenoy, 2018) • Limited specialist availability slows adoption (Birkhoff et al., 2021; Guo & Li, 2018; Paranjape et al., 2019; Wahl et al., 2018) • Supportive resources help overcome resistance rooted in attitudes, culture, and workload concerns (Granja et al., 2018; Jauk et al., 2021; Lambert-Kerzner et al., 2018)	• Access • Accuracy • Costs • Data infrastructure • Ethics and regulation
Social opportunity	• Data safety and privacy concerns influence collective acceptance (Bajwa et al., 2021; Habehh & & Gohel, 2021; Vayena et al., 2018) • Clinical practice norms and limited integration foster resistance (Granja et al., 2018; Jauk et al., 2021; Lambert-Kerzner et al., 2018) • Regulatory frameworks, or lack thereof, shaping practice standards (Alexopoulos et al., 2019; Bajwa et al., 2021; Qayyum et al., 2020; O'Sullivan et al., 2019; Owoyemi et al., 2020)	• Context of patient needs • Data infrastructure • Ethics and regulation • Practice norms • Privacy & security • Skills/expertise • Work climate
Reflective motivation	• Shared decision-making and patient perspectives influence personal beliefs and willingness to change (Bilimoria et al., 2013; Davenport & Kalakota, 2019; Johnson et al., 2016) • Lack of trustworthy regulations undermines confidence, decreasing motivation to adopt new practices (O'Sullivan et al., 2019; Xiao et al., 2018) • Over-reliance on tools reduces personal agency and sustained motivation (Secinaro et al., 2021)	• *Context of patient needs* • *Ethics and regulation* • *Over-reliance* • *Policy and social infrastructure* • *Practice norms* • *Emotional resistance*

Development of survey instrument

To develop the survey instrument, we began with an initial pool of 28 open-ended items derived from the thematic analysis (described above) of literature on AI adoption, implementation barriers, and behavioural constructs aligned with the COM-B framework. The survey items were refined through an iterative, consensus-based process by our interdisciplinary team, which brought together expertise in qualitative research, information systems, and implementation science. We prioritised face and content validity, conceptual clarity, and full coverage of the key behavioural domains.

This process involved merging overlapping items, removing redundant or overly narrow items, and revising or splitting unclear items to better reflect distinct concepts. We also adapted the wording to ensure that each item aligned with its intended COM-B domain while still remaining broadly applicable across professional sectors—by, for example, replacing healthcare-specific terms such as "patient" and "clinical work" with more neutral alternatives such as "client" and "job tasks". These revisions preserved the theoretical integrity of the COM-B domains while enhancing the instrument's relevance across various professional contexts, including, but not limited to, higher education.

The AI-adoption barriers and facilitators identified in the literature were then mapped to the COM-B domains to generate 16 statements that could be surveyed via a 10-point Likert-type scale, with the scale measuring level of agreement with the statement (with 1 being the lowest level of agreement). The final 16 survey items used in the survey (see Table 2 below) explored factors influencing AI adoption in professional settings in terms of four COM-B domains drawn from Michie et al. (2011): psychological capability, physical opportunity, social opportunity, and reflective motivation.

- *Psychological capability* refers to an individual's perceived knowledge, skills, and cognitive abilities. In the context of AI, this includes understanding how to interpret AI outputs and operate AI tools effectively.
- *Physical opportunity* refers to environmental resources, time, and infrastructure. In the context of AI, this includes adequate infrastructure for data integration, data sharing, and privacy safeguards.
- *Social opportunity* refers to cultural norms, social influences, and peer support. In the context of AI, this includes clinical practice norms, workplace climate, regulatory frameworks, and best practices.
- *Reflective motivation* captures beliefs, attitudes, and intentions. In the context of AI, this includes trust in AI and ethical concerns.

Table 2: The 16 COM-B-aligned statements used in the survey

COM-B domain(s) (from Michie et al., 2011)	Survey statement	Potential facilitator or barrier
Reflective motivation	I am concerned about relying too much on AI tools for my professional decisions.	Barrier
Psychological capability	I have adequate skills to run AI tools in my industry.	Facilitator
Psychological capability & reflective motivation	When using AI tools in my industry, I understand and am confident in the results and/or output.	Facilitator
Psychological capability & social opportunity	I am prepared to address my clients' ethical issues regarding AI tools in my decision-making.	Facilitator
Physical opportunity	My workplace has adequate technological infrastructure to effectively use AI tools.	Facilitator
Physical opportunity	Our current computer systems easily integrate AI tools.	Facilitator
Physical opportunity	My workplace has adequate support systems to effectively implement AI tools.	Facilitator
Physical opportunity	Our AI tools comply with regulation and privacy laws.	Facilitator
Physical opportunity	I can integrate AI tools into my job tasks with minimal effort and time.	Facilitator
Physical opportunity	The costs of infrastructure and resources limit our ability to use AI tools.	Barrier

Reflective motivation	AI tools benefit our clients.	Facilitator
Reflective motivation	Our AI tools are accurate enough to inform our professional decisions.	Facilitator
Reflective motivation	The precision of AI tools impacts my willingness to use AI tools for professional decisions.	Barrier
Reflective motivation & social opportunity	We are concerned that clients may have ethical concerns about our use of AI tools in decision-making.	Barrier
Social opportunity	Integrating AI tools aligns with our industry's best practices and standards.	Facilitator
Social opportunity	Our workplace culture supports and rewards innovation.	Facilitator

Survey administration

In November 2024, the pilot survey was electronically distributed via Qualtrics, an online survey platform that enables secure distribution and collection of questionnaire responses (Qualtrics, 2024)[1], to 100 higher education employees affiliated with the Southern African Association for Institutional Research (SAAIR)[2], a regional professional body comprising university staff and faculty. SAAIR membership spans a wide range of institutions (e.g., research-intensive, teaching-focused, rural, and urban universities) across the Southern African region. The target respondents worked primarily in academic planning, institutional research, and policy roles, and had all previously expressed an interest in AI. The survey was administered shortly after a SAAIR conference forum focused on the impact of generative AI in higher education. As such, the potential respondents were likely to have a working knowledge of AI and a shared context for interpreting the survey questions. Of the 100 individuals contacted, 32 responded, yielding a 32% response rate. The survey had five demographic questions, covering age, job sector, job title, years of experience, and country of origin, followed by the 16 COM-B-framed items that respondents scored via a 10-point Likert-type scale. The scale ranged from 1 = strongly disagree to 10 = strongly agree.

Data analysis

The dataset generated by the survey responses on the Qualtrics platform was analysed using median scores, interquartile ranges (IQRs), and coefficient of quartile variations (CQVs). Medians highlighted central tendencies, with higher values indicating stronger agreement with the survey statement. Interquartile ranges (IQRs) were used to describe the spread of the data, indicating the range within which the middle 50% of responses fell. The coefficient of quartile variation (CQV) assessed relative variability around the median, with higher CQV values indicating greater relative dispersion within the data. Medians of 7 or above (on the 1–10 scale) were interpreted as indicating strong agreement with the survey statement. Medians between 5 and 6, near the midpoint of the scale, indicated neutral or mixed levels of agreement. Medians below 5 signified disagreement. CQV values below 0.4 indicated a strong consensus among respondents, while values above 0.6 suggested divergent views and lower consensus.

3. Results and discussion

Respondent demographics

As shown in Table 3, the respondents' years of experience in their current positions ranged between less than a year and more than 20 years, with the largest subgroup (7 respondents) having 11 to 15 years of experience. The ages of the respondents ranged between 25 and 65-plus, with the largest numbers of participants being in the 35–44 and 45–54 age bands. Respondents held diverse positions, with the role of programme director or manager being the most common (9 respondents), followed by planner or administrator (5 respondents), and teaching, learning or curriculum specialist (5 respondents), data analyst or administrator (4 respondents), researcher or consultant (4 respondents), quality assurance officer or consultant (3 respondents) and professor or lecturer (2 respondents).

[1] https://www.qualtrics.com

[2] https://www.saair-web.co.za

Table 3: Respondent demographics (N=32)

Years of experience in current position	N (%)
Less than 1 year	5 (15.6)
1 to 2 years	6 (18.8)
3 to 5 years	6 (18.8)
6 to 10	3 (9.4)
11 to 15	7 (21.9)
16 to 20	1 (3.1)
More than 20	4 (12.5)
Age	**N (%)**
25 to 34	4 (12.5)
35 to 44	9 (28.1)
45 to 54	9 (28.1)
55 to 64	7 (21.9)
65-plus	3 (9.4)
Position	**N (%)**
Programme director or manager	9 (28.1)
Planner or administrator	5 (15.6)
Teaching, learning, or curriculum specialist	5 (15.6)
Data analyst or administrator	4 (12.5)
Researcher or consultant	4 (12.5)
Quality assurance officer or consultant	3 (9.4)
Professor or lecturer	2 (6.3)

Findings from COM-B-aligned survey items

Potential facilitators of AI adoption

Median scores for potential facilitators on the 10-point Likert-type scale ranged from 5 to 7.5 (Table 4). The highest median score (7.5) indicated agreement on the presence of adequate technological infrastructure for using AI tools. Three other potential facilitators received strong median scores (7): preparedness to address clients' ethical issues regarding AI tools in the respondents' decision-making; the ability of the respondents' computer systems to integrate AI tools; and the benefit that respondents felt AI offered to their clients.

The lowest median scores (5) were observed for three factors: having adequate workplace support systems to effectively implement AI tools; ensuring AI tools comply with regulations and privacy laws; and confidence that AI tools are accurate enough to inform professional decisions. CQVs ranged from 0.17 to 0.73. The highest consensus (CQV = 0.17) pertained to the presence of adequate technological infrastructure, while the lowest consensus (CQV = 0.73) related to beliefs about workplace culture supporting innovation.

Through examining medians and CQVs simultaneously, it was found that three of the four items with high medians also reflected strong consensus, as indicated by CQVs below 0.4. These were: the presence of adequate technological infrastructure for using AI tools (median = 7.5, CQV = 0.17); preparedness to respond to clients' ethical concerns (median = 7, CQV = 0.32); and the benefit AI tools provide to clients (median = 7, CQV = 0.29). There was less consensus on the ease of respondents' computer systems integrating AI tools (median = 7, CQV = 0.43).

Table 4: Findings on potential facilitators of AI adoption

COM-B domain(s)	Survey statement *(1=strongly disagree to 10=strongly agree)*	Lowest score	Highest score	Median score	IQR	CQV	Variability
Physical opportunity	My workplace has adequate technological infrastructure to effectively use AI tools.	1	10	7.5	1.25	0.17	low
Psychological capability & social opportunity	I am prepared to address my clients' ethical issues regarding AI tools in my decision-making.	2	10	7	2.25	0.32	medium
Physical opportunity	Our current computer systems easily integrate AI tools.	2	10	7	3	0.43	medium
Reflective motivation	AI tools benefit our clients.	2	10	7	2	0.29	medium
Social opportunity	Integrating AI tools aligns with our industry's best practices and standards.	3	10	6.5	3	0.46	medium
Psychological capability	I have adequate skills to run AI tools in my industry.	1	10	6	3	0.5	high
Psychological capability & reflective motivation	When using AI tools in my industry, I understand and am confident in the results and/or output.	1	10	6	3.25	0.54	medium
Physical opportunity	I can integrate AI tools into my job tasks with minimal effort and time.	1	10	6	4	0.67	medium
Social opportunity	Our workplace culture supports and rewards innovation.	2	10	5.5	4	0.73	high
Physical opportunity	My workplace has adequate support systems to effectively implement AI tools.	1	9	5	3	0.6	high
Physical opportunity	Our AI tools comply with regulation and privacy laws.	1	10	5	3	0.6	high
Reflective motivation	Our AI tools are accurate enough to inform our professional decisions.	1	9	5	2	0.4	medium

Potential barriers to AI adoption

Median scores for potential barriers on the 10-point Likert-type scale ranged from 6 to 7 (Table 5). The higher median score (7) indicated agreement on three potential barriers: concerns about overreliance on AI for professional decisions; the uncertain precision of AI tools affecting willingness to use them; and clients' ethical concerns about using AI in decision-making. The lower median score (6) was observed for the barrier posed by costs of infrastructure and resources. CQVs ranged from 0.29 to 0.71. The highest consensus (CQV = 0.29) pertained to concern about clients' ethical concerns, while the lowest consensus (CQV = 0.71) related to concerns about overreliance on AI and about the costs of infrastructure and resources.

Through examining medians and CQVs together, we were able to identify one potential barrier with both a high median and a strong consensus (a CQV below 0.4). That potential barrier was that clients may have ethical concerns about the use of AI tools in decision-making (median = 7; CQV = 0.29). There was less consensus about two other potential barriers that had high median scores: overreliance on AI (median = 7; CQV = 0.71) and costs of infrastructure and resources (median = 6; CQV = 0.73).

Table 5: Findings on potential barriers to AI adoption

COM-B domain(s)	Survey statement (1=strongly disagree to 10=strongly agree)	Lowest score	Highest score	Median score	IQR	CQV	Variability
Reflective motivation	I am concerned about relying too much on AI tools for my professional decisions.	1	10	7	5	0.71	medium
Reflective motivation	The precision of AI tools impacts my willingness to use AI tools for professional decisions.	2	10	7	3	0.43	medium
Reflective motivation & social opportunity	We are concerned that clients may have ethical concerns about our use of AI tools in decision making.	2	10	7	2	0.29	high
Physical opportunity	The costs of infrastructure and resources limit our ability to use AI tools.	2	10	6	4.25	0.71	high

4. Conclusions and limitations

The results of this pilot study, as set out above, suggest that the COM-B model offers a potentially useful behaviourally grounded lens for evaluating AI adoption, thus potentially complementing traditional IS models such as TAM and UTAUT. Deployment of the COM-B model has the potential to contribute to the understanding of technology adoption's behavioural-intention construct—a construct that, in existing technology-adoption frameworks, is often limited to variables such as ease of use and perceived usefulness. In the evolving landscape of AI, the application of the COM-B model in adoption research has the potential to help assess readiness by identifying gaps in capability, opportunity, and motivation.

Because it was a pilot exercise, this study had several limitations. The small sample size (n = 33), and narrow focus on a particular grouping of Southern African higher education professionals, restricted generalisability. The sample was also not large enough to support formal psychometric validation. While COM-B is a widely validated framework, future research should assess the reliability and validity of this specific model in the AI adoption context via larger, more diverse samples. Also, the survey did not collect information on which AI tools respondents used or how they defined AI, which may have introduced variability in interpretation. Additionally, some education-specific or sectoral nuances may not have been fully captured. We recommend that future studies include items capturing respondents' AI usage and definitions and incorporate sector-specific validation. We also acknowledge that only a few items were used to assess each COM-B construct, which may have limited internal consistency. Also, the use of a 10-point Likert-type scale, despite offering granularity, may have introduced ceiling and/or floor effects.

AI declaration

The authors acknowledge the use of ChatGPT v4.0 to assist with editing and grammar, and Scholar GPT was used to assist with identifying literature.

Competing interests declaration

The authors have no competing interests to declare.

Data availability statement

The data supporting the findings of this study is available upon written request to the first-listed author at : pattersonmar@umkc.edu

Funding

This was unfunded work as part of a collaboration between the University of Missouri and University of the Western Cape.

Authors' contributions

M.E.P: Conceptualisation, methodology, data collection and analysis, validation, data curation, writing and revisions of the manuscript, student supervision, project management, project leadership activities. J.B.: Conceptualisation, methodology, data collection and analysis, validation, data curation, manuscript revisions, student supervision.

I.C.: Conceptualisation, methodology, validation, manuscript revisions.

References

Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database: The Journal of Biological Databases and Curation, 2020,* baaa010. https://doi.org/10.1093/database/baaa010

Akinwalere, S. N., & Ivanov, V. (2022). Artificial intelligence in higher education: Challenges and opportunities. *Border Crossing, 12*(1), 1–15. https://doi.org/10.33182/bc.v12i1.2015

Alexopoulos, C., Lachana, Z., Androutsopoulou, A., Diamantopoulou, V., Charalabidis, Y., & Loutsaris, M.A. (2019). How machine learning is changing e-government. In *ICEGOV 2019: Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance* (354–363). https://doi.org/10.1145/3326365.3326412

Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthcare Journal, 8*(2*),* e188–e194. https://doi.org/10.7861/fhj.2021-0095

Bilimoria, K. Y., Liu, Y., Paruch, J. L., Zhou, L., Kmiecik, T. E., Ko, C. Y., & Cohen, M. E. (2013). Development and evaluation of the universal ACS NSQIP surgical risk calculator: A decision aid and informed consent tool for patients and surgeons. *Journal of the American College of Surgeons, 217*(5), 833–42.e423. https://doi.org/10.1016/j.jamcollsurg.2013.07.385

Birkhoff, D. C., van Dalen, A. S. H. M., & Schijven, M. P. (2021). A review on the current applications of artificial intelligence in the operating room. *Surgical Innovation*, *28*(5), 611–619. https://doi.org/10.1177/1553350621996961

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*, *81*, 149–159. https://proceedings.mlr.press/v81/binns18a.html

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W. W. Norton & Company.

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. In *Proceedings of the ACM on Human-Computer Interaction*, *3 (CSCW)* (1–24). https://doi.org/10.1145/3359206

Cane, J., O'Connor, D., & Michie, S. (2012). Validation of the theoretical domains framework for use in behaviour change and implementation research. *Implementation Science, 7*(37), 1–17. https://doi.org/10.1186/1748-5908-7-37

Cocosila, M., & Archer, N. (2017). Practitioner pre-adoption perceptions of Electronic Medical Record systems. *Behaviour & Information Technology*, *36*(8), 827–838. https://doi.org/10.1080/0144929X.2017.1303083

Dada, O.A & Van Belle, J.P. (2023) Factors influencing the establishment of technology innovation hubs: A structured literature review. In *The 9th Annual ACIST Proceedings*. https://digitalcommons.kennesaw.edu/acist/2023/presentations/4

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Davis, F. D., Granić, A., & Marangunić, N. (2024). *The Technology Acceptance Model: 30 years of TAM.* Springer International Publishing. https://doi.org/10.1007/978-3-030-45274-2 de la Fuente Tambo, D., Moreno, S. I., & Ruiz, M. A. (2024). *Barriers and enablers for generative artificial intelligence in clinical psychology: A qualitative study based on the COM-B and theoretical domains framework (TDF) models*. Research Square. https://doi.org/10.21203/rs.3.rs-5309244/v1

Dwivedi, Y. K., Hughes, L., Ismagilova, E., & Aarts, G. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice, and policy. *International Journal of Information Management*, *57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Ganapathy, A., Heeks, R., & Iazzolino, G. (2024). *Theorizing digital inclusion and inequalities in ICT4D: Insights and implications for future research*. GDI Digital Development Working Paper No. 107. https://www.gdi.manchester.ac.uk/research/publications/di

Gesulga, J. M., Berjame, A., Moquiala, K. S., & Galido, A. (2017). Barriers to electronic health record system implementation and information systems resources: A structured review. *Procedia Computer Science*, *124*, 544–551. https://doi.org/10.1016/j.procs.2017.12.188

Granja, C., Janssen, W., & Johansen, M. A. (2018). Factors determining the success and failure of ehealth interventions: Systematic review of the literature*. Journal of Medical Internet Research*, *20*(5), e10235. https://doi.org/10.2196/10235

Guo, J., & Li, B. (2018). The application of medical artificial intelligence technology in rural areas of developing countries*. Health Equity, 2*(1), 174–181. https://doi.org/10.1089/heq.2018.0037

Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. *Current Genomics, 22*(4), 291–300. https://doi.org/10.2174/1389202922666210705124359

Huy, L. V., Nguyen, H. T., Vo-Thanh, T., Thinh, N. H. T., & Thi Thu Dung, T. (2024). Generative AI, why, how, and outcomes: A user adoption study. *AIS Transactions on Human-Computer Interaction*, *16*(1), 1–27. https://doi.org/10.17705/1thci.00198

Jauk, S., Kramer, D., Avian, A., Berghold, A., Leodolter, W., & Schulz, S. (2021). Technology acceptance of a machine learning algorithm predicting delirium in a clinical setting: A mixed-methods study*. Journal of Medical Systems*, *45*(4), 48. https://doi.org/10.1007/s10916-021-01727-6

Johnson, A. E., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D. A., & Clifford, G. D. (2016). Machine learning and decision support in critical care. In *Proceedings of the IEEE, 104*(2), 444–466. https://doi.org/10.1109/JPROC.2015.2501978

Keyworth, C., Epton, T., Goldthorpe, J., Calam, R., & Armitage, C. J. (2020). Acceptability, reliability, and validity of a brief measure of capabilities, opportunities, and motivations ("COM-B"). *British Journal of Health Psychology*, *25*(3), 474–501. https://doi.org/10.1111/bjhp.12417

Korpelainen, E., & Kira, M. (2013). Systems approach for analyzing problems in IT system adoption at work. *Behaviour & Information Technology*, *32*(3), 247–262. https://doi.org/10.1080/0144929X.2011.624638

Kruse, C. S., Kristof, C., Jones, B., Mitchell, E., & Martinez, A. (2016). Barriers to electronic health record adoption: A systematic literature review. *Journal of Medical Systems, 40*(12), 252. https://doi.org/10.1007/s10916-016-0628-9

Lambert-Kerzner, A., Ford, K. L., Hammermeister, K. E., Henderson, W. G., Bronsert, M. R., & Meguid, R. A. (2018). Assessment of attitudes towards future implementation of the "Surgical Risk Preoperative Assessment System" (SURPAS) tool: A pilot survey among patients, surgeons, and hospital administrators. *Patient Safety in Surgery*, *12*, 12. https://doi.org/10.1186/s13037-018-0159-z

Leeds, I. L., Rosenblum, A. J., Wise, P. E., Watkins, A. C., Goldblatt, M. I., Haut, E. R., Efron, J. E., & Johnston, F. M. (2018). Eye of the beholder: Risk calculators and barriers to adoption in surgical trainees. *Surgery*, *164*(5), 1117–1123. https://doi.org/10.1016/j.surg.2018.07.002

Luo, C., Yang, C., Yuan, R., Liu, Q., Li, P., & He, Y. (2024). Barriers and facilitators to technology acceptance of socially assistive robots in older adults: A qualitative study based on the capability, opportunity, and motivation behavior model (COM-B) and stakeholder perspectives. *Geriatric Nursing, 58*, 162–170. https://doi.org/10.1016/j.gerinurse.2024.05.025

Marsh, R. J., Brent, A. C., & De Kock, I. H. (2021). Understanding the barriers and drivers of sustainable construction adoption and implementation in South Africa: A quantitative study using the Theoretical Domains Framework and COM-B model. *Journal of the South African Institution of Civil Engineering*, *63(4)*, 11–23. https://doi.org/10.17159/2309-8775/2021/v63n4a2

Michie, S., van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science, 6*, 42. https://doi.org/10.1186/1748-5908-6-42

Michie, S., & West, R. (2013). Behaviour change theory and evidence: A presentation to Government. *Health Psychology Review*, *7*(1), 1–22. https://doi.org/10.1080/17437199.2011.649445

Nelson, G. S. (2019). Bias in artificial intelligence. *North Carolina Medical Journal*, *80*(4), 220–222. https://doi.org/10.18043/ncm.80.4.220

Obrenovic, B., Gu, X., Wang, G., Godinic, D., & Jakhongirov, I. (2024). Generative AI and human–robot interaction: Implications and future agenda for business, society, and ethics. *AI & Society*, *40*, 677–690. https://doi.org/10.1007/s00146-024-01889-0

Opesemowo, O. A. G., & Adewuyi, H. O. (2024). A systematic review of artificial intelligence in mathematics education: The emergence of 4IR. *Eurasia Journal of Mathematics, Science & Technology Education*, *20*(7), em2478. https://doi.org/10.29333/ejmste/14762

O'Sullivan, S., Nevejans, N., Allen, C., Blyth, A., Leonard, S., Pagallo, U., Holzinger, K., Holzinger, A., Sajid, M. I., & Ashrafian, H. (2019). Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery, 15*(1), e1968. https://doi.org/10.1002/rcs.1968

Owoyemi, A., Owoyemi, J., Osiyemi, A., & Boyd, A. (2020). Artificial intelligence for healthcare in Africa. *Frontiers in Digital Health*, *2,* 6. https://doi.org/10.3389/fdgth.2020.00006

Paiva, J. O. V., Andrade, R. M. C., de Oliveira, P. A. M., Duarte, P., Santos, I. S., Evangelista, A. L. P., Theophilo, R. L., de Andrade, L. O. M., & Barreto, I. C. H. C. (2020). Mobile applications for elderly healthcare: A systematic mapping. *PloS one*, *15*(7), e0236091. https://doi.org/10.1371/journal.pone.0236091

Panch, T., Mattie, H., & Celi, L. A. (2019). The "inconvenient truth" about AI in healthcare. *npj Digital Medicine, 2*, 77. https://doi.org/10.1038/s41746-019-0155-4

Paranjape, K., Schinkel, M., Nannan Panday, R., Car, J., & Nanayakkara, P. (2019). Introducing artificial intelligence training in medical education. *JMIR Medical Education*, *5*(2), e16048. https://doi.org/10.2196/16048

Peiffer-Smadja, N., Rawson, T. M., Ahmad, R., Buchard, A., Georgiou, P., Lescure, F. X., Birgand, G., & Holmes, A. H. (2020). Machine learning for clinical decision support in infectious diseases: A narrative review of current applications. *Clinical Microbiology and Infection, 26*(5), 584–595. https://doi.org/10.1016/j.cmi.2019.09.009

Polyportis, A., & Pahos, N. (2024). Understanding students' adoption of the ChatGPT chatbot in higher education: The role of anthropomorphism, trust, design novelty and institutional policy. *Behaviour & Information Technology*, *44*(2), 315–336. https://doi.org/10.1080/0144929X.2024.2317364

Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2021). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, *14*, 156–180. https://doi.org/10.1109/RBME.2020.3013489

Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2017). Deep learning for health informatics. *IEEE Journal of Biomedical and Health informatics*, *21*(1), 4–21. https://doi.org/10.1109/JBHI.2016.2636665

Schwalbe, N., & Wahl, B. (2020). Artificial intelligence and the future of global health. *Lancet, 395*(10236), 1579–1586. https://doi.org/10.1016/S0140-6736(20)30226-9

Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., & Biancone, P. (2021). The role of artificial intelligence in healthcare: A structured literature review. *BMC Medical Informatics and Decision Making, 21*(1), 125. https://doi.org/10.1186/s12911-021-01488-9

Siau, Keng and Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal, 31*(2), 47-53. https://ink.library.smu.edu.sg/sis_research/9371

Sohn, K., & Kwon, O. (2020). Technology acceptance theories and factors influencing artificial intelligence-based intelligent products. *Telematics and Informatics*, *47,* 101324. https://doi.org/10.1016/j.tele.2019.101324

Tan, T. F., Li, Y., Lim, J. S., Gunasekeran, D. V., Teo, Z. L., Ng, W. Y., & Ting, D. S. (2022). Metaverse and virtual health care in ophthalmology: Opportunities and challenges. *Asia-Pacific Journal of Ophthalmology*, *11*(3), 237–246. https://doi.org/10.1097/APO.0000000000000537

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS medicine*, *15*(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

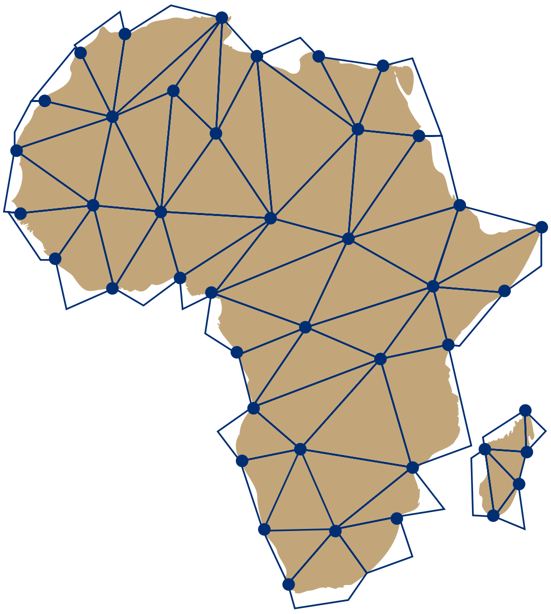
Wahl, B., Cossy-Gantner, A., Germann, S., & Schwalbe, N. R. (2018). Artificial intelligence (AI) and global health: How can AI contribute to health in resource-poor settings? *BMJ Global Health*, *3*(4), e000798. https://doi.org/10.1136/bmjgh-2018-000798

West, R., & Michie, S. (2021). A brief introduction to the COM-B model of behaviour and the PRIME theory of motivation. *Qeios*, *3*, 1-5. https://www.qeios.com/read/WW04E6.3

Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases, 66*(1), 149–153. https://doi.org/10.1093/cid/cix731

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *Journal of the American Medical Informatics Association, 25*(10), 1419–1428. https://doi.org/10.1093/jamia/ocy068

THEMATIC SECTION:
BRICS COUNTRIES AND AI SOVEREIGNTY



How are BRICS countries building AI sovereignty? Introduction to Thematic Section

Luca Belli
Professor of Law and Director, Center for Technology and Society (CTS) and CyberBRICS project, Fundação Getulio Vargas (FGV) Law School, Rio de Janeiro
https://orcid.org/0000-0002-9997-2998

Abstract

In this article, the author provides an introduction to the AJIC Thematic Section: BRICS Countries and AI Sovereignty.

Keywords

artificial intelligence (AI), sovereign AI, AI sovereignty, digital sovereignty, BRICS

DOI: https://doi.org/10.23962/ajic.i35.23192

Recommended citation

Belli, L. (2025). How are BRICS countries building AI sovereignty? Introduction to Thematic Section. *The African Journal of Information and Communication (AJIC)*, *35*, 1–4. https://doi.org/10.23962/ajic.i35.23192

1. Introduction

What is sovereign artificial intelligence (AI), and how are BRICS countries shaping their AI sovereignty narratives and experiments? This Thematic Section of *The African Journal of Information and Communication (AJIC)*, in combination with the work presented in *AJIC* Issue 34 of 2024,[1] provides an introduction to the early findings of the CyberBRICS project's ongoing research on AI sovereignty in the BRICS countries. The research is anchored in the premise that AI sovereignty constitutes a critical facet of digital sovereignty (Belli, 2025; Jiang & Belli, 2024), reflecting the imperative of states to exercise self-determination, regulatory authority, and strategic autonomy over the development, deployment, and governance of AI systems. Conspicuously, the research provides much-needed context for understanding the relevance of the BRICS members' declarations, at the 2025 BRICS Brazil Summit, with respect to AI matters (BRICS, 2025a, 2025b).

Drawing upon the corpus of research generated under the auspices of the CyberBRICS project,[2] this Thematic Section provides three contributions. The first two explore the modalities through which two of the BRICS emerging economies, India (in the article by Vipra) and Russia (in the article bv Ignatov and Kerimi), are navigating the complex terrain of AI sovereignty. The third item in the section (by Sengupta, Barbosa and Samdub) provides a comparative perspective on how two BRICS countries, India and Brazil, are leveraging digital public infrastructure (DPI) as a facilitator of AI governance and AI sovereignty.

As highlighted in *AJIC* 34's Thematic Section introduction (Belli, 2024), which should be read in conjunction with this one, we at the CyberBRICS project posit that the assertion of AI sovereignty plays an instrumental role in the preservation of national agency and technological autonomy, and in the mitigation of structural dependencies on exogenous technological actors. This analysis foregrounds the strategic significance of AI sovereignty as a precondition for the effective comprehension, regulation, and endogenous development of AI technologies. Our analytical framework also highlights the fact that a systemic approach and a critical

[1] https://ajic.wits.ac.za/issue/view/1251
[2] All CyberBRICS publications are available on an open access basis at https://cyberbrics.info/cyberbrics-publications

perspective are always needed to ascertain the extent to which AI sovereignty initiatives are concretely successful in achieving their stated purposes, and to scrutinise what could be such initiatives' collateral effects.

As we have previously argued, AI sovereignty can be situated within the broader genus of digital sovereignty, thus allowing us to elucidate its conceptual contours and juridical underpinnings. The CyberBRICS inquiry examines the legislative and regulatory instruments, as well as the industrial policy tools, that are being leveraged by the BRICS states to assert power, agency and control over their digital infrastructures and to attenuate reliance on foreign technological ecosystems. From this perspective, we contend that governance, regulation, and industrial policy must be construed as interdependent and mutually reinforcing mechanisms essential to the realisation of AI sovereignty.

Critical appraisal of selected case studies allows our research to illustrate how BRICS countries exploit alternative technological and regulatory strategies. These include, as discussed in this Thematic Section's article by Sengupta, Barbosa and Samdub, establishment of DPI as a techno-regulatory substratum conducive to digital innovation. Other strategies used are facilitative regulatory measures such as tax incentives, the designation of special industrial zones, and targeted capacity-building initiatives. Examination of such strategies illustrates the variegated successes and limitations encountered by BRICS states in their pursuit of digital sovereignty, as well as the risk of AI sovereignty initiatives being co-opted, as in the Russian case discussed in the article by Ignatov and Kerimi, to implement securitisation and control agendas.

By systematically analysing iterative regulatory practices and the trial of alternative governance models across the BRICS jurisdictions, the CyberBRICS project explores the extent to which adaptive and context-sensitive regulation can enhance the efficacy of AI governance and fortify the juridical foundations of AI sovereignty. Our outputs, including the three articles that follow in this Thematic Section, provide some of the concrete context that is necessary to situate the most recent AI-related declarations issued by the BRICS leaders.

2. Foregrounding of AI at the 2025 BRICS Summit

The increasing significance of AI governance for the BRICS nations was emphatically underscored by the outcomes of the 17th BRICS Summit in Rio de Janeiro in early July 2025. The Summit's Declaration, Strengthening Global South Cooperation for a More Inclusive and Sustainable Governance, included the following statement:

> 16. We recognize that Artificial Intelligence (AI) represents a milestone opportunity to boost development towards a more prosperous future. To achieve that goal, we underscore that global governance of AI should mitigate potential risks and address the needs of all countries, including those of the Global South. A collective global effort is needed to establish an AI governance that upholds our shared values, addresses risks, builds trust, and ensures broad and inclusive international collaboration and access, in accordance with sovereign laws, including capacity building for developing countries, with the United Nations at its core. To support a constructive debate towards a more balanced approach, we agreed on the BRICS Leaders' Statement on the Global Governance of Artificial Intelligence, which aims to foster responsible development, deployment, and use of AI technologies for sustainable development and inclusive growth, in compliance with national regulatory frameworks, the UN Charter and respecting the sovereignty of States. (BRICS, 2025a)

As alluded to in the Declaration, at the Summit the BRICS leadership collectively adopted a formal statement on AI matters, entitled the BRICS Leaders' Statement on the Global Governance of Artificial Intelligence (BRICS, 2025b), thereby marking a critical juncture in the bloc's engagement with international technology policy. In addition, the Leaders' Statement articulates a comprehensive vision that situates AI not merely as technological innovation but also as a transformative opportunity capable of advancing equitable development on a global scale, contingent upon the establishment of governance frameworks that are inclusive, representative, and attentive to the particular needs of developing countries. The document thus lays the groundwork for a robust BRICS approach to AI governance.

The Leaders' Statement reflects a nuanced understanding of the current international AI landscape, which is characterised by fragmented—or, in some respects, absent—governance mechanisms. The Statement posits that multilateralism constitutes an indispensable approach to remedy this governance deficit and to preclude a deleterious "race to the bottom" among states and corporate actors. In this context, the United Nations is identified as the central institution capable of orchestrating a coordinated response to the challenges posed by AI, thereby ensuring that regulatory frameworks are harmonised and that shared (technical) standards are upheld globally, while also promoting cooperation on research and development (R&D) and on innovative AI-governance mechanisms. In addressing the modalities of AI governance, the Statement emphasises the critical role of open-source collaboration and the development of inclusive, interoperable international standards. Such mechanisms are envisaged as essential enablers of innovation, particularly for countries with limited technological and financial resources. The document further highlights the necessity of confronting market distortions, monopolistic practices, and technological exclusion, which presently impede equitable access to AI technologies.

The Statement identifies environmental sustainability, decent work, and the enhancement of digital infrastructure as foundational pillars for the responsible deployment of AI—thus establishing a clear link between the BRICS AI-related initiatives and the UN Sustainable Development Goals (SDGs), and underscoring the potential of AI to contribute meaningfully to advancements in access to health, agriculture, energy, and education. This approach reflects a development-centred perspective that seeks to harness digital transformation as a means of reducing existing inequalities and addressing structural asymmetries within and among nations. Conspicuously, the principle of fair and equitable access to AI technologies and computing infrastructure is emphasised as a prerequisite for enabling widespread adoption and development.

Importantly, the proposed BRICS vision also delineates the need for a balanced approach to data governance,[3] which must safeguard the public interest while respecting intellectual property rights and copyright protections. This balance is deemed necessary to prevent exploitative data extraction and violations of privacy, which could undermine trust and the ethical use of AI systems. Moreover, the BRICS leaders express concern regarding algorithmic bias, particularly its impact on marginalised groups, and caution against the proliferation of misinformation and the misuse of generative AI technologies. They advocate for the development of enhanced detection tools and the promotion of media literacy as essential measures to mitigate these risks.

Underpinning the statement are several guiding principles that reflect the collective approach of the BRICS members. These include the pursuit of a shared approach and common vision for AI governance grounded in consensus-based decision-making, as well as full respect for the digital sovereignty of each member state, which entails the right to regulate AI in accordance with national policies and priorities. The document further commits to openness, transparency, accountability, and the equitable sharing of information and resources, which are deemed as instrumental conditions to foster a more trustworthy AI ecosystem. Finally, the statement affirms a commitment to mutually beneficial cooperation within BRICS and extends this spirit of collaboration to the broader Global South, advocating for win-win partnerships that transcend regional boundaries.

The adoption of this Leaders Statement was the culmination of extensive negotiations and is a unique consensus document shaped by the leading economies of the Global Majority. The Statement's development-oriented and sovereignty-respecting governance framework challenges existing paradigms dominated by developed countries and multinational corporations, thereby asserting the interests of emerging economies and developing nations in shaping the future trajectory of AI governance.

[3] This topic is explored in Belli and Gaspar (2025). For an introduction to data governance in the BRICS, see Belli and Doneda (2023).

3. Conclusion

By emphasising national regulatory frameworks grounded in the UN Charter and respecting sovereignty, the BRICS Rio Declaration (BRICS, 2025a) quoted above articulates a governance model that balances global cooperation with the autonomy of individual states. However, these statements always need to be analysed critically, with an eye to understanding the extent to which the words reflect reality. The CyberBRICS project's research and outputs—including the three outputs provided in the Thematic Section that follows—aim to provide the necessary critical lens.

With respect to the AI sovereignty ambitions of BRICS nations, our critical framing emphasises that such sovereignty is not merely about control over technology but also about creating the capacity to understand, develop and regulate AI systems. To achieve these purposes, the BRICS leaders consider it essential to establish data governance, open scientific collaboration, and capacity-building tailored to the specific needs each member country. This type of sovereignty perspective reinforces the BRICS commitment to a multilateral yet decentralised global AI governance structure, where each nation has the possibility to shape AI policies aligned with its social, economic, and cultural contexts while contributing to a shared effort towards a global vision.

References

Belli, L. (2024). BRICS countries and AI sovereignty: Introduction to Thematic Section. *The African Journal of Information and Communication (AJIC)*, *34,* 1–6. https://doi.org/10.23962/ajic.i34.20864

Belli, L. (2025). Exploring the Key AI sovereignty enablers (KASE) of Brazil, to build an AI sovereignty stack. In L. Belli & W. B. Gaspar (Eds.), *The quest for AI sovereignty, transparency and accountability: Official outcome of the UN IGF Data and Artificial Intelligence Governance Coalition*. Springer Nature. https://doi.org/10.2139/ssrn.5204010

Belli, L., & Doneda, D. (2023). Data protection in the BRICS countries: Legal interoperability through innovative practices and convergence. *International Data Privacy Law, 13*(1), 1–24. https://doi.org/10.1093/idpl/ipac019

Belli, L., & Gaspar, W. B. (Eds.). (2025). *Personal data architectures in the BRICS countries*. Oxford University Press.

BRICS. (2025a). Rio de Janeiro Declaration: Strengthening Global South Cooperation for a More Inclusive and Sustainable Governance. 6 July. https://brics.br/en/documents/presidency-documents/250705-brics-leaders-declaration-en.pdf

BRICS. (2025b). BRICS Leaders' Statement on the Global Governance of Artificial Intelligence. 6 July. https://cyberbrics.info/brics-leaders-statement-onthe-global-governance-of-artificial-intelligence/

Jiang, M., & Belli, L. (Eds). (2024). *Digital sovereignty from the BRICS countries: How the Global South and emerging power alliances are reshaping digital governance*. Cambridge University Press. https://doi.org/10.1017/9781009531085

Towards AI sovereignty: The good, the bad, and the ugly of AI policy in India

Jai Vipra

Non-Resident Fellow, CyberBRICS project, Center for Technology and Society (CTS), Fundação Getulio Vargas (FGV) Law School, Rio de Janeiro; and PhD Student, Department of Science and Technology Studies, Cornell University, Ithaca, New York

https://orcid.org/0009-0007-6220-0154

Abstract

India's approach to artificial intelligence (AI) policy reflects a mix of ambition, creativity, and inconsistency. While the country has made significant strides in areas such as computational capacity, data protection fundamentals, and connectivity, its AI sovereignty efforts are hampered by a lack of strategic coherence, inadequate cybersecurity, and an absence of algorithmic-accountability legislation. This article evaluates India's AI policy through the lens of the key AI sovereignty enablers (KASE) framework (Belli, 2023; CyberBRICS, 2024), highlighting positive precursors, opportunities for improvement, and fundamental shortcomings of the Indian approach. It argues that India's reactive and fragmented policymaking, coupled with frequent shifts in direction, undermines its potential to achieve AI sovereignty. The article concludes with recommendations for a more cohesive and forward-looking AI strategy that aligns with India's long-term interests.

Keywords

artificial intelligence (AI), AI sovereignty, India, policy, computational capacity, data protection, connectivity, cybersecurity, algorithmic accountability

DOI: https://doi.org/10.23962/ajic.i35.21263

Recommended citation

Vipra, J. (2025). Towards AI sovereignty: The good, the bad, and the ugly of AI policy in India. *The African Journal of Information and Communication (AJIC)*, 35, 1–11. https://doi.org/10.23962/ajic.i35.21263

1. Introduction

In April 2023, the Indian Ministry of Electronics and Information Technology (MeitY) stated in response to a parliamentary question that it was not planning to regulate artificial intelligence (AI), and that it sought to promote the growth of the sector in India (Singh, 2023a). In July 2023, the Telecom Regulatory Authority of India (TRAI) recommended urgently setting up a regulatory framework for AI (Aulakh, 2023). The next year, in March 2024, MeitY issued an advisory that included a provision requiring technology firms that were deemed "significant" to obtain government permission before releasing AI models (MeitY, 2024). The advisory was not meant to be legally binding. Fifteen days later, responding to criticism, the government withdrew this part of the advisory and thus no longer required firms to obtain permission before releasing AI models (Agrawal, 2024).

These instances are emblematic. India's approach to AI regulation and AI sovereignty has been characterised by infectious enthusiasm, creditable ambition, and even some good policies, but a lack of strategic coherence or overarching vision. Much AI-related policymaking in India has been in reaction to matters in which the government has had a political interest, such as in the representation of certain politicians by large language models (LLMs), resulting in frequent changes in direction.

This article aims to provide a broad overview of India's policies on, and with relevance to, AI, and to assess how these policies relate to AI sovereignty. I adopt a definition of AI sovereignty based on the key AI sovereignty enablers (KASE) framework set out in Belli (2023), in terms of which AI sovereignty is the ability of a country to "understand, develop, and regulate AI systems" in order to exercise "control, agency, and self-determination" over them. The KASE framework treats AI sovereignty as a stack, with the overall effect of reducing the unilateral impact of foreign actors on the country's choices (Belli, 2023; CyberBRICS, 2024). To identify and evaluate India's AI sovereignty stack, I map India's policy—both in its status quo and in its recent developments—along the KASE elements, which are:

1. Sound personal data governance;
2. Sound algorithmic governance;
3. Strong computational capacity;
4. Meaningful connectivity;
5. Reliable electrical power;
6. A digitally literate population;
7. Solid cybersecurity;
8. An appropriate regulatory framework, and
9. AI-ready digital public infrastructure.

The first eight KASE elements listed are from Belli (2023), and the ninth—AI-ready digital public infrastructure—was added by CyberBRICS (2024).

I evaluate India's AI policy measures in terms of whether they are positive precursors, opportunities for improvement, or fundamental shortcomings towards the goal of AI sovereignty, based on the KASE elements listed above. There is no clean separation between the categories of precursors, opportunities, and issues; they are instead on a continuum of beneficial or not for AI sovereignty. India's performance on KASE elements is evaluated both in terms of the historical evolution of these elements, as well as the paths that they open up for AI sovereignty in the future. I demonstrate India's mixed, sometimes inconsistent, and evolving approach to AI sovereignty, which can be linked to the government's domestic-capital-directed approach to sovereignty in general (Varadarajan, 2025). I then conclude that this mixed approach has led to a situation where India's dependency on foreign technological developments is only sometimes negated by domestic developments, and that the latter often fall short due to a lack of overarching strategy.

2. An overview of AI policy in India

Policy analysts have referred to India's approach to AI policy as being both oscillatory and broadly pro-innovation (Mohanty & Sahu, 2024). Mohanty and Sahu (2024) also show how a multiplicity of ministries and agencies regulates AI applications and articulates AI policy stances in India.

A few key policy initiatives and documents frame AI policy discussions in India. One of these is a 2018 report by NITI Aayog (a government think-tank), called National Strategy for Artificial Intelligence (NITI Aayog, 2018). This report identifies resource and policy constraints to equitable access to AI, and recommends investments in research, skilling, security, privacy, and promoting AI adoption, as well as closer collaboration with the private sector. In 2021, NITI Aayog followed up with a two-part report on responsible AI, outlining principles for responsible AI in India and setting out methods to operationalise these principles (NITI Aayog, 2021a; 2021b). The first part of the report assesses the direct and indirect impacts of AI, which lead to systems and societal considerations. The second part recommends actions to be taken by the government (providing an appropriate regulatory environment, etc.) and the private sector (incentivising ethics, etc.).

The flagship policy and spending vehicle of the Indian government for AI is the IndiaAI mission. Launched in March 2024, this mission aims to catalyse AI innovation by encouraging investment in, and to some extent subsidising, the various inputs to AI: computational power, data, skills, finance, and ethics frameworks (IndiaAI, n.d.; Prime Minister of India, 2024). In 2025, MeitY published a report on AI governance guidelines development (MeitY, 2025a). The report urges that AI actors be seen as constituting an ecosystem, and recommends the effective enforcement of existing laws, traceability and transparency for regulators, and a central policy coordination mechanism for AI. This last recommendation in particular might address the oscillatory nature of India's AI policymaking.

3. Promising precursors

India has made uneven progress along the KASE framework. The progress on some of these enablers has been remarkable and worth highlighting. Indian AI policy, and its digital policy more broadly, has often been characterised by creativity and unorthodoxy. This section highlights the elements of the KASE framework in which Indian AI policy has particularly excelled in its attempts to make policies with fresh perspectives. These elements serve as examples of the potential of Indian AI policy to serve sovereignty objectives—a potential that, as we shall see, remains largely unrealised.

Proactive policy on computational power

Adequate computational capacity is crucial for training and running AI models, and for conducting AI experiments. India has a comprehensive and flexible computational-power policy set that has gone through productive iterations. Computational policy is a domain in which changes in orientation have been appropriately responsive to changing conditions, rather than being reversals of hastily made policy. India's efforts to build effective computational capacity for AI have three main prongs, which I now discuss in three sub-sections.

GPU procurement subsidy

Many of today's AI models need to be trained on, and also often run on, graphics processing units (GPUs). As part of the IndiaAI mission, the government has allocated around INR50 billion (approx. USD584 million) to partially subsidise the procurement of GPUs for Indian companies (Mishra, 2024). This planned subsidy is meant to be demand-driven, in that the companies will decide the kinds of GPUs that they need to procure, and the government will ensure that the purchases that it subsidises are not misused or resold. There is debate in India around the exact model of government spending that can reliably catalyse AI innovation (Suraksha & Lohchab, 2024). Such a debate has led to a change in stated government policy from the creation of a GPU cluster through government procurement to the current policy of subsidisation (Prime Minister of India, 2024). At the time of writing, in early 2025, the government is about to launch a portal to help businesses, non-profits, government organisations, and others to access subsidised use of 18,000 GPUs (*BW Businessworld*, 2025).

Production- and design-linked incentive schemes (PLIS and DLIS)

In the period 2021–22, the government unveiled production- and design-linked incentive schemes (PLIS and DLIS, respectively) to subsidise the production and design of semiconductor chips in India (MeITY, 2023b). Since India does not possess the technology or know-how to design or produce advanced GPUs, these schemes target the design and production of chips that lag behind the frontier but are nevertheless useful. Under these schemes, the government reimburses up to 50% of costs (and incentivises a portion of sales) for companies that are majority Indian-owned in order to nurture a domestic ecosystem for compute power (MeitY, n.d.). These programmes have not had the expected level of success, particularly due to the reluctance of foreign semiconductor producers and designers to transfer technology, and due to the nascent level of the semiconductor market in India (Vipra, 2024). Nevertheless, these schemes demonstrate that Indian policymakers understand the importance of domestic capabilities in chip design and production.

Modernising the state-owned Semi-Conductor Laboratory (SCL)

In 2023, the government invited proposals to modernise the state-owned SCL (MeitY, 2025b). The state expects such modernisation to follow one of, or a combination of, two broad paths: turning SCL into a research and development hub, and/or turning it into an at-scale manufacturer of chips. At present, SCL produces older semiconductor models at a relatively low volume, but these chips are critical for India's defence needs. In an environment where chip design and production are concentrated in a few private companies outside India, this focus on developing the capabilities of a state laboratory is an important pillar of India's efforts towards AI sovereignty.

Useful data-governance experiments

While numerous countries now have data-protection laws that look broadly similar, the laws' philosophical and practical foundations differ. In India, the Digital Personal Data Protection Act of 2023 recognises data protection as a fiduciary duty (Republic of India, 2023). This means that the entity that collects or uses personal data must observe a duty of care towards the "data principal", i.e., the person to which the data pertains, and must act in the person's best interest. This requirement to act in the data principal's best interest has the scope to be interpreted in ways favourable to people, rather than corporations, even if the technology, manner, or purpose of data use changes. Specifically, in the case of AI, such a foundation opens up avenues to challenge the use of personal data to train AI models that eliminate jobs, undermine individual intellectual property, and/or degrade public or private services. In this respect, a *fiduciary* relationship is a better approach than an approach based on data as *property*, of which people can be dispossessed through unequal power relationships.

However, as is explained later in this article, other aspects of India's data protection law undermine Indians' right to privacy in the context of widespread AI use. In addition, as Bailey and Goyal (2020) point out, the use of the term "data fiduciary" has not translated into fiduciary-like obligations in the law for technology companies. Prasad (2019) presents some options to invoke fiduciary obligations, for instance, on large technology companies, but also concludes that provisions in the data protection law do not correspond to fiduciary-like obligations.

India has also experimented with other data governance practices that have the potential to protect AI sovereignty. For instance, Indian regulators like the Reserve Bank of India have had strong stances on data localisation, requiring that some types of data be stored locally in India (Reserve Bank of India, 2018). Local data storage requirements in select cases might ensure that data protection rules are more effectively applied to the data. Such requirements might also create more opportunities to build AI sovereignty by providing leverage for international negotiations on digital issues, much like national control over other resources such as minerals. Multinational technology firms prefer no local storage requirements, and countries like India use data localisation as a non-ideal method of regaining national rights over data (Basu, 2025).

Another example is India's consent management framework. Section 2(g) of India's Digital Personal Data Protection Act, 2023, provides for an intermediary called a consent manager, which enables a person to "give, manage, review and withdraw her consent through an accessible, transparent and interoperable platform" (Republic of India, 2023). The Act holds consent managers accountable to data principals (section 6(8)). Consent managers are expected to help data principals to avoid consent fatigue and provide interoperability (Kazia et al., 2025). Like consent managers, account aggregators in the financial sector aim to provide seamless flows of financial services data (Kazia et al., 2025). Whether data is actually a resource or not, the consent management framework recognises that data is used like a resource and attempts to shape a more equitable market for this resource-style use by allowing data to be transferred according to the wishes of the data principal. This is in contrast to the dominant model where, once data is collected, people have little to no control over its movement.

Respectable connectivity foundations

According to government statistics, India has 954 million internet subscribers, which is about 68% of India's population (Ministry of Communications, 2024). The average data cost fell by almost 100% in the period 2014–24, and 4G coverage of remote villages in difficult terrains neared 25% in the same period (Ministry of Communications, 2024). Through a mobile-first strategy, India has managed to connect millions of people to the internet in a short period of time.

Indian policymakers and regulators also seem to be prepared for new issues arising from the connectivity infrastructure requirements of digital technology. The Telecom Regulatory Authority of India (TRAI) has recently studied India's domestic ownership and capabilities in submarine cables and landing stations, noting that regulatory complexity has made it difficult to conduct business in this sector (Arya, 2021; TRAI, 2023). India has also opened up the space sector to private players, encouraging innovation in satellite and other related communications infrastructure (ISRO, 2023).

Caveats regarding good policy

One aspect to note in relation to all the policies I have classified here as "promising precursors" is that they are not necessarily tied to other policy measures such that they might work together towards the goal of AI sovereignty. The provision of computational capacity would return more dividends if it were paired with appropriate investments in AI education, such that the computational capacity is used optimally and to its potential. Such use is not guaranteed through the mere provision of resources. In a world where computational power is expensive and relatively scarce, investments in using it optimally require multi-pronged approaches. Similarly, data policy must go beyond the "unlocking", "leveraging", or "making available" data for AI that is so often foregrounded in policy documents and business documents aimed at policy changes (Gupta, 2025; Saxena, 2021; Singh et al., 2025; Suri, 2025).

Strong data protection laws—which, for instance, ban surveillance pricing where companies price goods individually based on data collected—can make such practices prohibitive. The discouragement of this tendency of targeted advertising and pricing, engaged in by platform business models, could perhaps redirect AI innovation towards more productive and public-interest endeavours. At the very least, it can spur experimentation with newer business models. Opening up the space sector to private players and presenting India as a space-investment destination (Modi, 2025) is not necessarily wrong, but must be accompanied by careful national security and sovereignty considerations. Not protecting against ceding control of communications infrastructure to foreign players is ill-advised in a digital economy.

4. Opportunities for improvement

Connectivity gaps

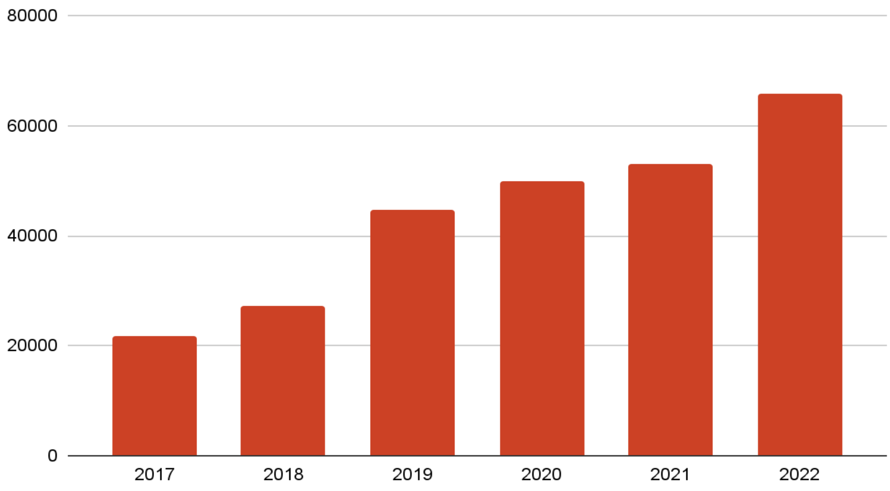
Much of India's increase in connectivity over the last decade has been the result of a price war among providers that is leading to the monopolisation and subsequent ill-health of the telecommunications market in India (Chandrasekhar & Ghosh, 2021). In other words, connectivity has been provided to a large proportion of the population by increasing monopolisation in the telecommunications sector, which might also threaten market competition in other sectors through the control of data in the telecommunications sector (Chandrasekhar & Ghosh, 2021). Thus, while connectivity gains have been achieved through this price war, the manner of their achievement might yet impose costs on consumers in the future.

According to a government survey conducted between 2018 and 2020, only a third of Indian women had ever used the internet, compared to more than half of Indian men (McDougal et al., 2022). Rural women are held back from the internet even more, with only 3 in 10 women having ever accessed the internet (Sheriff, 2020). Regional disparities in access are significant. In the national capital, there are 186 internet subscribers per 100 people (many individuals have more than one subscription), while in the state of Jharkhand, there are only 10 internet subscribers per 100 people (Parsheera, 2022). In all states, except Kerala, urban areas have a higher internet subscription density than rural areas (Parsheera, 2022). As long as access to the internet and therefore to AI technology is unequal, the government cannot achieve unbiased, fair outcomes of AI use.

Inadequate cybersecurity

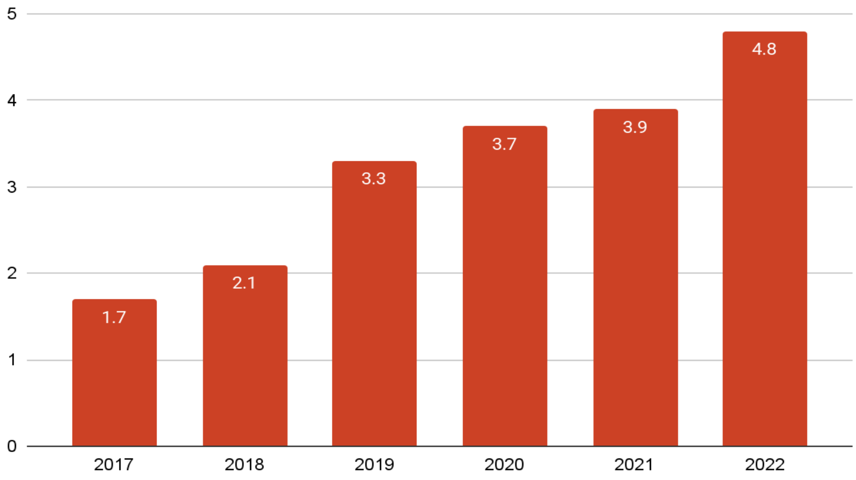
An important dimension of AI sovereignty is a country's ability to ensure the security of its AI systems and of critical infrastructure where AI is embedded. A country must be able to govern crime that occurs through digital means; with the proliferation of AI, such crimes become easier to commit and scale, through methods such as voice-cloning (Hernandez, 2023). India faces a large number of cyberattacks, and its cybersecurity strategy does not seem to be keeping pace. It ranks 47th in terms of preparedness in a composite cybersecurity index (SEON, 2023). According to one index, India was the 10th largest hotspot of cybercrime in the world (Bruce et al., 2024). Figures 1 and 2 below show that cybercrime has been steadily rising in India in recent years.

Figure 1: Cybercrimes in India (annual totals)



Source: Compiled by author based on data from National Crime Records Bureau (NCRB) (2018–2023)

Figure 2: Cybercrimes in India (annual totals per 100,000 people)



Source: Compiled by author based on data from NCRB (2018–2023)

India's legal framework for cybersecurity includes some provisions in the Information Technology Act, 2000, the Digital Personal Data Protection Act, and guidelines by sectoral regulators (MeitY, 2025a). These provisions and their current enforcement mechanisms may be inadequate to deal with the scale and sophistication of the cybersecurity threats that generative AI might lead to (MeitY, 2025a). Despite the statistics cited above, the IndiaAI mission does not include a strong focus on cybersecurity. Moreover, the draft National Cybersecurity Strategy, prepared in 2021 and reformulated in 2023, has not yet been adopted (ETTelecom, 2023). Also, there are no specific protection policies for India's AI-related infrastructure, and there is an ongoing administrative conflict between different departments over the governance of India's nodal cybersecurity agency, the Indian Computer Emergency Response Team (CERT-In) (Barik, 2024).

Narrow AI talent and education policies

Many of the individual policies that comprise India's AI strategy focus on talent development and education. For instance, the IndiaAI mission has a component called FutureSkills, which plans to increase AI programmes at all levels of education and includes a focus on AI courses in smaller cities (IndiaAI, n.d.; Prime Minister of India, 2024). The MeitY IndiaAI Expert Group has recommended the creation of a model curriculum for AI, "upskilling" the non-IT workforce, faculty training in AI, encouraging faculty to collaborate with industry, and the creation of an India-specific AI community (MeitY, 2023). There are various government funding mechanisms for AI training and Centres of Excellence in educational institutions. The Department of Science and Technology has instituted technology innovation hubs in various colleges (DST, n.d.).

Despite AI talent and education policies that appear to tick all the necessary boxes, India's approach lacks strategic direction. Many engineers and scientists who work at the foremost AI companies in the world have studied at Indian institutes, but a large proportion of Indian AI scientists prefer to work in the United States (Zwetsloot et al., 2021). Nonetheless, globally India ranks second in terms of AI talent, behind only the US (Mostrous et al., 2024). That this high prevalence of AI-relevant skills is not translating into global leadership in AI innovation for India is a policy failure. A reading of policy documents, including the IndiaAI expert group recommendations referred to above, shows that India's AI talent and education policies are limited to training students and workers in the technology and methods already developed in other countries, rather than promoting the development of new AI technology and methods in India. India can do much more with its large science and technology talent base, including the public funding of ambitious projects like AI development beyond deep-learning methods towards more fundamental explorations into other approaches to AI.

5. Fundamental issues

Absence of algorithmic accountability

Algorithmic accountability is a critical KASE pillar. Algorithms and their functioning can both lead to social problems, and be instruments of regulation (Belli, 2023). Despite the proliferation of algorithmic decision-making in various parts of daily life in India, and the emergence of AI technology that is likely to increase this prevalence, India does not have any laws that provide for the general governance of algorithmic decision-making. Some sectoral regulations govern a narrow slice of activity, for instance, the regulation of software as medical devices (Lenin, 2024). A proposed Digital India Act is likely to contain provisions on algorithmic transparency and periodic risk assessments, but its introduction has been repeatedly delayed (Sur, 2024). The vast majority of algorithmic decision-making remains ungoverned. It is particularly concerning that AI-driven state surveillance is conducted without a clear legal basis, such as through the use of facial recognition technology by the police, or through the use of a Covid-19 contact-tracing app made mandatory in many contexts during the pandemic (Bhandari & Rahman, 2020; Jauhar, 2021).

In the past few years, Indian gig workers have protested against the arbitrary blocking of their accounts, opaque rating systems, unilateral changes in payment structures, and burdensome requirements to prove eligibility for monetary incentives (Singh, 2023b). Gig workers in India work much longer than eight-hour days, are locked into platforms due to penalties for rejecting gigs, and make far less than minimum wage. The commonality driving these elements of exploitation is management through algorithms that are increasingly AI-driven. Algorithmic management is not limited to platform work—call centre workers, IT employees, and even lawyers are also subjected to granular, digitalised control over their work. Similarly, algorithms affect medical decisions, financial markets, and criminal justice in India, and underpin its surveillance architecture, in particular the use of facial recognition technology at airports and by law enforcement.

Weak data protection

While the robust approach to data protection in India, grounded in fiduciary responsibility, has been outlined above, the actual data protection law fails to rise to the challenges of widespread AI use. The Digital Personal Data Protection Act does not apply to personal data that is made publicly available (Article 3(c)(ii)). This means that no protections are available for such data, including protection against the use of this data for training AI models in ways that potentially violate the principles of purpose limitation and data minimisation (Pahwa, 2023). People who post data, creative work, or intellectual work on the internet do not do so with the expectation that this work will feed into the training of AI models, particularly without compensation. India's data protection law also does not enshrine any rights in relation to automated decision-making, unlike many other data protection laws around the world (Apacible-Bernardo et al., 2023). With AI being implemented to automate decisions across various sectors, this omission is significant.

Frequent internet shutdowns

Another aspect of Indian technology policy that casts a shadow on its ambition for growth and inclusion is the frequent state-directed internet shutdowns. India consistently has the highest number of internet shutdowns globally. Table 1 below, based on data from Access Now (n.d.), shows the annual number of internet shutdowns in India between 2016 and 2023, compared to the country with the second-largest number of shutdowns in the same year. Since Access Now started collecting annual data on internet shutdowns in 2016, India has always had the highest number.

Table 1: Internet shutdowns in India and in country with the second largest no. of shutdowns

Year	No. of shutdowns in India	Shutdowns in the country with second-largest no.	Country with second-largest no. of shutdowns
2016	30	8	Pakistan
2017	69	10	Pakistan
2018	134	13	Pakistan
2019	121	12	Venezuela
2020	108	6	Yemen
2021	106	15	Myanmar
2022	84	22	Ukraine
2023	116	37	Myanmar

Source: Compiled by the author using data from Access Now (n.d.)

India's shutdowns tend to be concentrated in specific regions. For instance, more than 60% of the country's internet shutdowns in 2022 were imposed in Jammu and Kashmir, a region subject to militarisation and crackdowns on political activity (Mogul, 2023). It is difficult to argue that digital services, including AI services, can be provided in a reliable, non-exclusionary, and unbiased manner in a country that is subjected to so many internet shutdowns.

6. Conclusions

It has become clear that some policy areas require more urgent attention than others, while all policy areas could benefit from greater inter-linking and orientation towards common goals. India could benefit from a more explicit political understanding of global AI dominance and AI political–economic sovereignty, and how to leverage its relationships with the US, China, and the EU to protect its own interests. Public subsidies for computational power are potentially good, but we should be careful not to shift the risks of AI-building to the public sector, while allowing the private sector to capture the rewards through building applications that may not be in the public interest. Options for government equity and benefit-sharing are currently underexplored in India's provision of public financial support to private-sector AI development.

Finally, it is inadvisable for India to merely follow the lead of the US (and, indeed, other countries including China and the EU nations) in both technology and technology policy. US technology policy is (understandably) in flux, with priorities shifting from one administration to the next. For instance, antitrust, which was a priority for the Biden administration, is not a priority for the current Trump administration. Meanwhile, India continues work on its Digital Competition Bill (Kumar, 2024). Realities of concentration in AI and digital markets have not changed, and with a revival of mergers, acquisitions, and overall monopolisation in the US, Indian policymakers need to redouble their efforts towards promoting competition domestically.

Acknowledgements

The author thanks Luca Belli, Larissa Galdino de Magalhaes Santos, Amrita Sengupta, Mila Samdub, and two anonymous reviewers for their comments and suggestions.

Funding

Funding for this study was received from the FGV Law School's 2024 CyberBRICS Fellowship Program.

Data availability

The data supporting the results of this study is available upon written request to the author at jv474@cornell.edu

AI declaration

The author did not use any AI tools for conducting the research or writing the article.

Competing interests declaration

The author has no competing interests to declare.

References

Access Now. (n.d.). *Ending internet shutdowns.* Retrieved February 11, 2025, from https://www.accessnow.org/issue/internet-shutdowns

Agrawal, A. (2024, March 15). In revised AI advisory, IT ministry removes requirement for govt permission. *Hindustan Times*. https://www.hindustantimes.com/india-news/in-revised-ai-advisory-it-ministry-removes-requirement-for-government-permission-101710520296018.html

Apacible-Bernardo, A., Sonkar, S., & Chakraborty, S. (2023). Top 10 operational impacts of India's DPDPA – Comparative analysis with the EU General Data Protection Regulation and other major data privacy laws. International Association of Privacy Professionals. https://iapp.org/resources/article/operational-impacts-of-indias-dpdpa-part6

Arya, A. (2021). Submarine telecommunication cable infrastructure regime in India: An analysis on the Indian legal and regulatory regime. *The Indian Journal of Projects, Infrastructure, and Energy Law*, *1*(1), 102–114. https://ijpiel.com/wp-content/uploads/2022/02/9_Submarine-Telecommunication-Cable-Infrastructure-Regime-in-India.pdf

Aulakh, G. (2023, July 20). *Trai recommends regulatory framework for AI, risk-based framework for AI specific use cases*. Mint. https://www.livemint.com/technology/tech-news/trai-issues-recommendations-on-ai-says-regulatory-framework-for-development-of-responsible-ai-urgently-needed-11689859911432.html

Bailey, R., & Goyal, T. (2020, January 13). *Fiduciary relationships as a means to protect privacy: Examining the use of the fiduciary concept in the draft Personal Data Protection Bill, 2019*. The Leap Blog. https://blog.theleapjournal.org/2020/01/fiduciary-relationships-as-means-to.html

Barik, S. (2024, July 13). Both Home and IT ministries pitch for control of nodal cyber security watchdog Cert-In. *The Indian Express.* https://indianexpress.com/article/india/both-home-and-it-ministries-pitch-for-control-of-nodal-cyber-security-watchdog-cert-in-9450203

Basu, A. (2025, May 13). *Data diplomacy: Rethinking cross-border data flows for a more equitable global digital economy*. Planetary Politics. https://www.newamerica.org/planetary-politics/blog/data-diplomacy-rethinking-cross-border-data-flows-for-a-more-equitable-global-digital-economy

Belli, L. (2023). Exploring the key AI sovereignty enablers (KASE) of Brazil, towards an AI sovereignty stack. Pre-print version. In Carnegie Endowment for International Peace (Ed.), *Digital Democracy Network Conference 2023 essay collection*. Buenos Aires. https://doi.org/10.2139/ssrn.4465501

Bhandari, V., & Rahman, F. (2020, May 25). *Constitutionalism during a crisis: The case of Aarogya Setu*. The Leap Blog. https://blog.theleapjournal.org/2020/05/constitutionalism-during-crisis-case-of.html

Bruce, M., Lusthaus, J., Kashyap, R., Phair, N., & Varese, F. (2024). Mapping the global geography of cybercrime with the World Cybercrime Index. *PLoS ONE*, *19*(4). https://doi.org/10.1371/journal.pone.0297312

BW Businessworld. (2025, January 30). IndiaAI Mission: Govt seeks proposals for foundational models, 18K GPU facility to debut soon. https://www.businessworld.in/article/indiaai-mission-govt-seeks-proposals-for-foundational-models-18k-gpu-facility-to-debut-soon-546497

Chandrasekhar, C., & Ghosh, J. (2021, August 23). The rising spectre of a telecom monopoly. *The Hindu Business Line.* https://www.thehindubusinessline.com/opinion/columns/c-p-chandrasekhar/the-rising-spectre-of-a-telecom-monopoly/article36063279.ece

CyberBRICS. (2024). *Digital public infrastructure for sovereign AI: A view from the BRICS* [Webinar]. https://cyberbrics.info/webinar-digital-public-infrastructure-for-sovereign-ai

Department of Science and Technology (DST). (n.d.). *25 Technology Innovation Hubs across the country through NM-ICPS are boosting new and emerging technologies to power national initiatives.* https://dst.gov.in/25-technology-innovation-hubs-across-country-through-nm-icps-are-boosting-new-and-emerging

ETTelecom. (2023, February 20). National Cybersecurity Strategy 2023 may come out soon: Pant. *The Economic Times.* https://telecom.economictimes.indiatimes.com/news/national-cybersecurity-strategy-2023-may-come-out-soon-pant/98093316

Indian Space Research Organisation (ISRO). (2023). *Indian Space Policy – 2023.* https://www.isro.gov.in/media_isro/pdf/IndianSpacePolicy2023.pdf

Jauhar, A. (2021). *Indian law enforcement's ongoing usage of automated facial recognition tech – ethical risks and legal challenges*. Vidhi Centre for Legal Policy. https://vidhilegalpolicy.in/wp-content/uploads/2021/08/210805_FRT_Paper1_Primer-Lit-Review_final.pdf

Kazia, N. A., Sinha, S., & Agarwal, S. (2025). *Consent managers: An Indian solution for managing consent.* International Bar Association. https://www.ibanet.org/consent-managers-Indian-solution

Kumar, D. (2024, July 1). Impact of Digital Competition Bill on India's homegrown startup ecosystem. *Business Standard.* https://www.business-standard.com/companies/start-ups/impact-of-digital-competition-bill-on-india-s-homegrown-startup-ecosystem-124070101008_1.html

Lenin, B. (2024). *India – Regulating software as medical devices – Navigating hurdles one byte at a time.* Conventus Law. https://conventuslaw.com/report/india-regulating-software-as-medical-devices-navigating-hurdles-one-byte-at-a-time

McDougal, L., Raj, A., & Singh, A. (2022, January 16). The digital divide and is it holding back women in India? *Hindustan Times.* https://www.hindustantimes.com/ht-insight/gender-equality/the-digital-divide-and-is-it-holding-back-women-in-india-101641971745195.html

Ministry of Communications. (2024). *Universal Connectivity and Digital India initiatives reaching all areas.* https://static.pib.gov.in/WriteReadData/specificdocs/documents/2024/aug/doc202486366801.pdf

Ministry of Electronics and Information Technology (MeitY). (n.d.). *FAQs.*

MeitY. (2023a). *IndiaAI 2023: Expert Group Report – First edition.* https://indiaai.gov.in/news/indiaai-2023-expert-group-report-first-edition

MeitY. (2023b). *Production Linked Incentive Scheme – PLI 2.0 for IT Hardware*. https://www.meity.gov.in/static/uploads/2024/02/Production-Linked-Scheme-2.0-for-IT-Hardware-notification_0.pdf

MeitY. (2024). *Due diligence by Intermediaries / Platforms under the Information Technology Act, 2000 and Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021.* https://docs.google.com/document/d/1nc45Bqk2gU3GC9Say7rIB0xJulfcd7jdjdHaMSrHXPs/mobilebasic

MeitY. (2025a). *Report on AI governance guidelines development*. https://indiaai.s3.ap-south-1.amazonaws.com/docs/subcommittee-report-dec26.pdf

MeitY. (2025b). *Request for proposal (RFP) for augmentation & enhancement of existing 8-inch fab of Semi-Conductor Laboratory (SCL), India.* https://www.meity.gov.in/static/uploads/2025/02/d60485e4181a949761bd4d4b6ab2799e.pdf

Mishra, A. (2024, July 4). Govt to use 50% of India AI mission funds for GPU procurement: MeitY. *Business Standard.* https://www.business-standard.com/technology/tech-news/govt-to-use-50-of-india-ai-mission-funds-for-gpu-procurement-meity-124070400728_1.html

Modi, N. [@narendramodi]. (2025, January 30). *When it comes to the space sector, bet on India!* [Post]. X. https://x.com/narendramodi/status/1884970434405535865

Mogul, R. (2023, March 1). India, world's largest democracy, leads global list of internet shutdowns. *CNN.* https://edition.cnn.com/2023/03/01/tech/internet-shutdowns-india-report-intl-hnk/index.html

Mohanty, A. & Sahu, S. (2024). *India's advance on AI regulation*. Carnegie India. https://carnegieendowment.org/research/2024/11/indias-advance-on-ai-regulation?lang=en

Mostrous, A., White, J., & Cesareo, S. (2024). *The Global Artificial Intelligence Index.* Tortoise. https://www.tortoisemedia.com/2024/09/19/the-global-artificial-intelligence-index-2024

National Crime Records Bureau (NCRB). (2018–2023). *Crime in India.* Open Government Data (OGD) Platform India. https://www.data.gov.in/ministrydepartment/National%20Crime%20Records%20Bureau%20(NCRB)

NITI Aayog. (2018). *National Strategy for Artificial Intelligence: #AIforAll*. https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf

NITI Aayog. (2021a). *Responsible AI. #AIforAll. Approach Document for India: Part 1 – Principles for Responsible AI.* https://niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf

NITI Aayog. (2021b). *Responsible AI. #AIforAll. Approach Document for India: Part 2 - Operationalizing Principles for Responsible AI*. https://www.niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf

Pahwa, N. (2023, September 1). *How will India's Digital Personal Data Protection Law impact artificial intelligence?* [Video]. MediaNama*.* https://beta.medianama.com/2023/09/223-india-data-protection-law-impact-ai

Parsheera, S. (2022). Understanding state-level variations in India's digital transformation. *The African Journal of Information and Communication (AJIC)*, *30*, 1–9. https://doi.org/10.23962/ajic.i30.15082

Prasad, S. K. (2019). *Information fiduciaries and India's Data Protection Law*. Data Catalyst. https://datacatalyst.org/wp-content/uploads/2020/06/Information-Fiduciaries-and-Indias-Data-Protection-Law.pdf

Prime Minister of India. (2024, March 7). Cabinet approves ambitious IndiaAI mission to strengthen the AI innovation ecosystem. https://www.pmindia.gov.in/en/news_updates/cabinet-approves-ambitious-indiaai-mission-to-strengthen-the-ai-innovation-ecosystem

Republic of India. (2023). The Digital Personal Data Protection Act, 2023. https://egazette.gov.in/WriteReadData/2023/248045.pdf

Reserve Bank of India. (2018). *Storage of payment system data*. DPSS.CO.OD No.2785/06.08.005/2017-2018. https://www.rbi.org.in/Scripts/NotificationUser.aspx?Id=11244&Mode=0

Saxena, P. (2021, June 11). *RAISE 2021 – Leveraging data for AI towards social empowerment*. IndiaAI. https://indiaai.gov.in/article/leveraging-data-for-ai-towards-social-empowerment

SEON. (2023). *Global cybercrime report: Which countries are most at risk in 2023?* https://seon.io/resources/global-cybercrime-report/ and https://assets.cdn.seon.io/uploads/2023/04/Cybersecurity_countries-min.pdf

Sheriff, K. (2020, December 14). NFHS data shows urban-rural, gender gaps in Internet use. *The Indian Express.* https://indianexpress.com/article/india/nfhs-data-shows-urban-rural-gender-gaps-in-internet-use-7103710

Singh, J. (2023a, January 25). *India's gig economy drivers face bust in the country's digital boom*. TechCrunch. https://techcrunch.com/2023/01/25/india-gig-workers-problems/?guccounter=1

Singh, M. (2023b, April 5). *India opts against AI regulation*. TechCrunch. https://techcrunch.com/2023/04/05/india-opts-against-ai-regulation

Singh, M. (2024, March 3). *India reverses AI stance, requires government approval for model launches*. TechCrunch. https://techcrunch.com/2024/03/03/india-reverses-ai-stance-requires-government-approval-for-model-launches

Singh, T., Shivraj, A., Madaan, S., & Nigam, M. (2025). *Unlocking AI's potential in India: Transforming agriculture and healthcare*. Boston Consulting Group. https://www.bcg.com/publications/2025/india-unlocking-ai-potential-in-india-transforming-agriculture-and-healthcare

Sur, A. (2024, July 1). *Digital India Bill likely to be delayed, government may opt for smaller, urgent regulations*. Money Control. https://www.moneycontrol.com/technology/digital-india-bill-likely-to-be-delayed-government-may-opt-for-smaller-urgent-regulations-article-12759435.html

Suraksha, P., & Lohchab, H. (2024, April 10). AI compute mission best served by marketplace model, say experts. *The Economic Times.* https://economictimes.indiatimes.com/tech/tech-bytes/marketplace-model-best-option-for-govt-to-offer-ai-compute-capacity-to-innovators-say-experts/articleshow/109169175.cms

Suri, A. (2025). *The missing pieces in India's AI puzzle: Talent, data, and R&D*. Carnegie India. https://carnegieendowment.org/research/2025/02/the-missing-pieces-in-indias-ai-puzzle-talent-data-and-randd?lang=en

Telecom Regulatory Authority of India (TRAI). (2023, June 19). *TRAI releases recommendations on 'Licensing Framework and Regulatory Mechanism for Submarine Cable Landing in India'* [Press release]. https://trai.gov.in/sites/default/files/2024-08/PR_No.54of2023.pdf

Varadarajan, L. (2025). Imperialism, the Third World and the fundamental continuities in Indian foreign policy. *Studies in Indian Politics*, *13*(1), 75–85. https://doi.org/10.1177/23210230251325599

Vipra, J. (2024). *A compute agenda for India*. Evam Law & Policy. https://cyberbrics.info/a-compute-agenda-for-india

Zwetsloot, R., Zhang, B., Anderljung, M., Horowitz, M., & Dafoe, A. (2021). *The immigration preferences of top AI researchers: New survey evidence.* Centre for the Governance of AI. https://www.governance.ai/research-paper/the-immigration-preferences-of-top-ai-researchers-new-survey-evidence

Russia's securitised approach to AI sovereignty

Alexander Ignatov
Senior Research Fellow, Center for International Institutions Research, Russian Presidential Academy of National Economy and Public Administration (RANEPA), Moscow; and Visiting Scholar, CyberBRICS project, Center for Technology and Society (CTS), Fundação Getulio Vargas (FGV) Law School, Rio de Janeiro
https://orcid.org/0000-0001-6740-4454

Danil Kerimi
Doctoral Candidate, School of Social Sciences and Technology, Technical University of Munich
https://orcid.org/0009-0002-7235-4257

Abstract

In the context of the world's major powers competing for dominance in the artificial intelligence (AI) realm, Russia aims to become a global leader in AI development. This article evaluates Russian AI governance through the lenses of the key AI sovereignty enablers (KASE) framework and the Copenhagen School's securitisation theory. The Russian government's approach to AI governance, in line with its broader approach to digital governance, grants extensive powers to state security and law enforcement entities, while major domestic AI market players are state-influenced. This securitised approach to AI sovereignty and governance stems from concerns about the country's stability, alongside a high degree of politicisation of digital governance. The article argues that the likely impact of Russian securitisation of AI governance will be further consolidation of state control over AI innovations and a narrowing of the space for non-state technological developments.

Keywords
Russia, artificial intelligence (AI), AI sovereignty, securitisation, digital economy

DOI: https://doi.org/10.23962/ajic.i35.20626

Recommended citation

Ignatov, A., & Kerimi, D. (2025). Russia's securitised approach to AI sovereignty. *The African Journal of Information and Communication (AJIC)*, *35*, 1–11. https://doi.org/10.23962/ajic.i35.20626

1. Introduction

As the heir to the Soviet Union's scientific legacy, Russia has been keen to highlight its modern digital power and its prominent global AI ambitions. As an integral component of the country's efforts towards digital sovereignty, AI-based solutions attract growing attention from Russia's leadership, as exemplified by an ambitious agenda in terms of which Russia is to reach the top ranks of global AI powers by 2030 (President of Russia, 2019b). Nonetheless, in addition to the already tough competition for global AI leadership, Russia's efforts are further complicated by geopolitical/military conflict and economic sanctions.

In this study, we reviewed Russia's efforts to achieve AI sovereignty through the lens of Belli's key AI sovereignty enablers (KASE) framework (Belli, 2023), which is grounded in Belli's (2023) framing of AI sovereignty as "the capacity of a given country to understand, develop and regulate AI systems" (2023, p. 1). We also reviewed Russia's AI governance through the lens of securitisation theory, as set out by the Copenhagen School (Buzan et al., 1998). This securitisation lens allowed us to explore the differentiated weight that the Russian government places on the KASE dimensions.

As set out in this article, we found that, as a product of Russia's current geopolitical context, AI is often viewed by the country's leadership as a national military, political, and economic security priority, i.e., AI is viewed as a security/securitisation matter. Consequently, the state actors that deal with cyber, information, data, and energy security matters are given considerable financial support. We conclude that it is highly likely that, in the foreseeable future, the under-prioritised (from the securitisation perspective) dimensions will follow the same path, thus completing the securitisation of AI within Russia's public discourse and policy framework. This article also provides insight into the general stance of Russia's leadership towards emerging AI-based solutions and potential approaches to the development of a market regulatory framework.

2. Challenges to Russian AI development

Russia's proclaimed goal to become a global AI leader aligns with the country's overarching ambitions of building a strong digital state (President of Russia, 2019b). Russia's political and economic elites are determined not to miss out on current technological trends that are firmly driven by AI. In the 2020s, the laissez-faire approach to digital regulation ended in many parts of the world. Russia is among the countries taking control of the next stage of technological evolution—away from the business community and towards the state (Zinovieva, 2024). However, Russia is faced with additional challenges in respect of talent, compute power, and capital dimensions. These challenges include: the brain drain due to fears generated by the conflict with Ukraine, an underdeveloped hardware ecosystem, and budgetary pressures due to sanctions and competing investment priorities.

According to a statement on the recently approved state budget proposal for 2025, Russia is increasing its military budget by 25% to RUB13.5 trillion (USD145 billion), which is equivalent to 6.31% of the country's GDP (Miller, 2024). As the military budget grows, other areas will inevitably suffer. Thus, according to the same budget proposal, civilian research will shrink by a quarter (Statista, 2024). According to the Institute for Statistical Studies and Economics of Knowledge of the Russian Higher School of Economics, the RUB458 billion (USD4.9 billion) that was dedicated to applied research in 2024 will be reduced to RUB362 billion (USD3.5 billion) in 2025, and to RUB260 billion (USD2.6 billion) in 2026 (Gerden, 2024). To put these figures into perspective, Google's parent company Alphabet alone spends 10 times more than Russia on applied research. In dollar terms, Russia will spend about the same amount on combined research and development (R&D) in 2025 as Portugal. On average, in recent years, Russia has spent about 1% of its GDP on R&D activities. This is less than what is spent by countries such as Malaysia and Egypt and is less than half of the average in Organisation for Economic Co-operation and Development (OECD) countries (OECD, 2023).

For Russia's Federal AI Programme, the same budget proposal provides RUB1.145 trillion (USD11.5 billion) allocated in 2025, which is similar to what is spent annually on R&D by a single Chinese technology conglomerate, Tencent (2024). For 2026, the budget projects RUB1.25 trillion (USD12 billion) and in 2027 RUB1.5 trillion (USD14 billion) (D-Russia, 2024). At the same time, however, additional funds will be allocated to defence-related AI through the aforementioned increased military budget.

In addition to financing AI development, another challenge faced by Russia is an AI talent shortage. In 2022–23, the outflow of IT specialists was estimated at more than 20,000 individuals (*Realnoye Vremya*, 2024).[1] Another estimate claimed that, in the first half of 2022, the outflow number surpassed 40,000 (RBC, 2022). The Ministry of Digital Development, Communications and Mass Media reported that, in 2022, at least 100,000 IT specialists left the country and only 10,000 returned (Inclient, 2022). While some of the leading Russian software companies claim that the outflow of IT specialists is not affecting them (Telecom Daily, 2024), the labour market shows a growing demand for specialists with no considerable or even any relevant experience, as hiring requirements soften to accommodate the shortage of personnel. In Tatarstan, where one of the biggest Russian IT hubs, Innopolis, is located, the growth in demand for IT specialists has been estimated at 103% (*Realnoye Vremya*, 2024).

[1] The shortage of capable AI specialists was on the agenda even before this period (see Nadibaidze, 2022).

Perhaps the biggest challenge that Russia must overcome on the road to a robust AI development ecosystem is the lack of hardware availability. This inherited weakness of the Russian computing ecosystem has been further worsened by the sanctions imposed by foreign countries (Kolomychenko, 2024). Russia's domestic production of electronic components is negligible by global standards. Within the framework of the state programme to support the electronic industry, projects such as the Baikal microprocessors production facility (Baikal Electronics, n.d.; Bendett, 2024) have been implemented to organise the production of microprocessors using domestic technologies, with progress towards localisation of the production chain announced in March 2024 (Kholupova, 2024). Nevertheless, the main production of Russian processors continues to be outsourced to labs outside the country, such as Taiwan's TSMC (Urusov, 2023).

Russia's major digital market players are either state-owned or have significant ties to the state, which means that commercial practices are often influenced by the state's national digital-sovereignty priorities (Petrella et al., 2021). For instance, Russian IT companies are, as in most other countries, obliged to give law enforcement and intelligence agencies access to users' data only with court authorisation. However, there are frequent reports in the independent media of Russian digital platform firms such as VK granting access to users' data based on a simple "telephone call" from a law enforcement or intelligence agency—even when such data transfer should result in a criminal prosecution for the firm concerned (Sidelnikova, 2024).

3. Analytical tools: The KASE framework and securitisation theory

We deployed two analytical tools in our evaluation of the state of Russian AI governance: the KASE framework and securitisation theory.

KASE framework

The KASE framework put forward by Belli (2023) sets out eight dimensions as crucial to a country's progress towards AI sovereignty:

- data governance;
- algorithmic governance;
- computational capacity;
- meaningful connectivity;
- reliable electrical power;
- digitally literate population;
- strong cybersecurity; and
- appropriate regulatory framework.

In our KASE evaluation we used a mapping tool (see Appendix) that we developed with colleagues in the CyberBRICS project (CyberBRICS, n.d.).

Securitisation theory

In addition to the KASE framework and mapping, we see securitisation theory as a useful analytical lens for exploring AI sovereignty dimensions in various countries, and particularly in Russia (Stix, 2022). In many countries, the officials who are now dealing with AI regulation were previously in charge of cyber policy (**Ü**nver, 2024). Cyber policy, for its part, was in many cases built upon counter-terrorism work (UNICRI & UNOCT, 2021a; 2021b). Since the early 2000s, at national and international levels, the work of counter-terrorism experts has been compelled to evolve into AI-focused responsibilities. A line can be traced from post9/11 (2001) counter-terrorism capacity-building through to cybersecurity regimes (e.g., the Tallinn Manual and DHS cyber strategies), with personnel and frameworks then migrating, both operationally and institutionally across national and multilateral levels, into the nascent domain of AI governance (Bianchi & Greipl, 2022; Pfaff, 2025; Tallberg et al., 2023; US Department of the Treasury, 2024).

For example, the G7's 2023 AI Principles emphasise AI security risk management and trace their heritage to cybersecurity norms originally designed for counter-terrorism-inspired threats (EC, 2023). Similarly, OECD statements underscore growing synergies between cybersecurity and AI governance, shaped by counter-terrorism risk frameworks and cyber-risk protocols (OECD, 2024). Meanwhile, in the private sector, those overseeing, for instance, anti-money laundering in the financial services sector, are the ones at the forefront of AI adoption with significant technology budgets (US Department of the Treasury, 2024).

Securitisation theory is closely associated with works by Buzan, Wæver, and de Wilde, collectively referenced as the Copenhagen School (Buzan et al., 1998). The theory explains how a part of objective reality becomes viewed as a threat to a referent object, e.g., a state, a person ora group of people dependent on a sphere of interest, namely the economy, society, the military, policy, or the environment. Grounded in identification of the threat by the state, securitisation presents an argument whereby the state advocates implementation of extraordinary measures to counter the threat, even when the measures may contradict established rules. Securitisation can also be viewed as the process by which non-politicised issues (issues not talked about, or not part of public debate) or politicised issues (issues already publicly debated) are elevated to security issues that need to be dealt with as a matter of urgency and that reqjuire bypassing of procedures for public debate and democratic engagement.

According to Charrett (2009), a prominent example of a securitised issue is terrorism. The dramatic changes in US foreign policy after the 9/11 attacks of 2001, which eventually resulted in a US-led invasion of Iraq in 2003, became possible due to US President George W. Bush's use of enhanced executive powers grounded in a securitisation of the "meaning of 9/11" (Charrett, 2009) as something requiring harsh, responsive actions by a state unfettered by normal procedural checks and balances. The proclamation of the US response to 9/11 as a "Global War on Terror" (National Archives, n.d.) led to significant expansion of presidential powers, spying on ordinary Americans, detention of Muslims and Arabs, and establishment of a secretive military tribunal system—with most of these elements remaining in place despite protracted debate and sustained efforts to roll them back in order to ensure the separation of powers and stability of the democratic order (Charisle, 2021).

In summary, a completed securitisation means that: (1) the state provides the public with an argument framing a referent object as threatened; (2) there is a stated demand to exercise extraordinary measures to protect the referent object; and (3) justification is provided for the state to break established rules in order to protect the referent object. In this study, we employed the securitisation lens as a means to explore bureaucratic and structural tendencies in the Russian state's approach to AI development, and to build a picture of the future of the country's AI governance model.

4. KASE findings

Data governance

Russia does not have a specialised data governance strategy, but it has a comprehensive framework with clearly assigned responsibilities and practical regulatory systems. The Ministry of Digital Development, Communications and Mass Media leads data management, security, and regulatory policies, alongside Roskomnadzor (the Federal Service for Supervision of Communications, Information Technology and Mass Media) and the Federal Security Service (FSB). The primary data governance law is the Federal Law on Personal Data (No. 152-FZ) (Russian Federation, 2006), supported by additional laws on information and critical infrastructure protection (an overlap with the cybersecurity domain). Core funding comes from government-affiliated funds such as the Russian Science Foundation and the Skolkovo Foundation.

Although Russia lacks an explicit international strategy for AI and data governance, its stance in the international arena—in institutions such as BRICS, the G20, and the UN/UNESCO—has some fundamental features that can be taken as bearing strategic significance, namely Russia's adherence to state-centric multilateralism and its rejection of multistakeholder approaches in which states and non-state actors cooperate.

Algorithmic governance

Leading Russian national enterprises such as Sber (GigaChat), Yandex (Neuro) and VK (all three are directly or indirectly managed by the government) have developed their own large language models (LLMs) and drive AI innovation in Russia, alongside a growing AI startup ecosystem that often collaborates with larger corporations and research institutions. Although an "algorithm strategy" is not specified, the National

Strategy for the Development of Artificial Intelligence (President of Russia, 2019b) emphasises deploying algorithms in priority spheres such as healthcare, education, and transportation, with involvement from government agencies such as the Ministry of Science and Higher Education and the Ministry of Digital Development, Communications and Mass Media. At the time of writing, in early 2025, AI regulation discussions were ongoing in Russia's State Duma (the Parliament's lower chamber), particularly around matters of transparency and accountability, but comprehensive algorithm-specific or LLM-focused laws had yet to be promulgated.

Computational capacity

Russia's Strategy for the Development of the Electronic Industry until 2030 (Government of Russia, 2020) emphasises expanding hardware production, including storage solutions and server hardware, with multiple ministries involved, led by the Ministry of Industry and Trade. Import substitution is a priority, targeting the production of processors, controllers, and memory, and the advancement of silicon technologies to the 5 nanometre (5nm) level for eventual domestic production. State funding for AI and microelectronics R&D has begun to increase significantly, with 2024 investments reaching RUB5.2 billion (USD51.6 million) for AI projects (Norem, 2024). Russia has six supercomputers in the global TOP500 index, with Yandex's Chervonenkis ranked highest among the six, in 75th position globally (TOP500, n.d.). Government-supported enterprises such as Rostec, and private-sector-led (with varying degrees of state ownership) entities such as Sber (50% state-owned), drive growth in computational capacity in the domestic AI sector.

Meaningful connectivity

Infrastructure is considered a backbone of Russia's security, grounded in the notion of "critically important information infrastructure" (Consultant Plus, 2017). The FSB is directly involved in providing protection for critical information infrastructure. The International Telecommunication Union (ITU) (n.d.-a) ranks Russia highly for internet affordability, with the country offering some of the lowest internet costs globally. In 2024, the entry-level fixed-broadband basket cost in Russia was 0.57% of GNI per capita, compared with the global average of 2.66% (2023) (ITU, n.d.-b). Over 92% of the Russian population (with both genders equally represented) use the internet regularly, with 83.1% of rural households and 89.5% of urban residents having internet access at home. Younger users (aged 15–24) have a high internet usage rate (98.7%), while engagement is lower (89.2%) among older generations (25–74 years). Russia's Strategy for the Development of the Communications Industry until 2035 (Government of Russia, 2023), led by the Ministry of Digital Development, states that the fixed telecommunications sector needs more investment due to high costs and potential infrastructure challenges. Russia is connected to multiple submarine cables, most of them domestic, with several of the domestic cables, such as the Polar Express, designed to enhance internal connectivity across regions.

Reliable electrical power

According to the International Energy Agency (IEA, n.d.), Russia's electricity production primarily depends on natural gas (45.1%), with nuclear energy (19.4%), hydropower (17.3%), and coal (16.3%) also playing significant roles. Renewable energy, comprising wind and solar, contributes a small share (3.54% in 2021). The Strategy for the Development of the Electric Power Industry of the Russian Federation (Government of Russia, 2013) aims to modernise and diversify the energy sector, with oversight by the Ministry of Energy. Key regulatory bodies include the Federal Grid Operator, which manages electricity transmission, and the Federal Antimonopoly Service, which maintains competition in the electricity market. Electricity market regulation and the energy industry stability at large are considered matters of utmost importance, with national security concerns involved (Government of Russia, 2019).

Digitally literate population

In 2023, Russian President Vladimir Putin directed an update to Russia's National Strategy for the Development of Artificial Intelligence through 2030 (President of Russia, 2019b), emphasising support for AI research centres along with increased government expenditure. The Russian AI market grew by 18% in 2022, reaching RUB650 billion (USD6.4 billion) (Consultant Plus, 2017), and the government planned to invest RUB5.2 billion (USD51.6 million) in AI in 2024 (Interfax, 2025). The AI Strategy promotes comprehensive AI education, aiming to integrate AI topics across educational levels, to develop specialised degrees, and to enhance practical training. By 2023, Russia had approximately 17,000 AI graduates (2021–2023) (ComNews, 2024), 70% growth in AI publications in top-tier journals (2019–2023) (Analytical Center, 2024), and 96 approved AI standards (2019–2023) (Government of Russia, n.d.).

The Ministry of Science and Higher Education is tasked by the National Strategy with implementing educational aspects, supported by partnerships with major universities. Furthermore, Russia's AI Alliance, including major tech firms such as Sber and Yandex, supports talent development initiatives (AI Alliance Russia, n.d.). Due to geopolitical tensions, international AI collaboration is limited, with BRICS serving as the primary partner. Russia is aiming for a significant rise in AI-skilled graduates and high AI-readiness across priority economic sectors by 2030 (President of Russia, 2019b), but the lack of skilled labour, mentioned earlier in this article, constitutes a significant obstacle in this respect.

Strong cybersecurity

In Russia, cybersecurity is guided by the National Security Strategy (President of Russia, 2021) and the Doctrine of Information Security (President of Russia, 2016), with the country's Security Council playing a central role in strategy oversight. Regulatory entities, such as Roskomnadzor (Federal Service for Supervision of Communications, Information Technology and Mass Media) and the Cybersecurity Department in the Ministry of Digital Development, are responsible for enforcing cyber regulations. Government funding supports R&D to reduce reliance on foreign technology, focusing on building a skilled domestic workforce and domestic cybersecurity solutions. Geopolitical pressures have led the country's private and public sectors to favour Russian-developed technologies (ComNews, 2023), with companies like Kaspersky Lab and Positive Technologies leading the industry (Kurasheva, 2023).

Appropriate regulatory framework

At the time of writing, in early 2025, Russia had not yet enacted any significant regulation targeting AI. The State Duma's major party Edinaya Rossia (United Russia) is said to have been working on a draft law on AI regulation since 2023—a law that would, inter alia, define AI solutions developers' responsibilities and prevent the use of AI for fraud (*Kommersant*, 2023). Also, in 2023, a draft law was presented to protect AI users against harm arising from AI. In July 2024, President Putin promulgated a law forcing AI developers to provide insurance against possible harm caused by their AI-based products (TASS, 2024). In early 2025, the State Duma created a working group on AI that has a mandate until 2026 to develop regulations (Dorofeeva et al., 2025).

5. Securitisation findings

AI technologies are often viewed as a source of threat to Russia's sovereignty and, especially, to the country's military security. According to President Putin, AI development "shall be constrained" as it would "inevitably lead to a point where they [AI technologies] may begin to pose a threat to humanity—comparable to the development of nuclear capabilities", with national governments around the world taking the lead in the process (President of Russia, 2023). AI as a threat is presented in the national AI strategy, e.g., the 2019 Presidential Decree (with 2024 amendments) approving the strategy includes a notion of AI as a tool for spreading "prohibited information" (President of Russia, 2019b; 2024).

Under Russia's current AI policy dispensation, most of the KASE framework dimensions are either already viewed through the securitisation prism or are on track to soon be viewed in such a manner. An important factor to consider in this process is the balance of power between ministries/agencies subscribing to securitisation and those subscribing to development, i.e., the guns versus butter paradigm. As in other parts of the Russian regulatory and budgetary apparatus, the *siloviki* (security agency personnel) at entities such as the FSB are partly responsible for data governance and cybersecurity policy implementation.[2] The aforementioned Roskomnadzor serves as a media supervisor and is also deeply involved in data governance. Meanwhile, the market champions include state ownership stakes and operate under state supervision, e.g., Sber, formerly Sberbank and by far the largest Russian bank (Ross, 2024); VK, the largest

[2] There is a widespread belief among scholars and policy experts that the influence of Russian security agencies extends broadly across the country's entire IT sector (see Epifanova & Dietrich, 2022).

Russian social media platform (SimilarWeb, 2021); and Yandex, the largest Russian search engine, with 72% of Russia’s market share. This supervision is conducted either directly, when the state enacts its powers as an owner, e.g., in Sber, with 50% of its shares owned by the state (Petrella et al., 2021); or indirectly, via proxy “oligarchs”, who are company owners tied to the state. The major funds supporting prominent innovation projects are mostly affiliated with the state.

As reliable energy supply has become a major concern for the development of AI worldwide, Russia is not unique in considering energy market stability as a matter of highest importance for AI development. Like other energy-rich countries (e.g., Saudi Arabia, the US), Russia seeks to showcase its capabilities as an “energy superpower”—referring to its ability to influence the global energy market and, in turn, the international agenda (Rutland, 2008). Russia’s Energy Security Doctrine of 2019 (President of Russia, 2019a) cites shrinking external markets, difficulties in reaching new markets, and the international climate and environmental agenda as major threats to the country’s stability. Also cited in this Doctrine is the wrongful use of information and communication technologies (ICTs) against information infrastructure in ways that may hamper the functionality of energy facilities.

At the time of writing in early 2025, the only examples of non-fully securitised KASE dimensions that our study had identified were (1) algorithmic governance; and (2) the regulatory framework. With respect to both these dimensions, the discussions that were in progress indicated that security considerations were poised to take the lead. Russian algorithmic governance, which some might view as an area characterised by public–private dialogue aimed at finding an appropriate common ground, is in reality a domain heavily influenced by the state, with IT champions serving as proxies. Once this dimension is fully recognised as a potential threat to stability, it is very likely that algorithmic governance will follow the same path as data governance, i.e., politicisation, followed by securitisation. With respect to the regulatory framework, the scarce insights available in early 2025 regarding the ongoing discussions of the AI draft law suggested that security matters were likely to prevail over market interests.

6. Conclusion

Russia aspires to reach a leading position among global AI powers. However, the country’s ambition is constrained by shortages of available resources, including compute power, capital and talent. A distinct feature of Russia’s AI governance model is the strong influence of law enforcement bodies, namely the FSB and Roskomnadzor, in AI governance. This influence, which goes beyond these agencies’ basic responsibilities, serves as an illustration of the ongoing securitisation of numerous aspects of the country’s digital-economy governance. Digital technologies, and AI in particular, are viewed by the Russian leadership as sources of risk. The response is the government’s politicisation and securitisation of AI-related matters and its supervision of non-state market actors’ activities. We expect that the coming years will see more restrictions imposed by the government, justified by the state as a means to protect Russia’s AI sovereignty and broader digital sovereignty. The likely impact of the restrictions will be further marginalisation of non-state actors in the Russian AI sector, thus consolidating state control over digital innovation and narrowing the space for open technological development.

Acknowledgement

The authors acknowledge the support of the CyberBRICS project, Center for Technology and Society (CTS), Fundação Getulio Vargas (FGV) Law School, Rio de Janeiro.

Funding declaration

The research was conducted as part of the CyberBRICS project’s fellowship programme.

Data availability

The data supporting the results of this study is available upon written request to the first-listed author at ignatov-aa@ranepa.ru.

AI declaration

The authors did not use any generative AI tools for the research covered in this article or in the preparation of this article.

Competing interests declaration

The authors have no competing interests to declare.

Author contributions

AI: conceptualisation; methodology; validation; writing – the initial draft (including substantive translation).
DK: conceptualisation; data collection; validation; writing – revisions.

References

AI Alliance Russia. (n.d.). *AI Alliance Russia*. Retrieved April 8, 2025, from https://a-ai.ru/?lang=en

Analytical Center. (2024). *2024 Analytical report on publication activity of Russian specialists at A* level conferences in the field of artificial intelligence for the period from 2019 to 2023*.

Baikal Electronics. (n.d.). *About the company*. Retrieved November 20, 2024, from https://www.baikalelectronics.ru/about/

Belli, L. (2023). *Exploring the key AI sovereignty enablers (KASE) of Brazil, towards an AI sovereignty stack*. https://cyberbrics.info/wp-content/uploads/2023/08/AI-sovereignty-updated-CLEAN.pdf

Bendett, S. (2024). *The role of AI in Russia's confrontation with the West*. Center for New American Security (CNAS). https://www.cnas.org/publications/reports/the-role-of-ai-in-russias-confrontation-with-the-west

Bianchi, A., & Greipl, A. (2022, November 17). *States' prevention of terrorism and the rule of law: Challenging the "magic" of artificial intelligence (AI).* International Centre for Counter-Terrorism (ICCT). https://icct.nl/publication/states-prevention-terrorism-and-rule-law-challenging-magic-artificial-intelligence-ai

BRICS. (2024). XVI BRICS Summit Kazan Declaration. http://static.kremlin.ru/media/events/files/en/RosOySvLzGaJtmx2wYFv0lN4NSPZploG.pdf

Buzan, B., Wæver, O., & de Wilde, J. (1998). *Security: A new framework for analysis*. Lynne Rienner.

Charisle, M. (2021, September 11). How 9/11 radically expanded the power of the U.S. government. *TIME*. https://time.com/6096903/september-11-legal-history/

Charrett, C. (2009). *A critical application of securitization theory: Overcoming the normative dilemma of writing security*. Working Paper No. 2009/7. International Catalan Institute for Peace. http://dx.doi.org/10.2139/ssrn.1884149

ComNews. (2023, May 25). *57% of Russian companies have switched to domestic software*. https://www.comnews.ru/projects/import-substitution/news/226354/57-rossiyskikh-kompaniy-pereshli-otechestvennoe

ComNews. (2024, March 22). *With the growing demand for AI specialists, only a few Russian universities are graduating qualified personnel*. https://www.comnews.ru/content/232211/2024-03-22/2024-w12/1007/pri-rastuschey-potrebnosti-specialistakh-ii-tolko-neskolko-rossiyskikh-vuzov-vypuskayut-profprigodnye-kadry

Consultant Plus. (2025). *Federal Law "On the Security of the Critical Information Infrastructure of the Russian Federation" dated 26.07.2017 N 187-FZ (latest revision)*. https://www.consultant.ru/document/cons_doc_LAW_220885

CyberBRICS. (n.d.). *About us*. https://cyberbrics.info/about-us

Dorofeeva, E., & Arialina, M. (2025, April 8). The State Duma has created a working group to regulate artificial intelligence. *Vedomosti.* https://www.vedomosti.ru/society/articles/2025/04/08/1103157-v-gosdume-poyavilas-po-regulirovaniyu-iskusstvennogo-intellekta

D-Russia. (2024, October 1). *What digital expenditures are included in Russia's draft budget for 2025–27?* https://d-russia.ru/kakie-cifrovye-rashody-zalozheny-v-proekte-bjudzheta-rossii-na-2025-27-gg.html

Epifanova, A., & Dietrich, P. (2022). *Russia's quest for digital sovereignty*. German Council on Foreign Relations (DGAP). https://dgap.org/sites/default/files/article_pdfs/DGAP-Analyse-2022-01-EN_0.pdf

European Commission (EC). (2023, October 30). *Commission welcomes G7 leaders' agreement on Guiding Principles and a Code of Conduct on Artificial Intelligence* [Press release]. https://digital-strategy.ec.europa.eu/en/news/commission-welcomes-g7-leaders-agreement-guiding-principles-and-code-conduct-artificial

Gerden, E. (2024, August 29). Russia set to cut research spending by 25%. *Science*. https://www.science.org/content/article/russia-set-cut-research-spending-25

Government of Russia. (n.d.). *AI development*.

Government of Russia. (2013). Strategy for the Development of the Electric Power Industry of the Russian Federation. http://static.government.ru/media/acts/files/0001201304080048.pdf

Government of Russia. (2019). Energy Security Doctrine of the Russian Federation. https://minenergo.gov.ru/ministry/energy-security-doctrine

Government of Russia. (2020). Strategy for Development of the Electronic Industry of the Russian Federation for the Period until 2030. http://static.government.ru/media/files/1QkfNDghANiBUNBbXaFBM69Jxd48ePeY.pdf

Government of Russia. (2023). Strategy of Development of the Telecommunications Industry of the Russian Federation until 2035. http://static.government.ru/media/files/Pc7fHuejbNvqv17b0RJNv0RIqTo20lUV.pdf

Inclient. (2022). *Statistics on the outflow of IT specialists from Russia in 2022*. Retrieved November 20, 2024, from https://inclient.ru/outflow-it-specialists

Interfax. (2025, September 27). *Russia to spend over 5 bln rubles to support development of AI technology in 2024 – Mishustin*. https://interfax.com/newsroom/top-stories/94917

International Energy Agency (IEA). (n.d.). Energy system of Russia. Retrieved November 20, 2024, from https://www.iea.org/countries/russia

International Telecommunication Union (ITU). (n.d.-a). Russian Federation. Retrieved November 20, 2024, from https://datahub.itu.int/data/?e=RUS

ITU. (n.d.-b). Russian Federation fixed-broadband internet basket. Retrieved April 8, 2025, from https://datahub.itu.int/data/?e=RUS&c=701&i=34616

Kholupova, K. (2024, March 26). Baikal processor localises one of the production stages. *Vedomosti.* https://www.vedomosti.ru/technology/articles/2024/03/26/1027924-razrabotchik-protsessorov-baikal-lokalizuet-odin-iz-etapov-proizvodstva

Kolomychenko, M. (2024). The impact and limits of sanctions on Russia's telecoms industry. *DGAP.* Retrieved June 17, 2025, from https://dgap.org/en/research/publications/impact-and-limits-sanctions-russias-telecoms-industry

Kommersant. (2023, April 14). Woe from wit: The use of AI to be regulated by law. https://www.kommersant.ru/doc/5928661

Kommersant. (2024, April 9). Law abiding intelligence. https://www.kommersant.ru/doc/6621034

Kurasheva, A. (2023, July 28). Foreign companies still hold 30% of the Russian cybersecurity market. *Vedomosti*. https://www.vedomosti.ru/technology/articles/2023/07/28/987508-zarubezhnie-zanimayut-30

Miller, A. (2024, September 30). Russia to allocate a record 13.5 trillion rubles for war in 2025. *DW*. https://www.dw.com/ru/rossia-vydelit-v-2025-godu-na-vojnu-rekordnye-135-trln-rublej/a-70368182

Nadibaidze, A. (2022). *Russian perceptions of military AI, automation, and autonomy.* Foreign Policy Research Institute (FPRI). https://www.fpri.org/wp-content/uploads/2022/01/012622-russia-ai-.pdf

National Archives. (n.d.). *Global war on terror*. Retrieved November 20, 2024, from https://www.georgewbushlibrary.gov/research/topic-guides/global-war-terror

Norem, J. (2024, April 22). *Russia is working on a 128-core supercomputing platform: Report*. ExtremeTech. https://www.extremetech.com/computing/russia-is-working-on-a-128-core-supercomputing-platform-report

Organisation for Economic Co-operation and Development (OECD). (2023). *OECD science, technology and innovation outlook 2023.* https://www.oecd.org/en/publications/oecd-science-technology-and-innovation-outlook-2023_0b55736e-en.html

OECD. (2024). *New perspectives on measuring cybersecurity.* https://www.oecd.org/en/publications/new-perspectives-on-measuring-cybersecurity_b1e31997-en.html

Petrella, S., Miller, C., & Cooper, B. (2021). Russia's artificial intelligence strategy: The role of state-owned firms. *Orbis*, *65*(1), 75–100. https://doi.org/10.1016/j.orbis.2020.11.004

Pfaff, C. A. (Ed.) (2025). *The weaponization of AI: The next stage of terrorism and warfare.* Centre for Excellence Defence Against Terrorism (COE-DAT). https://www.tmmm.tsk.tr/publication/researches/21-TheWeaponizationofAI-TheNextStageofTerrorismandWarfare.pdf

President of Russia. (2016). Decree of the President of the Russian Federation No. 646 dated 5 December 2016 on the Approval of the Information Security Doctrine of the Russian Federation. http://www.kremlin.ru/acts/bank/41460

President of Russia. (2019a). Decree of the President of the Russian Federation No. 216 dated 13 May 2019 on the Approval of the Energy Security Doctrine of the Russian Federation. http://static.kremlin.ru/media/events/files/ru/rsskwUHzl25X6IijBy20Doj88faOQLN4.pdf

President of Russia. (2019b). Decree of the President of the Russian Federation No. 490 dated 10 October 2019 on the Development of Artificial Intelligence in the Russian Federation. http://www.kremlin.ru/acts/bank/44731

President of Russia. (2021). National Security Strategy of the Russian Federation. http://www.scrf.gov.ru/media/files/file/l4wGRPQJvETSkUTYmhepzRochb1j1jqh.pdf

President of Russia. (2024). Decree of the President of the Russian Federation No. 124 dated 15 February 2024 on Amendments to the Decree of the President of the Russian Federation No. 490 dated 10 October 2019 on the Development of Artificial Intelligence in the Russian Federation. http://publication.pravo.gov.ru/document/0001202402150063

RBC. (2022, May 28). *Experts assess the impact of IT specialists towards the end of the last half year*. https://www.rbc.ru/technology_and_media/28/05/2022/628fa85d9a7947dabe3b3e30

Realnoye Vremya. (2024, August 4). Over the past two years, the drain of IT specialists in Russia has decreased. https://realnoevremya.ru/news/314247-v-rossii-uluchshilas-obstanovka-otnositelno-otezda-it-specialistov

Ross, S. (2024, August 28). *The 5 biggest Russian banks*. Investopedia. Retrieved November 20, 2024, from https://www.investopedia.com/articles/investing/082015/6-biggest-russian-banks.asp

Russian Federation. (2006). Federal Law on Personal Data (as amended as of 8 August 2024). https://docs.cntd.ru/document/901990046

Rutland, P. (2008). Russia as an energy superpower. *New Political Economy*, *13*(2), 203–210. https://prutland.faculty.wesleyan.edu/files/2015/08/Russia-as-an-energy-superpower.pdf

Sidelnikova, D. (2024, February 27). *Three more Yandex services have been included in the register of user tracking. What does this mean and how can you protect yourself?* Takie Dela. https://takiedela.ru/notes/utechka-dannykh/

SimilarWeb. (2021). *Top websites ranking*. Retrieved November 20, 2024, from https://www.similarweb.com/top-websites/russian-federation/computers-electronics-and-technology/social-networks-and-online-communities/

Starchak, M. (2024, August 16). *Russian defense plan kicks off separate AI development push*. DefenseNews. https://www.defensenews.com/global/europe/2024/08/16/russian-defense-plan-kicks-off-separate-ai-development-push/

Statista. (2024). *Leading countries by gross research and development (R&D) expenditure worldwide in 2022*. Retrieved November 20, 2024, from https://www.statista.com/statistics/732247/worldwide-research-and-development-gross-expenditure-top-countries/

Stix, C. (2022) Foundations for the future: Institution building for the purpose of artificial intelligence governance. *AI Ethics*, 2, 463–476. https://doi.org/10.1007/s43681-021-00093-w

TAdviser. (2025). *390 billion rubles of investment and 2,000 AI developers: The Ministry of Economic Development summed up the results of the federal project "Artificial Intelligence"*. Retrieved April 8, 2025, from https://shorturl.at/z1NBm

Tallberg, J., Erman, E., Furendal, M., Geith, J., Klamberg, M., & Lundgren, M. (2023). The global governance of artificial intelligence: Next steps for empirical and normative research. *International Studies Review*, *25*(3). https://academic.oup.com/isr/article/25/3/viad040/7259354

TASS. (2024, July 8). *AI developers will insure against risks for potential harm to life from technology*. https://tass.ru/obschestvo/21307449

Telecom Daily. (2024). *Programmer attrition has ceased to be a problem for Russian IT companies*. Retrieved November 20, 2024, from https://shorturl.at/utQNw

Tencent. (2024). *Corporate overview*. Retrieved November 20, 2024, from https://static.www.tencent.com/uploads/2024/08/14/20913f7ba15aacb47b51d502f1cc1da4.pdf

TOP500. (n.d.). *The list*. Retrieved April 8, 2025, from https://top500.org

Tyunyaeva, M. (2023, December 14). What AI threats did Vladimir Putin warn about? *Vedomosti*. https://www.vedomosti.ru/technology/articles/2023/12/14/1011153-ugrozah-ii-putin

UN Interregional Crime and Justice Research Institute (UNICRI) & UN Office of Counter-Terrorism (UNOCT). (2021a). *Algorithms and terrorism: The malicious use of artificial intelligence for terrorist purposes.* https://unicri.org/News/Algorithms-Terrorism-UNICRI-UNOCCT

UNICRI & UNOCT. (2021b). *Countering terrorism online with artificial intelligence.* https://www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/countering-terrorism-online-with-ai-uncct-unicri-report-web.pdf

Ünver, H. A. (2024). *Artificial intelligence (AI) and human rights: Using AI as a weapon of repression and its impact on human rights*. European Parliament. https://www.europarl.europa.eu/RegData/etudes/IDAN/2024/754450/EXPO_IDA(2024)754450_EN.pdf

Urusov, P. (2023, July 25). *Vital microchip sanctions will hit Russian computing power hard*. Carnegie Politika. https://carnegieendowment.org/russia-eurasia/politika/2023/07/vital-microchip-sanctions-will-hit-russian-computing-power-hard?lang=en

US Department of the Treasury. (2024). *Managing artificial intelligence-specific cybersecurity risks in the financial services sector*. https://home.treasury.gov/system/files/136/Managing-Artificial-Intelligence-Specific-Cybersecurity-Risks-In-The-Financial-Services-Sector.pdf

Zinovieva, E. (2024, October 31). *What's wrong with the Global Digital Compact?* Russian Foreign Affairs Council. https://russiancouncil.ru/analytics-and-comments/analytics/chto-ne-tak-s-globalnym-tsifrovym-dogovorom

Annexure: KASE mapping tool

General questions	KASE dimensions	KASE dimension-specific questions
1. Is there a strategy? If so, which public entity (e.g., ministry) defines and implements it? 2. Is there any regulation? If so, which public entity (e.g., regulatory authority) oversees regulating? 3. Is there a funding mechanism stimulating R&D and innovation? If so, which entities orchestrate the funding mechanism? Which mechanisms exist to incentivise innovation? 4. Which are the key private-sector or non-governmental stakeholders (e.g. national champion(s), dominant actors, or non-governmental bodies)? Which are their main interests (provide examples)? Is there any foreign private entity with particular relevance in the sector? 5. Is there a strategy of international cooperation or expansion of national sector?	Data (personal, non-personal, critical, confidential, etc.)	1. How is the country’s census infrastructure in terms of capacity and diversity? 2. Are high-quality, diverse data sets easily available? 3. Are there AI-ready datasets? 4. Is there a strategy for data commons?
	Algorithms (including models, etc.)	1. Is there any policy for open-source software development? 2. Does the public administration use proprietary software developed domestically or by foreign players, or open software? 3. What are the AI procurement rules, if any? 4. Is there any public–private partnership mechanism to incentivise development and deployment of algorithms?
	Computing capacity value chain (including servers, storage resources,	1. Which kind of public computing capacity is there? 2. Are there public supercomputers? 3. What is the largest computing cluster? 4. Are they available for private sector use? 5. Are there any components manufactured in the country? 6. What are the most notable investments in the various elements of the computing capacity value chain? 7. Is there a strategy for capacity building for cutting-edge work in the computation supply chain?
	Connectivity infrastructure (including submarine cable, terrestrial, and satellite infrastructure)	1. How meaningful is connectivity (affordability, zero-rating in place, proportion of access by type of device, by gender, by economic segment, etc.)?
	Electricity infrastructure (including renewables and batteries, etc.)	1. Is there a stable, reliable, and affordable electrical power supply throughout the country? 2. Are there relevant discrepancies within the country in terms of energy supply and infrastructure? 3. What is the proportion of electricity produced via renewable sources? 4. Is there any regulation for the use of electricity for specific types of technology?
	Education, talent promotion, and retention	1. What is the digital literacy rate? 2. How many computer scientists and engineers graduate per year? 3. Are there specific degrees (Bachelor’s and Master’s) specifically targeting AI from public universities? 4. Is there any public initiative to foster AI studies? 5. Are there specific courses or certifications for AI for public servants? 6. Is it within the public administration? 7. What are the immigration patterns of AI scientists? 8. Is the country importing or exporting AI talent?
	Cybersecurity	1. Are there specific protection policies for AI-related infrastructure (such as supercomputers)? 2. Is there a public body fostering coordination among agencies and public administration with competences on cybersecurity?
	Digital public infrastructure (DPI) (DPI for AI, and AI for DPI)	1. Is there a definition of DPI? 2. Are there AI components within major DPIs (digital ID, payment methods, data sharing platforms)? 3. Is AI used in other public software platforms that could be considered DPIs? 4. Are there specific AI software and hardware labelled as DPI? 5. Has the government developed or promoted the development of any generative LLM?

Source: CyberBRICS (n.d.)

Understanding interrelationships between AI and digital public infrastructure (DPI) in India and Brazil

Amrita Sengupta
Fellow, CyberBRICS project, Center for Technology and Society (CTS), Fundação Getulio Vargas (FGV) Law School, Rio de Janeiro
https://orcid.org/0009-0003-8612-090X

Alexandre Costa Barbosa
Associate Researcher, Weizenbaum Institut, Berlin; and Fellow, CyberBRICS project, Center for Technology and Society (CTS), Fundação Getulio Vargas (FGV) Law School, Rio de Janeiro
https://orcid.org/0000-0002-9893-9678

Mila T. Samdub
Fellow, CyberBRICS project, Center for Technology and Society (CTS), Fundação Getulio Vargas (FGV) Law School, Rio de Janeiro
https://orcid.org/0000-0001-8826-4550

Abstract

The conversation around artificial intelligence (AI) has gained tremendous momentum, especially since the onset of widespread use of generative AI. While certain countries dominate AI's discourse, development and deployment, others are rushing to see how AI strategies can be built for their economies, in an effort to avoid missing out on potential benefits. To that end, many countries, including the BRICS nations, are seeking to develop their own competencies for AI development and are working towards greater AI sovereignty. In this context, the conversation around digital public infrastructure (DPI) is critical, given that both AI and DPI have the potential, when implemented well, to mutually advance the public good. In this article, we discuss the interrelationships between AI and DPI, with a particular emphasis on how this interrelationship is being operationalised in India and Brazil. We suggest two broad frames—AI for DPI, and DPI for AI—and examine the frameworks that integrate AI and DPI. We also point to some emergent risks and future considerations that countries and their policymakers need to take into account when considering interactions between AI and DPI.

Keywords

artificial intelligence (AI), digital public infrastructure (DPI), digital public goods, AI sovereignty, public values, AI governance

DOI: https://doi.org/10.23962/ajic.i35.20800

Recommended citation

Sengupta A., Barbosa, A.C., & Samdub, M.T. (2025). Understanding interrelationships between AI and digital public infrastructure (DPI) in India and Brazil. *The African Journal of Information and Communication (AJIC), 35*, 1–11. https://doi.org/10.23962/ajic.i35.20800

1. Introduction

In September 2024, the UN Summit of the Future adopted the Global Digital Compact, which suggests that digital public infrastructure (DPI) is key to inclusive growth and pushes for greater investments to this end (UN, 2024b). DPI is a subset of a more general category, public infrastructure, which comprises fundamental services, public goods, and long-term systems including, but not limited to, railways, roads, telecommunications, and public transport (Bowker et al., 2010). Such infrastructure is frequently cited in political and economic discourses as essential for comprehensive, large-scale solutions crucial to a population's quality of life (Edwards et al., 2009). Within the digital subset of public infrastructure—the DPI—a core current dimension may be the role played by artificial intelligence (AI). Accordingly, the focus of this article is on interrelationships between AI and DPI, with a focus on how these interrelationships are linked to the AI sovereignty aspirations of two members of the BRICS bloc of countries, India and Brazil. These two nations are leaders, along with fellow BRICS member South Africa, in foregrounding DPI within the BRICS bloc and, more broadly, within the G20 bloc of the world's largest economies.

DPI

The concept of DPI, as a specific class of digital infrastructure, is emergent and contested, particularly with respect to the notion of "public" (Mazzucato et al., 2024; Samdub, 2025a; Samdub & Rajendra-Nicolucci, 2024). For the purposes of this article, we adopt an understanding of DPI as open, interoperable software development at scale on a platform architecture that has several hardware dependencies, often promoted by state mandate. This combines a normative definition of DPI, as adopted in international fora, with a critical analysis of the forms actually being taken by DPI around the world. The DPI agenda achieved a measure of global consolidation during India's G20 presidency in 2023, with the G20 New Delhi Declaration framing DPI as "an evolving concept" that refers to

> a set of shared digital systems, built and leveraged by both the public and private sectors, based on secure and resilient infrastructure, [which] can be built on open standards and specifications, as well as open-source software [that] can enable the delivery of services at societal scale. (G20, 2023)

Building upon that consensus, while also localising the concept, the Brazilian Government, chair of the G20 in 2024, defined DPI as

> structuring solutions that adopt networked technology standards for the public interest. They are designed to be used by various entities in the public and private sectors, following the principles of universality and interoperability. (Federal Government of Brazil, 2024c)

While DPI can be defined broadly, the dominant version of DPI is associated with systems for digital identification, payments, and data exchange (Samdub, 2025a) built on a platform architecture that can be accessed by a range of ecosystem actors. Such DPI systems are active in India and Brazil. Systems with wide-scale adoption, including India's Aadhaar biometric identification project and Brazil's Pix digital payments system, have attained scale due to state mandates. While other BRICS countries have built advanced systems for ID, payments, and data exchange, they have generally not used the term DPI to refer to them.

While also considering software as infrastructure, the field of information infrastructure studies emphasises material aspects as key to defining infrastructure (Star, 1999). Other dimensions are also central to infrastructure, such as transparency, embeddedness, and modularity, with infrastructure providing a foundation for multidimensional effects (Frischmann, 2012). Digital infrastructure in general includes submarine and terrestrial cables, optical fibre, towers, satellites, and the internet, as well as technical standards and, as in the case of the domain name system (DNS), organisations that maintain the technical standards. Such infrastructure enables data flow, nationally and internationally (Bowker et al., 2010). These hardware and technical infrastructures are critical dependencies for the functioning of software DPIs.

AI

Agreeing upon and defining AI precisely has involved considerable confusion and numerous challenges, and its definition has also seen significant evolution over the years. As early as 1950, Alan Turing defined AI as "the science and engineering of making intelligent machines, especially intelligent computer programs" (as cited by Pellicelli, 2023, p.140). More recently, the Organisation for Economic Co-operation and Development (OECD) 2024 AI definition specifies as follows:

> [a] machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment. (OECD, 2024, p. 4)

UNESCO's (2021) definition of AI as "information-processing technologies that integrate models and algorithms" provides potential overlaps with DPI. Critical and structural components of AI systems, such as open AI models, may be considered DPI, as recognised by the aforementioned UN Global Digital Compact (UN, 2024b). Moreover, AI technologies may be applied to DPI, for purposes such as detecting fraud in cash transfer programmes or enabling identity verification. Although researchers have noted the links and potential synergies between DPI and AI, there have not, to our knowledge, been systematic critical accounts of the interrelationships between DPI and AI (Nagar & Eaves, 2024).

AI sovereignty

AI sovereignty, as defined by Belli (2023), refers to a nation's capacity to "understand, muster and develop AI systems, while retaining control, agency, and, ultimately, self-determination over such systems" (2023, p. 29). Contemporary AI development depends on access to and control over resources that are highly concentrated in the hands of US-based tech firms, namely Microsoft, Google, Meta, and Amazon (AI Now Institute, 2023). These include, among others, high-quality datasets, cloud-computing infrastructure, high-performance semiconductors, and large-scale models. DPI, when implemented well, may play a critical role in creating greater autonomy and control over technological systems by nations (via the actions of state and/or non-state actors) while promoting the public good. It is through this lens that we have prepared this article, trying to understand DPI's potential for furthering AI sovereignty while proposing a structure for making sense of the relationships between DPI and AI.

It is important to note that the notion of AI sovereignty—or sovereign AI—is also increasingly being cited by Big Tech companies seeking to provide AI as an offering to governments. NVIDIA, a large tech corporation with the largest GPU market share (Yahoo Finance, 2025), defined sovereign AI as "nations' capabilities to produce artificial intelligence using its own infrastructure, data, workforce and business networks" (Lee, 2024). In NVIDIA's Q2 2025 earnings press release NVIDIA's CEO mentioned sovereign AI as a promising multibillion-dollar vertical future market (Keegan, 2024). While not using the term sovereignty, OpenAI recently announced a new initiative called "OpenAI for Countries", which pledges to support national AI development across various points of the AI value chain (including data centres and customised ChatGPT, among others) (OpenAI, 2025). While Belli's framework of AI sovereignty looks at developing autonomy at different layers of the AI stack, the use of sovereign AI as a term of art by global Big Tech companies, whose very dominance is partly what AI sovereignty seeks to challenge, has the potential to act as a counterforce.

In this article, we categorise two possible linkages—DPI for AI and AI for DPI—and present examples from India and Brazil to explain these phenomena. We then propose an analytical framework to explore these relationships further and also to situate them in the context of AI sovereignty. We then conclude by offering some insights into the potential risks of DPI-for-AI and AI-for-DPI approaches, and some questions that could be explored when developing these solutions.

2. AI for DPI

AI for DPI refers to the ways in which AI is used to enable or extend the functioning of DPI. Different kinds of AI are being and can be applied at various sites along the chain of DPI implementation. DPI is often described as the "rails" of digital society. We understand AI for DPI as a situation in which AI technologies are used in the construction of these rails. We now turn to exploration of a range of AI integrations into DPI, from the most rudimentary to the most sophisticated.

If we define AI in its broadest sense as automated decision-making (ADM) systems, it is essential to the functioning of DPI. For example, AI enables DPI for identification. One of the key operations in DPI for identification is the process of de-duplication, i.e., the determination and deletion of repeat citizen records across government databases. The determination of these repeat records takes place using ADM techniques, matching possible duplicates with each other and flagging them for deletion. Many implementations of identification DPI, such as the Indian Aadhaar system, depend on biometric authentication to verify identity. The matching of a user's fingerprint, retina, or face to a record in a database is a probabilistic process that returns a confidence percentage rather than a definitive yes/no answer (Ranganathan, 2020). As such, these algorithmic processes are rudimentary forms of AI that are ubiquitous in DPI.

AI systems are also used to automate administrative tasks in DPI. AI can be applied to detect fraud in financial transactions, supporting anti-money-laundering schemes and easier know-your-customer (KYC) procedures. It can also be used to facilitate the eligibility of beneficiaries within a given social information management system. For example, Brazil's CadUnico is a database and beneficiary identification tool that differentiates the needs of target populations according to the characteristics of each family. The entire procedure takes place through a single gateway and with a single application, storage, and governance scheme. In 2023, CadUnico was integrated with Brazil's National Social Information Registry (CNIS), a pre-existing system that supports the granting of social security benefits. AI is used to identify inconsistent and updated registries within CadUnico and the beneficiaries of social protection programmes (Grossman, 2025). We consider such relatively rudimentary uses of ADM as AI for DPI because they enable the automated processing of information at a large scale and volume.

As part of their modular, extensible, and interoperable architecture, several DPIs offer application programming interface (API) access, enabling government agencies and private companies to build on top of their "rails". Such AI systems, built on top of DPI, may be used for citizen–state interactions with the goal of improving citizens' access to public services. These forms of AI for DPI are similar to the DPI-for-AI examples discussed in the next section, in that they "plug in" to DPI. However, we analytically distinguish them from DPI for AI based on the following distinction: the goal of AI for DPI is to enhance access to DPI and deliver public value, while the goal of DPI for AI is to provide support for AI development that caters to domestic public needs or creates an enabling environment for domestic AI development.

The widespread promotion of chatbots in public service delivery (Garcia, 2024) is a key example of AI for DPI. These chatbots interact with citizens to impart knowledge about government services, with the promise of improving access. For example, India's Jugalbandi chatbot, developed by AIforBharat and Microsoft, makes information about government schemes available in vernacular Indian languages. The recently launched Hello UPI! conversational payments technology in India layers an AI conversational chatbot on top of the Unified Payments Interface (UPI) payments system (Ministry of Finance, 2023). While previously UPI needed to be accessed using an app interface, Hello UPI! uses API access to UPI to allow users to make payments using voice commands, with the AI providing automatic speech recognition, language translation, intent verification, and voice output. The goal of this feature is to increase financial inclusion by easing access to the financial system, especially for people who are not literate.

3. DPI for AI

DPI for AI comprises ways in which DPI is leveraged to advance a country's AI-related interests. Examples range from the creation, collection, and collation of large datasets for AI training to the Open Cloud Compute (OCC) system proposed by India's People+ai. DPI for AI can aid in creating access to large datasets, which is an ongoing priority for Global Southern countries seeking to build their AI sovereignty. However, such DPI-for-AI uses do not preclude the possibility of coded harms of bias facilitated through algorithmic decision-making, and they raise concerns around, inter alia, privacy, data security, compliance with local regulations of data storage, and data minimisation.

We now turn to consideration of two DPI-for-AI examples in India—the aforementioned OCC system, and the BhashaDaan function of the Bhashini language translation platform—as well as Brazilian policy directions with relevance to DPI for AI. It is important to note that the use of DPI to improve or enable AI is a nascent idea at this stage, with few rollouts and limited evidence of the public value that it generates or other success parameters. However, we anticipate that the ongoing convergence of DPI and AI in global forums necessitates this current discussion.

Market concentration has been a major strand of study and investigation across industries in both economics and law. The Sherman Act of 1890 in the US was one of the first major acts by Congress with the aim of combating anti-competitive practices, reducing monopolistic power, and preserving economic competition (Micelli, 2009). Even in the AI industry, market concentration in the hands of a few tech giants has been a cause for major concern. Market concentration, while being impacted by various factors traditionally, particularly suffers from the role of network effects in social media platforms, which is now also seen with AI companies. While network effects have various definitions in economic theory, Church, Gandal, and Krause (2002) emphasised, building on the contribution of others, that a "network effect exists when consumption benefits depend positively on the total number of consumers who purchase compatible products" (2002, p. 1). This framing of network effects applies well to generative AI companies, where the greater use of a generative AI product leads to more data collection and learning, which may lead to better model performance.

At the heart of the AI sovereignty conversation is, then, the issue of taking back some of the control that US Big Tech firms enjoy. Even in Global Northern contexts, competition authorities (e.g., in the US and the UK) have been looking at the close links between generative AI firms and their Big Tech investors. In April 2024, the Competition and Markets Authority in the UK raised concerns about an "interconnected web" of over 90 partnerships and strategic investments established by Google, Apple, Microsoft, Meta, Amazon, and NVIDIA in the market for generative AI foundation models (Kersley, 2024).

Fundamental to AI is the need for computing infrastructure. The global cloud compute market is estimated at USD500 billion annually and is expected to have a value of USD1.5 trillion by 2030 (Yahoo Finance, 2023). One of India's responses to the issue of the current bundled model of mega data centres, and diminished bargaining power for end-users when dealing with large cloud service providers, is the OCC initiative (People+ai, 2024). OCC is slated to be a network of interoperable, micro data centres that are built on common standards, which facilitates India in building its requisite computing infrastructure. The team at People+ai and EkStep Foundation, the organisations facilitating the creation of this network, suggests that through OCC, a digital infrastructure approach to AI is being taken. OCC has been framed as a DPI for compute power and is also seen as an effort to enable "faster processing, lower latency,[1] and stronger data sovereignty" (India Times, 2024). It aims to create an open network of providers, governed through protocols. The promise of OCC is presented as the ability for small businesses to discover various kinds of compute service offerings, and they have the option to select services on the basis of their requirements. As of May 2024, the OCC project had 24 partners, including Oracle, Dell, Tata, and E2E Networks (Mohanty, 2024).

[1] Latency refers to the time that it takes for data to travel between the user and a server.

Bhashini is an AI-driven language translation system that aims to create accessibility to public services in different Indian languages (Fidel Softech, 2024). In addition, it hopes to create access to open-source data and efficient translation tools. The Bhashini platform is slated to be listed as a digital public good, with the aim of contributing to “linguistic accessibility and technological empowerment on a national scale” (Digital India Bhashini Division, 2024). Under Bhashini sits the project of BhashaDaan, which is an initiative to crowdsource language inputs for diverse Indian languages from citizens so as to build an open repository of data in multiple languages. In line with India’s interest in creating localised datasets and improving its AI capability, the intention here is to create large datasets for Indian languages, which can be used to train AI models for use by different stakeholders. The intention of the creation of these products is listed as being the “betterment of society”, which we consider below (Vikaspedia, n.d.). As suggested, the datasets created from BhashaDaan can be used for training various AI models. In this way the larger Bhashini project can also act as DPI for AI.

Brazil’s Minister of Management and Public Service Innovation, Esther Dweck, recognised the role of DPI for AI during the summit of Digital Public Infrastructure Safeguards convened by the Office of the UN Secretary-General’s Envoy on Technology. The Minister stated that this convergence is key for digital sovereignty, especially regarding developing autonomous capacity for a Brazilian Portuguese-trained AI model (Federal Government of Brazil, 2024a). The Brazilian AI Plan 2024–2028 allocates a total budget of BRL23 billion (equivalent to roughly USD4.5 billion), with approximately 25% directed toward AI infrastructure and development. The final version was published in June 2025 (CGEE, 2025). One-quarter of the total amount of investment is expected to be on infrastructure, with the AI Plan stating that “[w]e aim to establish Brazil as a global reference in sustainable AI infrastructure, with innovative models of energy efficiency and the responsible use of natural resources.”

In addition to developing AI models in Brazilian Portuguese, the Plan also supports National Data Infrastructure, which can be seen as an example of DPI for AI. One of the pillars of the National Data Infrastructure is the consolidation of a “Sovereign Cloud” to store and manage the data generated in the country; another is the expansion of supercomputers in the country. The two major Brazilian Federal IT companies have promoted a “Government Cloud” programme, guided by the Ministry of Management and Public Services Innovation. However, the system relies on the services of mainstream cloud companies’, such as Google, Oracle, Amazon, and Huawei (TI Inside, 2025). It also encompasses the creation of a unified education database for the development of applications and the use of AI in that sector. These educational digital infrastructures are relevant convergent factors within an agenda for digital sovereignty (Barbosa & Gonsales, 2024).

4. Layered integration of AI and DPI

Today, both AI and DPI development take place on a platform architecture. This architecture is characterised by API access points that enable other applications to be built on top of them in a stack (Plantin et al., 2018). Due to this platform architecture, DPI for AI and AI for DPI are integrated with each other across layers and iteratively. That is, DPI can be used as a foundation for AI, which in turn may be used to promote DPI, and vice versa. For example, as described in the previous section, the Indian state-promoted BhashaDaan linguistic database is a DPI that offers access to language data through APIs. This data is used by the Jugalbandi initiative alongside OpenAI’s GPT model to provide a broad AI platform to develop vernacular language chatbots. This Jugalbandi platform, itself described as a stack, is in turn used to build chatbots that promote vernacular-language access to sectoral DPIs in law, healthcare, and government services (Jugalbandi, n.d.).

Where and how DPIs and AI relate to each other in their respective stacks is crucial in determining the outcomes of such systems. The DPI stack consists of foundational DPI for payments, identity and data exchanges, as well as sectoral DPI in health, travel, education, and other domains. The AI stack is composed of data, compute, and applications. This iterative and layered integration means that integrating AI and DPI at more foundational levels in their respective stacks has the potential to multiply impact. For example, the successful use of AI for foundational DPI, such as for identity, can increase the value of all applications built on top of it. Conversely, harms can also be multiplied: for example, someone wrongly identified by an

AI identification system may be excluded from all downstream systems. More targeted AI for DPI, such as the Hello UPI! system described in a previous section, will have fewer knock-on effects, both positive and negative.

As seen in section 3 above, DPI for AI has the potential to occupy the data (Bhashini) and compute (OCC) layers of AI development. If successful, DPI may have the potential to disrupt the hyper-centralised and consolidated power structures of AI dominated by US hyperscalers, enhancing competition and AI sovereignty (AI Now Institute, 2023). If these DPI are transparent and accountable, this could lead to more democratic AI development. We now turn to the existing and potential risks of integrating AI and DPI.

5. Existing and emerging risks of AI and DPI convergence

In both AI and DPI, there is an ongoing conversation about the risks and harms that these technologies may pose. The UN's Universal DPI Safeguards Framework, for example, categorises DPI risks into inclusion, safety, and structural vulnerability (UN, 2024a). AI risks and harms have been the subject of far more debate, and they may include various social, political, and economic harms (Acemoglu, 2021). In this section, we turn our attention to risks at the intersection of AI and DPI.

Within the Indian DPI ecosystem, one of the metrics used to display success has been the number of enrolments to systems like Aadhaar, or the number of transactions via UPI. However, a system designed for public benefit should aim to benefit those at the greatest extremes of marginalisation. This means that it is essential to ask the question as to whom such systems ultimately exclude. Efficiency has been a central issue in the discourse on public administration and the current global discourse on DPI creates an expectation of inclusion, affordability, and access. However, as has been seen in the case of various DPI rollouts in India—denial of services and welfare benefits to those without Aadhaar to compulsory enrolments for new digital health IDs under Ayushman Bharat Digital Mission (ABDM)—there continue to be significant pain points to citizens (Parsheera, 2024).

In several cases, DPI exclusions are linked to incorrect decisions made by AI systems that have, for example, denied people access to the welfare to which they have a constitutional right; AI integration that increases efficiency has the potential to exacerbate harms such as exclusion. For example, the centrality of fingerprinting algorithms in Aadhaar has led to manual labourers whose fingerprints are worn out, and who do not return a positive biometric match, to be denied access to welfare (Frayer & Khan, 2018). As with all AI systems, the use of these systems opens up questions about transparency, accountability, and redress that have not been adequately addressed. Even as these AI systems may increase neutrality and efficiency in public service delivery, they may also lead to "barriers in access to welfare or the exercise of individual rights, and the dispossession of people's claims and entitlements to varying degrees" (Joshi, 2021).

In the context of AI for DPI in Brazil, there has been limited information about the use of automated systems in major digital public infrastructures, such as PIX, the instant payment system led by the Central Bank of Brazil, and GOV.br, the digital government ecosystem that includes the country's legal digital identity scheme. Further analysis should include a thorough examination of the newly approved rights- and risks-based AI regulation by the Brazilian Senate (2023), which includes guidelines on using biometric identification for security purposes. Additionally, the Brazilian Artificial Intelligence Plan, 2024–2028, relies heavily on the national identity card database (Federal Government of Brazil, 2024b). Moreover, private banks have begun integrating AI with PIX to interpret clients' intentions and enable automated transactions, while also leveraging machine-learning to detect data patterns, without an explicit impact assessment, indicative of risky behaviour (Nubank, 2023).

In early 2024, India made commitments to invest upwards of USD1.2 billion (₹10,300 crores) over five years on AI projects, including but not limited to computing infrastructure (*Reuters*, 2024). It is evident that there is an interest in moving towards greater AI sovereignty for the nation, and in building infrastructure to that end. It is important to consider the financing models that are being used to build this infrastructure, and the extent to which the infrastructure will serve the public interest. It is also important to note that, while there have been several public announcements about initiatives at the intersection of AI and DPI, their adoption and long-term use is unclear. For example, there have been no public updates about the aforementioned Jugalbandi since its launch in 2023 and the service's website is no longer accessible (Samdub, 2025b). DPI and AI financing must be based not on one-time costs but on lifetime costs, so as to provide clarity on the full extent of costs and also allow for benchmarking between countries (Eaves & Kedia, 2024).

Also requiring consideration are the negative environmental impacts that AI can have, such as impacts on energy and water resources, particularly given that India and Brazil both already experience extreme weather conditions. Environmental sustainability therefore must, inter alia, be factored into plans to expand compute power. Also important is cognizance of the fact that while DPI can contribute to levelling the playing field to compete with US Big Tech, it must not be allowed to generate domestic monopolies. DPI carries the risk of promoting "alt Big Tech" entities that are no more accountable to the public than foreign Big Tech (Parsheera, 2024). Careful, people-centric design and governance choices are essential to avoid this outcome.

6. Conclusion

This article has outlined an approach for making sense of the interrelationships between two key dimensions of the digital world—AI and DPI—with a focus on examples from two of the leading BRICS countries, India and Brazil. We have also explored how certain DPIs are creating technological systems that may contribute to achievement of AI sovereignty. We have categorised AI—DPI interactions as either AI for DPI or DPI for AI. Our framing has shown that AI and DPI are not independent; rather, DPI and AI are integrated at various levels via APIs, thus forming a layered structure. We have also highlighted existing and potential harms present in integrations of DPI and AI. While the discussion around DPI and AI is still somewhat nascent, it is important to begin to thoroughly investigate current gaps in public service delivery that DPI and AI can help to bridge, to foster a more concerted approach to integrating AI and DPI. Better impact evaluations are also needed, to allow for improved understanding of the successes and challenges of such approaches.

Finally, while our focus on the Indian and Brazilian cases has privileged the nation-state as level of analysis, it is important to note that individuals and communities should be the ultimate beneficiaries of digital technologies. In order to not lose sight of that goal, a layered and iterative integration of AI and DPI is required. Both AI and DPI are currently characterised by a concentration of power in the hands of a few organisations: in AI, power is largely in the hands of US Big Tech; in DPI, power is nominally in the hands of public entities, but in practice it is often held by private-sector actors. It is important to ensure that AI and DPI technologies expand, and not constrain, sovereignty—with sovereignty understood as the power to make choices about one's path—at multiple levels, from the nation to the community to the individual.

Acknowledgement
The authors acknowledge the support of the CyberBRICS project at FGV Law School, Rio de Janeiro.

Funding declaration

No funding was received for the study or the preparation of the article.

AI declaration

The authors did not use any AI tools in the research or in the preparation of this article.

Competing interests declaration

The authors have no competing interests to declare.

Authors' contributions
All three authors contributed equally to this study's conceptualisation, execution and writing.

References

Acemoglu, D. (2021). *Harms of AI*. National Bureau of Economic Research. https://doi.org/10.3386/w29247

AI Now Institute (2023). ChatGPT and more: Large scale AI models entrench Big Tech power. In *2023 landscape: Confronting tech power*. https://ainowinstitute.org/publication/large-scale-ai-models

Barbosa, A. C., & Gonsales, P. (2024). Technological infrastructures for education as a political project toward digital sovereignty. *EmRede Journal*, 11, 1085. https://doi.org/10.53628/emrede.v11i.1085

Belli, L. (2023). Exploring the key AI sovereignty enablers (KASE) of Brazil, to build an AI sovereignty stack. In L. Belli & W. B. Gaspar (Eds.), *The quest for AI sovereignty, transparency and accountability*. FGV Direito Rio. https://doi.org/10.2139/ssrn.4465501

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2010). Toward information infrastructure studies: Ways of knowing in a networked environment. In J. Hunsinger, L. Kliever, & R. Schroeder (Eds.), *International handbook of internet research* (pp. 97–117). Springer. https://doi.org/10.1007/978-1-4020-9789-8

Center for Strategic Studies and Management (CGEE). (2025). *Brazilian Artificial Intelligence Plan (PBIA)*. https://www.cgee.org.br/documents/10195/11009772/CGEE_PBIA.PDF

Church, J., Gandal, N., & Krause, D. (2002). *Indirect network effects and adoption externalities.* Foerder Institute for Economic Research Working Paper No. 02-30. http://dx.doi.org/10.2139/ssrn.369120

Digital India Bhashini Division. (2024). Request for empanelment (RFE) for BHASHINI system integrators for multimodal multilingual solution for Digital India BHASHINI Division. Digital India Corporation. https://bhashini.gov.in/static/media/Draft%20Request%20for%20Empanelment%20for%20Chatbot.2a98b42384b343180659.pdf

Eaves, D., & Kedia, M. (2024). *Exploring the different financing models for digital public infrastructure and why they matter*. Policy Brief No. 2024-6. Asian Development Bank Institute. https://doi.org/10.56506/VYDL5566

Edwards, P. N., Bowker, G. C., Jackson, S. J., & Williams, R. (2009). Introduction: An agenda for infrastructure studies. *Journal of the Association for Information Systems*, *10*(5), 364–374. https://doi.org/10.17705/1jais.00200

Federal Government of Brazil. (2024a). Brasil adere à iniciativa da ONU que promove uso universal da infraestrutura pública digital. *Agência Brasil*. https://agenciagov.ebc.com.br/noticias/202409/brasil-adere-a-iniciativa-da-onu-que-promove-uso-universal-da-infraestrutura-publica-digital

Federal Government of Brazil. (2024b). Plano Brasileiro de Inteligência Artificial (PBIA) 2024–2028. https://www.gov.br/lncc/pt-br/assuntos/noticias/ultimas-noticias-1/plano-brasileiro-de-inteligencia-artificial-pbia-2024-2028

Federal Government of Brazil. (2024c). Brasil Participativo: Consulta para Estratégia Nacional de Governo Digital. https://brasilparticipativo.presidencia.gov.br/processes/ENGD/f/77/

Fidel Softech. (2024). *The role of AI in Bhashini's language processing.* https://www.fidelsoftech.com/news-and-blogs/ai-in-bhashinis-language-processing

Frayer, L., & Khan, F. L. (2018, October 1). India's biometric ID has led to starvation for some poor, advocates say. *NPR*. https://www.npr.org/2018/10/01/652513097/indias-biometric-id-system-has-led-to-starvation-for-some-poor-advocates-say

Frischmann, B. M. (2012). *Infrastructure: The social value of shared resources*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199895656.001.0001

G20. (2023). G20 New Delhi Leaders' Declaration. https://www.mea.gov.in/Images/CPV/G20-New-Delhi-Leaders-Declaration.pdf

Garcia, P. (2024, July 9). Is AI the answer for better government services? *BBC*. https://www.bbc.com/news/articles/cmllxl89jlwo

Grossman, L. O. (2025, March 12). Dataprev: CadÚnico cruza bilhões de dados em minutos e terá biometria e inteligência artificial. *Convergencia Digital*. https://convergenciadigital.com.br/governo/dataprev-cadunico-cruza-bilhoes-de-dados-em-minutos-e-tera-biometria-e-inteligencia-artificial

Joshi, D. (2021). *Unpacking algorithmic harms.* AI Observatory. https://ai-observatory.in/admsharms

Jugalbandi. (n.d.). Live applications. Retrieved December 13, 2024, from https://web.archive.org/web/20240915072317, https://www.jugalbandi.ai/liveapplications

Keegan, J. (2024, October 24). *Why countries are seeking to build "sovereign AI"*. Sherwood News. https://sherwood.news/tech/why-countries-are-seeking-to-build-sovereign-ai

Kersley, A. (2024, April 15). Big tech's cloud oligopoly risks AI market concentration. *Computer Weekly*. https://www.computerweekly.com/feature/Big-techs-cloud-oligopoly-risks-AI-market-concentration

Lee, A. (2024, February 12). What is sovereign AI? *NVIDIA blog*. https://blogs.nvidia.com/blog/what-is-sovereign-ai

Manzoor, A. (2014). A look at efficiency in public administration: Past and future. *SAGE Open*, *4*(4), 1–5. https://doi.org/10.1177/2158244014564936

Mazzucato, M., Eaves, D., & Vasconcellos, B. (2024). *Digital public infrastructure and public value: What is "public" about DPI?* UCL Institute for Innovation and Public Purpose Working Paper Series: IIPP WP 2024-05. https://www.ucl.ac.uk/bartlett/sites/bartlett/files/iipp_wp_2024_05.pdf

Miceli, T. J. (2009). *The economic approach to law*. Stanford University Press.

Ministry of Finance. (2023, December 27). Ministry of Finance year ender 2023: Department of Financial Services. [Press release]. https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1990752

Mohanty, A. (2024, May 17). *Compute for India: A measured approach*. Carnegie India. https://carnegieindia.org/posts/2024/05/compute-for-india-a-measured-approach?lang=en&center=india

Nagar, S., & Eaves, D. (2024). *Interactions between artificial intelligence and digital public infrastructure: Concepts, benefits, and challenges.* arXiv. https://doi.org/10.48550/arXiv.2412.05761

Nubank. (2023, October 18). Nubank lança ferramenta que usa inteligência artificial para personalizar vida financeira dos clientes. https://blog.nubank.com.br/nubank-inteligencia-artificial-testes

OpenAI. (2025, May 7). *Introducing OpenAI for countries*. https://openai.com/global-affairs/openai-for-countries

Organisation for Economic Co-operation and Development (OECD). (2024). Explanatory Memorandum on the Updated OECD Definition of an AI System. https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_3c815e51/623da898-en.pdf

Parsheera, S. (2024, June 10). Digital public infrastructure and the jeopardy of "Alt Big Tech" in India. Center for the Advanced Study of India, University of Pennsylvania. https://casi.sas.upenn.edu/iit/smriti-parsheera-2024

Pellicelli, M. (2023). Managing the supply chain: Technologies for digitalization solutions. In *The digital transformation of supply chain management* (pp. 101–152). Elsevier. https://doi.org/10.1016/B978-0-323-85532-7.00002-5

People+ai (2024). *Creating open, innovative compute markets. The DPI way*. Concept paper. https://docs.google.com/document/d/1ZZd3d8CDu4qcZlrr_b3WW9QGVO-IPl2Ztph_8iQKlkc/edit?tab=t.0#heading=h.atxa4w43u3a0

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, *20*(1), 293–310. https://doi.org/10.1177/1461444816661553

Ranganathan, N. (2020). The economy (and regulatory practice) that biometrics inspires: A study of the Aadhaar project. In A. Kak (Ed.), *Regulating biometrics: Global approaches and urgent questions* (pp. 52–61). AI Now Institute. https://ainowinstitute.org/wp-content/uploads/2023/09/regulatingbiometrics-ranganathan.pdf

Reuters. (2024, March 7). India announces $1.2 bln investment in AI projects. https://www.reuters.com/technology/india-announces-12-bln-investment-ai-projects-2024-03-07

Samdub, M. T. (2025a). *"Digital public infrastructure" at a turning point: From definitions to motivations*. Open Future. https://openfuture.eu/publication/digital-public-infrastructure-at-a-turning-point

Samdub, M. T. (2025b). *India as the "AI Use Case Capital of the World" – Socio-Economic Development as AI Hype.* Tech Policy Press. https://www.techpolicy.press/india-as-the-ai-use-case-capital-of-the-world-socioeconomic-development-as-ai-hype

Samdub, M., & Rajendra-Nicolucci, C. (2024, November 25). *What is digital public infrastructure? Towards more specificity*. Tech Policy Press. https://www.techpolicy.press/what-is-digital-public-infrastructure-towards-more-specificity

Senado Federal. (2023). Projeto de Lei nº 2338, de 2023. https://www25.senado.leg.br/web/atividade/materias/-/materia/157233

Star, S. L. (1999). The ethnography of infrastructure. *American Behavioral Scientist*, *43*(3), 377–391. https://doi.org/10.1177/00027649921955326

The Economic Times. (2024, March 7). People+ai partners with 24 tech organisations to enable open cloud compute infra. https://economictimes.indiatimes.com/tech/technology/peopleai-partners-with-24-tech-organisations-to-enable-open-cloud-compute-infra/articleshow/109924739.cms

TI Inside. (2025, June 10). Management: Serpro and Dataprev launch government cloud services for federal agencies. https://tiinside.com.br/10/06/2025/gestao-serpro-e-dataprev-lancam-servicos-de-nuvem-de-governo-para-orgaos-federais

UN. (2024a). *The Universal Digital Public Infrastructure Safeguards Framework*: *A guide to building safe and inclusive DPI for societies.* https://dpi-safeguards-framework.org/frameworkpdf

UN. (2024b). Global Digital Compact. https://www.un.org/global-digital-compact/sites/default/files/2024-09/Global%20Digital%20Compact%20-%20English_0.pdf

UN Educational, Scientific and Cultural Organisation (UNESCO). (2021). Recommendation on the Ethics of Artificial Intelligence. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence

Vikaspedia. (n.d.). Bhashini. Retrieved December 13, 2024, from https://en.vikaspedia.in/viewcontent/e-governance/digital-india/bhashini

Yahoo Finance. (2023, November 3). Global cloud computing market expected to reach $1,554.94 billion by 2030, driven by adoption of cloud-native applications and advanced technologies. https://finance.yahoo.com/news/global-cloud-computing-market-expected-111300627.html

Yahoo Finance. (2025, June 10). Nvidia Secures 92% GPU Market Share in Q1 2025. https://finance.yahoo.com/news/nvidia-secures-92-gpu-market-150444612.html

The African Journal of Information and Communication (AJIC)

Published by the LINK Centre
University of the Witwatersrand (Wits)
Johannesburg, South Africa
https://www.wits.ac.za/linkcentre

ISSN 2077-7213 (online version)
ISSN 2077-7205 (print version)