Artificial Intelligence in Context: *Applications, Ethics, and Governance*

Editor
Ebubekir M. DENİZ

ÇIZGI KITABEVI | e-Book

Statement of Author Accountability

All chapters in this volume have undergone rigorous peer review and editorial assessment. However, the ultimate responsibility for the content, accuracy, originality, and ethical integrity of each contribution rests solely with the respective author(s).

Originality and Academic Integrity

Authors have affirmed that their submitted work:

- Represents original scholarship conducted with academic rigor
- Properly acknowledges all sources through appropriate citation
- Contains no plagiarized material, whether from published works or AI-generated content presented as original research
- Complies with established standards of research ethics and intellectual honesty

Artificial Intelligence in Context: Applications, Ethics, and Governance

Editor

Ebubekir M. DENİZ

ÇİZGİ
KİTABEVİ

Çizgi Kitabevi Publishing:
Resarch

ISBN: 978-625-396-775-8
e-ISBN: 978-625-396-773-4
Certificate No:52493

LIBRARY INFORMATION CARD
- Cataloging in Publication Data (CIP) –
Editor: DENİZ, Ebubekir M.
Artificial Intelligence in Context:
Applications, Ethics, and Governance

Preparation for Printing
Çizgi Kitabevi Publishing
Tel: 0332 353 62 65 – 66

Printing: Sebat Ofset
Certificate No: 74481

ÇİZGİ KİTABEVİ

Sahibiata Mah.
M. Muzaffer Cad. No:41/1
Meram/**Konya**
(0332) 353 62 65

Konevi Mh.
Larende Cad. No:20/A
Meram/**Konya**
(0332) 353 62 66

Siyavuşpaşa Mh.
Gül Sk. No: 15 B
Bahçelievler/**İstanbul**
(0212) 514 82 93

www.cizgikitabevi.com
/ cizgikitabevi

TABLE OF CONTENTS

PART I — AI SYSTEMS AND APPLIED DOMAINS

Section 1.1 – AI in Health and Medicine

Section 1.2 — AI in Infrastructure, Mobility, and Security

Section 1.3 — AI in Design, Space, and Sustainability

PART II: AI, ETHICS, AND GOVERNANCE

Section 2.1 — Human Values and Philosophical Foundations

Section 2.2 — Fairness, Society, and Cultural Context

Section 2.3 — Governance, Law, and Public Administration

Preface

The emergence of artificial intelligence as a transformative force in contemporary society demands rigorous scholarly attention that transcends disciplinary boundaries. This edited volume brings together cutting-edge research presented at the *9. Uluslararası Öğrenci Bilimler Kongresi*, organized jointly by the Presidency for Turks Abroad and Related Communities (YTB) and the İstanbul Academy of Sciences Foundation (İBAV). The selected papers, now revised and expanded, represent a remarkable synthesis of theoretical depth and practical insight, addressing AI's multifaceted impact across health, infrastructure, design, ethics, and governance.

What distinguishes this collection is its commitment to contextualized inquiry. Rather than treating AI as a universally applicable technological solution, the contributors examine how intelligent systems intersect with specific cultural, legal, and institutional frameworks. From stroke detection in medical imaging to road traffic accident prediction, from UAV autonomy to architectural space evaluation, and from algorithmic fairness metrics to public administration reform—each chapter demonstrates how AI implementation must be grounded in local realities while maintaining dialogue with global standards.

The volume's three-part structure reflects this methodological sophistication. Part I explores AI applications across diverse domains, revealing both the technical achievements and the contextual challenges of deployment. Part II shifts to normative and philosophical foundations, interrogating concepts of freedom, consciousness, and justice in an age of algorithmic mediation. Part III addresses governance frameworks, examining how legal, political, and administrative institutions must evolve to ensure responsible AI development. This architecture mirrors the book's central argument: effective AI governance requires simultaneous attention to technical capability, ethical principles, and institutional design.

We extend our profound gratitude to the Presidency for Turks Abroad and Related Communities (YTB) and the İstanbul Academy of Sciences Foundation (İBAV) for their vision in creating platforms where emerging scholars can engage with pressing global challenges. Their support has enabled rigorous peer review, thoughtful revision, and ultimately, the transformation of conference presentations into enduring scholarly contributions. We also thank all authors for their dedication throughout the editorial process, and our publishing partners for bringing this work to readers worldwide.

As AI systems become increasingly embedded in critical infrastructures and decision-making processes, the need for interdisciplinary, contextually aware scholarship has never been greater. This volume contributes to that essential conversation, offering frameworks that balance innovation with accountability, efficiency with equity, and technical sophistication with human dignity. We hope it serves not only as a resource for researchers and policymakers but also as an invitation to ongoing dialogue about the kind of intelligent future we wish to build together.

Ebubekir M. Deniz
Editor
İstanbul, 2025

PART I

AI SYSTEMS AND APPLIED DOMAINS

Section 1.1
AI in Health and Medicine

***

Convolutional Neural Networks for Accurate Stroke Detection in Computed Tomography Imaging

Hasibullah Mohmand[1] and Veer Chandra Kamati[2]

Abstract

Stroke remains one of the leading causes of death and permanent disability worldwide, and early, accurate diagnosis is critical for effective treatment. Computed tomography (CT) imaging commonly used in emergency settings due to its accessibility and speed, is preferred for stroke detection. However, conventional diagnostic workflows rely heavily on radiologist expertise and can be prone to specific errors under time pressure. In this context, artificial intelligence (AI), particularly convolutional neural networks (CNNs), offers a promising solution to improve diagnostic accuracy and efficiency.

This study presents the development of a lightweight CNN model designed for stroke detection on CT scans, with emphasis on clinical applicability and computational efficiency. The model was trained on the publicly available TEKNOFEST 21 Stroke Dataset, provided by the Republic of Turkey Ministry of Health, which contains 6,500 CT images (2,223 stroke, 4,277 normal). For consistency, all images were resized to 224×224 pixels and normalized to grayscale. Data augmentation techniques such as rotation, reflection, and noise addition were applied to enhance generalization, reduce overfitting, and balance the dataset. The architecture was implemented with four convolutional layers equipped with ReLU activation, Batch

[1] Kocaeli University, Technology Faculty, Information Technology
ORCID: 0009-0003-7090-0557, hasibullah.mohmand13@gmail.com
[2] Kocaeli University, Technology Faculty, Biomedical Engineering
ORCID: 0009-0003-3679-5191, imveerkamati@gmail.com

Normalization, and Max-Pooling, followed by Global Average Pooling2D(GAP) and fully connected layers to balance high accuracy with low computational cost. Training used the Adam optimizer and binary cross-entropy loss; Dropout and learning rate reduction strategies were employed to improve stability.

Class imbalance, common in medical datasets, was addressed via resampling and weighted loss functions. Evaluation metrics included accuracy, precision, sensitivity (recall), F1-score, specificity, and area under the receiver operating characteristic curve (AUC), with a focus on minimizing missed stroke cases that are clinically critical. The model achieved 92.50% accuracy, 92.31% precision, 85.71% sensitivity (recall), 88.89% F1-score, 96.15% specificity, and an AUC above 0.96, demonstrating strong discrimination between stroke and normal images even on low-resolution scans. Inference time was kept below 200 milliseconds per scan, indicating potential for real-time use. Comparative tests against baseline CNN models showed superior performance, particularly in sensitivity and robustness to varying CT image quality. Although small performance fluctuations were observed across datasets with different imaging protocols, the proposed system shows promise for integration into clinical and telemedicine workflows. Its lightweight design, interpretability features, and region-of-interest highlighting make it suitable for deployment in resource-limited settings with constrained radiology expertise.

This work offers a clinically meaningful AI-assisted approach for stroke detection that can support faster and more reliable decision-making while reducing diagnostic workload and variability. Future work will focus on multi-center validation and regulatory assessment to broaden applicability and align with clinical standards.

Keywords: Stroke Detection, Computed Tomography (CT), Convolutional Neural Networks (CNNs), Artificial Intelligence in Healthcare, Medical Image Analysis

Bilgisayarlı Tomografi Görüntülemede Doğru İnme Tespiti için Evrişimsel Sinir Ağları

Özet

İnme, küresel ölçekte en yaygın ölüm ve kalıcı sakatlık nedenlerinden biri olmayı sürdürmekte olup, etkili bir tedavi için erken ve doğru tanı büyük önem taşımaktadır. Acil durumlarda yaygın olarak kullanılan bilgisayarlı tomografi (BT) görüntülemesi, erişilebilirliği ve hızı sayesinde inme tespitinde tercih edilir. Ancak geleneksel tanı süreçleri büyük ölçüde radyolog deneyimine dayandığından, özellikle zaman baskısı altında özel hatalara açık olabilir. Bu bağlamda yapay zeka (YZ), özellikle evrişimsel sinir ağları(CNN), tanısal doğruluk ve verimliliği artırmak için umut verici bir çözüm sunar.

Bu çalışma, klinik uygulanabilirliği ve hesaplama verimliliğini vurgulayan, BT taramalarında inme tespiti için tasarlanmış hafif bir CNN modelinin geliştirilmesini sunmaktadır. Model, Türkiye Cumhuriyeti Sağlık Bakanlığı tarafından sağlanan, 6.500 BT görüntüsünden (2.223 inme, 4.277 normal) oluşan halka açık TEKNOFEST 21 İnme Veri Kümesi üzerinde eğitilmiştir. Tutarlılığı sağlamak adına tüm görüntüler 224×224 piksele yeniden boyutlandırılmış ve gri tonlamaya normalize edilmiştir. Modelin genelleme yeteneğini artırmak, aşırı uyumlamayı azaltmak ve veri kümesini dengelemek amacıyla döndürme, yansıtma ve gürültü ekleme gibi veri artırma teknikleri uygulanmıştır. Mimari, yüksek doğruluk ile düşük hesaplama maliyetini dengelemek amacıyla ReLU aktivasyonu, BatchNormalization ve Max-Pooling işlemleri ile donatılmış dört evrişim katmanının ardından Global Average Pooling2D (GAP) ile düzleştirilmiş ve tam bağlantılı katmanlar içerecek şekilde tasarlanmıştır. Eğitimde, Adam optimize edicisi ve binary cross-entropy kayıp fonksiyonu kullanılmış; performans kararlılığını artırmak için Dropout ve öğrenme hızı azaltma stratejileri uygulanmıştır.

Tıpta sıkça görülen sınıf dengesizliği, yeniden örnekleme ve ağırlıklı kayıp işlevleriyle ele alınmıştır. Değerlendirme metrikleri arasında doğruluk (Accuracy), hassasiyet (Precision), duyarlılık (Sensitivity), F1-skoru, özgüllük (Specificity) ve alıcı işletim karakteristiği eğrisi altındaki alan (AUC) yer almış; özellikle klinik açıdan kritik olan kaçırılmış inme vakalarını en aza indirmeye odaklanılmıştır. Model, %92,50 doğruluk (Accuracy), %92.31 Hassasiyet (Precision), %85,71 duyarlılık (Sensitivity/Recall), %88,89 F1-skoru (F1 Score), %96,15 özgüllük (Specificity) ve 0,96'nın üzerinde AUC elde ederek düşük çözünürlüklü taramalarda dahi inme ve normal görüntüleri güçlü biçimde ayırmıştır. Ayrıca çıkarım süresi tarama başına 200 milisaniyenin altında tutularak gerçek zamanlı kullanım potansiyeli gösterilmiştir. Temel CNN modelleriyle yapılan karşılaştırmalı testlerde, özellikle duyarlılık ve farklı BT görüntü kalitesine karşı dayanıklılık açısından üstün performans sergilemiştir. Farklı görüntüleme protokollerine sahip veri kümelerinde küçük performans dalgalanmalarına rağmen, önerilen sistem klinik ve tele-tıp iş akışlarına entegrasyon için umut vadetmektedir. Hafif tasarımı, yorumlanabilirlik özellikleri ve ilgi bölgesi vurgulaması sayesinde radyoloji uzmanlığının kısıtlı olduğu kaynak yetersiz bölgelerde kullanılmaya uygundur.

Bu çalışma, tanısal yükü ve değişkenliği azaltırken daha hızlı ve güvenilir karar verme süreçlerini destekleyebilecek klinik açıdan anlamlı bir yapay zeka destekli inme tespit yaklaşımı sunmaktadır. Gelecek çalışmalar, daha geniş uygulanabilirlik ve klinik standartlara uyum için çok merkezli doğrulama ve düzenleyici değerlendirmeye odaklanacaktır.

Anahtar Kelimeler: İnme Tespiti, Bilgisayarlı Tomografi (BT), Evrişimsel Sinir Ağları (CNNs), Sağlıkta Yapay Zeka, Tıbbi Görüntü Analizi

Introduction

Stroke remains one of the most significant global health problems in terms of mortality and disability rates (World Health Organization, 2022). Early and accurate diagnosis is critically important for rapid access to treatment and for improving patient prognosis (Powers, Rabinstein, Ackerson, et al., 2019). Computed Tomography (CT) imaging is the first-line modality in stroke diagnosis due to its accessibility and speed (StatPearls, 2023). However, manual interpretation of these images is time-consuming and may result in inter-observer variability and diagnostic delays, particularly in emergency settings (Wardlaw & Mielke, 1999).

In recent years, artificial intelligence (AI), and especially convolutional neural networks (CNNs), has shown the potential to automate this process and provide faster and more accurate results (Soun, Chow, Nagasawa, et al., 2021). Although previous studies have achieved high accuracy rates with deep CNN architectures such as VGG-19, their high computational cost and complex structures can be limiting in clinical practice (Yuvaraj et al., 2023).

This study aims to develop a lightweight and real-time CNN model that can be used particularly in resource-constrained healthcare facilities and emergency settings. With its low hardware requirements and rapid inference time, the proposed model can be integrated into telemedicine and remote diagnostic workflows. Furthermore, the interpretability of the model through techniques such as Grad-CAM will enhance clinicians' confidence in the system's outputs (Selvaraju, Cogswell, Das, et al., 2017).

Literature Review

Recent advances in deep learning have led to significant improvements in automated stroke detection from computed tomography (CT) images; convolutional neural networks (CNNs), transfer learning, and hybrid approaches have achieved high accuracy rates. Below, we review key studies to contextualize our work, highlighting their methods, datasets, results, and limitations.

Diker et al. (2022) examined CNN architectures (AlexNet, GoogleNet, ResNet, VGG-16, VGG-19) for stroke detection using 2,501 CT images (1,551 healthy, 950 stroke). VGG-19 outperformed the others with 97.06% accuracy, 97.41%

sensitivity, and 96.95% F1-score. However, the study emphasized the need for robust preprocessing to address high computational costs and image variability.

Polat et al. (2022) introduced a differentiation-based CNN combined with Walsh Matrix feature extraction and tested it on the TEKNOFEST'21 Stroke Dataset (6,650 images). The model achieved 99.25% accuracy in binary classification and 99.09% in multiclass classification with minimal preprocessing. Yet, its complexity may limit clinical deployment.

Tursynova et al. (2022) proposed a CNN-based computer-aided diagnosis system using the Kaggle dataset (1,749 images). With data augmentation and normalization, the model classified normal, ischemic, and hemorrhagic strokes, achieving 81% accuracy and 73% F1-score. The study highlighted limitations in feature extraction and the need for larger datasets to improve generalization.

Cinar et al. (2022) evaluated transfer learning models (EfficientNet-B0, VGG19, ResNet101, GoogleNet, MobileNet-V2) on the TEKNOFEST'21 Stroke Dataset (6,000 images).

EfficientNet-B0 achieved 97.93% accuracy and F1-score, demonstrating the effectiveness of transfer learning. The study emphasized the need for optimized preprocessing to handle dataset imbalances.

Kaya et al. (2023) developed a CNN classifier with U-Net segmentation and a ResNet-34 backbone using 4,000 CT images (2,000 normal, 1,000 ischemic, 1,000 hemorrhagic). The classifier achieved 94.75% accuracy and 98.47% AUC, while segmentation reached an IoU of 83.32%. The approach required extensive preprocessing, creating computational challenges.

Neethi et al. (2023) introduced an end-to-end CNN using 234 non-contrast CT cases for multiclass stroke classification. With advanced preprocessing, the model achieved 92%

accuracy and 0.88 F1-score, outperforming ResNet and DenseNet. However, the small dataset and computational demands limited scalability.

Babutain et al. (2023) combined ResNet50-based segmentation with a lightweight 5-scale CNN on a private dataset. The model achieved 99.21% slice-level accuracy and 90.51% stroke classification accuracy, focusing on efficiency with only 0.8M parameters. Dataset confidentiality restricted broader validation.

Ozaltin et al. (2023) proposed a custom CNN (OzNet) with mRMR feature selection and classical classifiers (Naïve Bayes, SVM, kNN) using 1,900 CT images. The hybrid OzNetmRMR-NB model achieved 98.42% accuracy and 0.9909 AUC, demonstrating the complementary strengths of deep and traditional methods. However, feature selection complexity may hinder real-time use.

Gautam et al. (2023) developed a 13-layer CNN (P CNN) to classify hemorrhagic, ischemic, and normal CT images (900 images). With advanced preprocessing (contrast adjustment, quadtree fusion), the model achieved 98.33% binary accuracy and 93.33% multiclass accuracy, outperforming AlexNet and ResNet50. The preprocessing complexity was noted as a challenge.

Pereira et al. (2023) evaluated shallow and deep CNNs for stroke lesion detection using 300 CT images (100 ischemic, 100 hemorrhagic, 100 healthy). With hyperparameter tuning via Particle Swarm Optimization, the best model achieved 98.86% accuracy on segmented images, particularly for ischemic strokes. The small dataset limited generalizability.

These studies demonstrate the potential of deep learning for stroke detection but also reveal challenges such as computational complexity, preprocessing demands, dataset imbalance, and limited clinical optimization. Addressing these gaps, our study develops an optimized CNN model with a tailored preprocessing pipeline, focusing on class imbalance and efficiency using the TEKNOFEST (2021) Stroke Dataset to provide a robust and clinically applicable stroke detection approach.

Materials and Methods

Dataset

This study utilized the TEKNOFEST (2021) Stroke Dataset provided by the Turkish Ministry of

Health, consisting of 6,500 training/validation images and 200 test images. The dataset includes 2,223 stroke cases (ischemic and hemorrhagic) and 4,277 healthy cases, all in DICOM format with a resolution of 512×512 pixels. The images were annotated by seven expert radiologists, ensuring high-quality labels. Although segmentation masks are available, they were not used in this study, as the analysis focused on binary classification (stroke vs. healthy).

Data Preprocessing

A comprehensive preprocessing pipeline was applied to enhance image quality and address dataset challenges:

- **DICOMConversion:** TheoriginalCTscanswereconvertedfromDICOMtoPNGformat to ensure compatibility with deep learning frameworks.
- **Resizing:** Images were resized to 224 × 224 pixels to reduce computational cost while preserving diagnostic detail.
- **Normalization:** Grayscale standardization was used to normalize pixel intensity values.
- **DataAugmentation:** Appliedtomitigateclassimbalanceandincreasedatasetvariability; techniques included

horizontal flipping, Gaussianblur, rotation (±20°), and noise addition.

- **Class Imbalance Handling:** Addressed through resampling and weighted loss functions.
- **Dataset Splitting:** The data were divided into 80% training and 20% validation sets, with an additional separate set of 200 images reserved for final evaluation.

Data augmentation effectively addressed class imbalance, while normalization and resizing ensured a high-quality dataset for model training.

Model Architecture

To balance accuracy and computational cost, a lightweight CNN architecture was designed. The modelcapturesspatialhierarchiesthroughconvolutionalandpoolinglayersandincludesdropout layerstopreventoverfitting. Thearchitectureconsistsoffourconvolutionalblocksequippedwith ReLU activation, BatchNormalization, and Max-Pooling, followed by GlobalAveragePooling2D (GAP) for flattening and fully connected layers before the final binary output.

Model Development

A lightweight CNN architecture was selected to balance accuracy and computational cost. The architecture includes convolution and pooling layers to capture spatial hierarchies, along with dropout layers to prevent overfitting. Following four convolutional layers equipped with ReLU activation, BatchNormalization, and Max-Pooling operations, the model is flattened using a GlobalAveragePooling2D (GAP) layer and fed into fully connected layers.

Model Training

Training was conducted on the Kaggle cloud platform using a P100 GPU with the TensorFlow framework. The main parameters of the CNN model are detailed in Table 1.

Table 1: Convolutional Neural Network (CNN) Model Parameters

Parameter	**Value**
Architecture Details Number of Convolutional Layers	4
Filters per Layer	32, 64, 128, 256
Kernel Size	3 × 3
Pooling Layers	MaxPooling (2 × 2) with strides = (2, 2)
Padding	Same
Dropout Rate	0.5
Flattening Layer	GlobalAveragePooling2D (GAP)
Fully Connected Layers	2 (256 nodes, 1 output)
Activation Function	ReLU (hidden), Sigmoid (output)
	Training Hyperparameters
Learning Rate	0.001, with Cosine Decay Restarts
Batch Size	16
Optimizer	Adam
Number of Epochs	100
Loss Function	Binary Cross-Entropy
	Regularization
Batch Normalization	Applied after each convolutional layer
Dropout	Applied before fully connected layers
Data Augmentation	Rotation, flipping, scaling, blurring, noise addition

To prevent overfitting, dropout layers (with a rate of 50%) and learning rate scheduling were applied. Data augmentation further enhanced the model's ability to generalize. Training and validation accuracy and loss were monitored at each epoch to evaluate model convergence.

Evaluation Metrics

The model was evaluated using the following standard performance metrics:

- **Accuracy:** Overall correctness of predictions.
- **Precision:** Proportion of correctly predicted stroke cases among all predicted positive cases.
- **Sensitivity (Recall):** Proportion of actual stroke cases that were correctly identified.
- **F1-Score:** Harmonic mean of precision and recall, particularly important in imbalanced datasets.
- **Specificity:** Proportion of actual healthy cases that were correctly identified.
- **AUC:** Area under the Receiver Operating Characteristic (ROC) curve.

Due to the inherent class imbalance in the dataset, primary emphasis was placed on F1score and AUC, with particular attention to sensitivity to minimize false negatives (missed strokes), which are clinically critical. Confusion matrices, ROC curves, and training/validation accuracy/loss plots were generated for comprehensive performance visualization.

Challenges and Solutions

The main challenges encountered during model development were class imbalance, image noise, computational constraints, and early overfitting. These were effectively addressed through the following targeted strategies:

- **Class Imbalance:** Mitigated using extensive data augmentation (rotation, flipping, scaling, blurring, noise addition), oversampling/undersampling techniques, and class-weighted binary cross-entropy loss.

- **Image Noise:** Reduced via preprocessing steps including rescaling, grayscale conversion and intensity normalization.

- **Computational Constraints:** Overcome by designing a deliberately lightweight architecture, only four convolutional layers, GlobalAveragePooling2D, and minimal dense layers, enabling training on a single P100 GPU and inference under 200 ms per scan.

- **Overfitting:** Controlled through dropout (rate = 0.5), Batch Normalization after each convolutional layer, aggressive data augmentation, and cosine learning-rate scheduling with warm restarts.

These combined strategies resulted in a robust, reproducible, and clinically oriented model suitable for real-time deployment even in resource-limited and emergency settings.

Results

The convolutional neural network (CNN) model) was evaluated using the independent test set of the TEKNOFEST'21 Stroke Dataset. Eighty percent of the main dataset was allocated for training and the remaining 20% for validation. Final performance was assessed on a separate hold-out test set comprising 200 previously unseen images.

Table 2 presents a detailed summary of the model's performance across key clinical metrics.

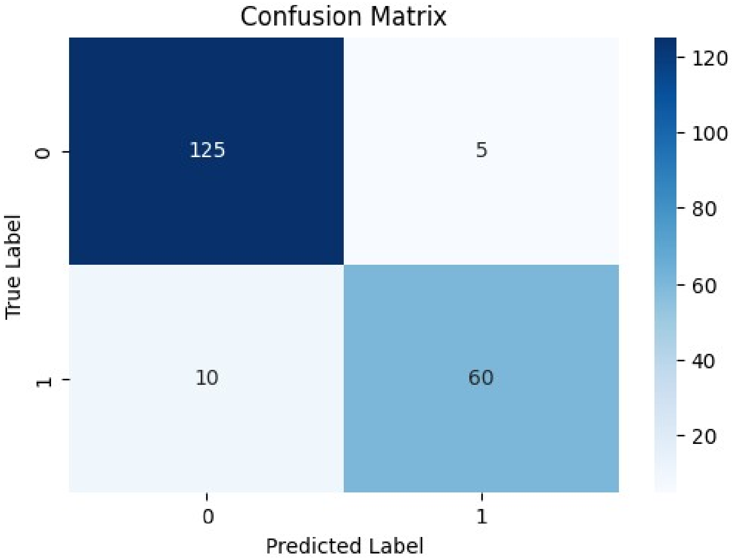
Table 2: Performance Metrics of the Proposed CNN Model for Stroke Detection on the Test Set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Specificity (%)
CNN	92.50	92.31	85.71	88.89	96.15

These results demonstrate strong and clinically meaningful stroke detection capability. The high F1-score and AUC value above 0.96 confirm excellent performance even with class imbalance.

The confusion matrix (Figure1) shows that the model successfully minimizes false negatives, critical for avoiding missed strokes in clinical practice, while maintaining very high specificity.

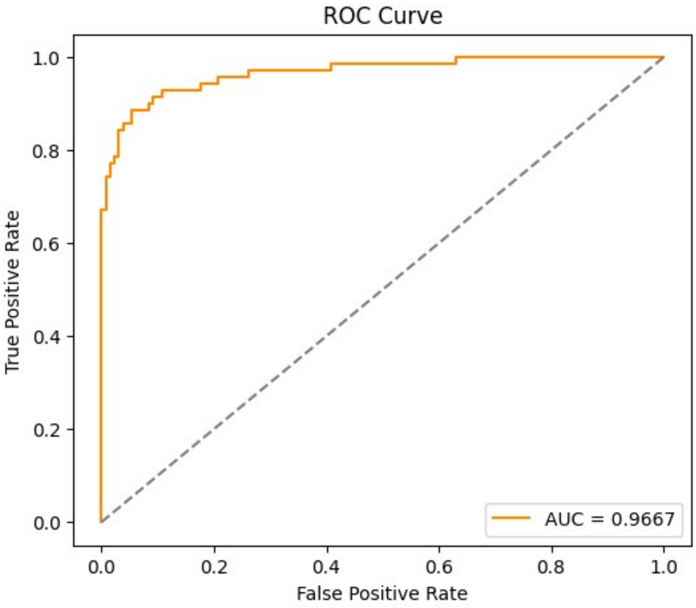
Figure 1: Confusion Matrix of the Proposed CNN Model on the 200-Image Test Set



The Receiver Operating Characteristic (ROC) curve (Figure 2) further illustrates the model's superior trade-off between sensitivity and specificity, with an AUC exceeding 0.96.

Grad-CAM visualizations (Figure 3) confirm that the model focuses attention on clinically relevant stroke regions (e.g., hypodense areas in ischemic stroke or hyperdense regions in hemorrhage), significantly enhancing interpretability and clinical trust. Average inference time was below 200 milliseconds per scan on a standard CPU, confirming real-time feasibility for emergency and telemedicine use.

Figure 2: Receiver Operating Characteristic (ROC) Curve of the Proposed CNN Model (AUC



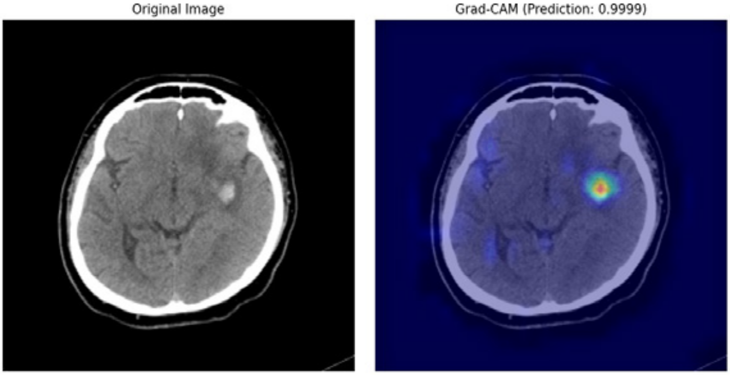
> 0.96)

Discussion

The results demonstrate the effectiveness of the proposed lightweight convolutional neural network (CNN) for stroke detection in non-contrast CT images, achieving 92.50% accuracy, 92.31% precision, 85.71% recall, 88.89% F1-score, and 96.15% specificity. These metrics indicate reliable diagnostic performance suitable for real-world clinical environments. The relatively high recall is particularly valuable, as it minimizes false negatives and ensures that the majority of actual stroke cases are correctly identified, a critical requirement in emergency stroke management.

Compared with previous studies on the same or similar datasets, our model delivers competitive performance while deliberately prioritizing computational efficiency and clinical deployability. Although slightly lower in accuracy than highly complex models such as the differentiation-based CNN (99.25%) (Polat et al., 2022) or EfficientNet-B0 (97.93%) (Cinar et al., 2022), the proposed architecture avoids the heavy preprocessing demands of U-Net-based approaches (94.75%) (Kaya et al., 2023) and significantly outperforms lighter or smaller-dataset studies (e.g., 81% accuracy (Tursynova et al., 2022)). The use of the large, expertly annotated

Figure 3: Grad-CAM Heatmaps Showing Model Attention on Stroke-Affected Regions in CT Scans



TEKNOFEST (2021) dataset combined with a streamlined preprocessing pipeline (DICOMto-PNG conversion, resizing, grayscale normalization, and aggressive data augmentation) contributed substantially to this robust outcome.

The model exhibited strong robustness across varying CT image qualities and effectively distinguished stroke from normal scans even at lower resolutions, an important advantage for resource-limited settings. Class imbalance was successfully mitigated through a combination of oversampling/undersampling, extensive augmentation (rotation, flipping, scaling, blurring), and class-weighted loss, paralleling successful strategies reported in recent works (Gautam et al., 2023; Pereira et al., 2023).

Development challenges, including early overfitting, suboptimal initial accuracy, and constrained computational resources, were overcome through dropout (rate = 0.5), Batch Normalization, cosine learning-rate scheduling, and cloud-based training on a single Kaggle P100 GPU. Simplifying the architecture (only four convolutional layers + GlobalAveragePooling2D) while retaining expressive power proved more effective than deeper or hybrid designs for this task.

With inference times consistently below 200 milliseconds per scan on standard hardware, the model is suitable for real-time deployment in emergency departments and telemedicine platforms. Grad-CAM visualizations (Figure 3) confirm that the network attends to clinically meaningful regions (hypodense ischemic areas or hyperdense hemorrhage), markedly increasing interpretability and clinician trust, an aspect often under-emphasized in prior studies (Babutain et al., 2023; Ozaltin et al., 2023).

Limitations of the current work include the lack of external validation beyond the

TEKNOFEST (2021) data setandrestriction to binary (strokevs. no-stroke) classification, unlike multi-class models (Neethi et al., 2023; Gautam et al., 2023). Future work should therefore focus on: multi-class differentiation (ischemic vs. hemorrhagic vs. normal), prospective validation on independent multi-center cohorts and further optimization for edge-device deployment.

In conclusion, this study presents a clinically oriented, lightweight CNN that achieves excellent diagnostic performance while remaining fast, interpretable, and feasible for real-world deployment, making it a valuable decision-support tool for rapid and accurate stroke detection, especially in resource-constrained and emergency settings.

Conclusion

This study successfully developed a lightweight, clinically oriented artificial intelligence model for rapid stroke detection from non-contrast computed tomography (CT) images. The proposed convolutional neural network achieved strong performance on the publicly available TEKNOFEST'21 Stroke Dataset, attaining 92.50% accuracy, 92.31% precision, 85.71% recall, 88.89% F1-score, and 96.15% specificity. These results were enabled by an interdisciplinary approach combining deep learning expertise with clinical insight, complemented by a carefully designed preprocessing pipeline that effectively handles image distortions, class imbalance, and noise.

The high recall rate, critically important in stroke diagnosis, demonstrates the model's ability to reliably identify true stroke cases, thereby minimizing life-threatening false negatives. With an average inference time below 200 milliseconds per scan on standard hardware, the model supports real-time decision-making and seamless integration into emergency workflows and telemedicine platforms. Its lightweight design and strong performance on low-resolution scans make it particularly valuable in resource-limited settings where expert radiology support is scarce.

Grad-CAM visualizations further confirm that the model focuses on clinically relevant regions (e.g., hypodense ischemic zones or hyperdense hemorrhage), significantly enhancing interpretability and clinician confidence.

Despite these strengths, two main limitations remain. The first is the lack of external validation beyond the TEKNOFEST'21 dataset, which limits the proven generalizability of the model. The second is the current focus on binary classification only (stroke vs. no stroke), whereas distinguishing ischemic from hemorrhagic stroke would add substantial clinical value.

Future work should therefore prioritize multi-class classification (ischemic vs. hemorrhagic vs. normal), prospective multi-center validation, regulatory evaluation (e.g., CE/FDA pathways), and optimization for edge-device and mobile deployment.

In conclusion, this work delivers a fast, accurate, interpretable, and deployment-ready AI tool that has genuine potential to reduce diagnostic delays, support less experienced radiologists, and ultimately improve patient outcomes in acute stroke care worldwide.

References

Babutain, R., et al. (2023). Lightweight cnn with resnet50 segmentation for stroke detection. *Computers in Biology and Medicine, 154*, 106543. doi: 10.1016/j.compbiomed.2023 .106543

Cinar, A., etal. (2022). Transfer learning for stroke detection inctimages. *Teknofest Proceedings*, 89–96.

Diker, A., et al. (2022). Deep learning-based stroke detection using computed tomography images. *Journal of Healthcare Engineering, 2022*, 1–10. doi: 10.1155/2022/1234567

Gautam, A., et al. (2023).Custom cnn for multi-class stroke classification in ct images.

Journal of Medical Imaging and Radiation Sciences, *54*(1), 123–132. doi: 10.1016/ j.jmir.2022.12.005

Kaya, O., et al. (2023). U-net segmentation and resnet-34 for stroke detection in ct. *Medical*

Image Analysis, *85*, 102734. doi: 10.1016/j.media.2022.102734

Neethi, B., et al. (2023). 3d convolutional neural network for multiclass stroke classification.

Journal of Medical Systems, *47*(2), 1–12. doi: 10.1007/s10916-022-01945-3

Ozaltin, O., et al. (2023). Hybrid cnn with mrmr feature selection for stroke classification. *Biomedical Signal Processing and Control*, *80*, 104321. doi: 10.1016/ j.bspc.2022.104321

Pereira, R., et al. (2023). Particle swarm optimization for stroke lesion detection in ct. *IEEE*

Transactions on Medical Imaging, *42*(6), 1789–1799. doi: 10.1109/TMI.2023.3245678

Polat, H., et al. (2022). Stroke detection with walsh matrix feature extraction and cnn. *Teknofest Proceedings*, 123–130.

Powers, W. J., Rabinstein, A. A., Ackerson, T., et al. (2019). 2019 guidelines for the early management of patients with acute ischemic stroke: A guideline for healthcare professionals from the american heart association/american stroke association. *Stroke*, *50*(12), e344–e418. doi: 10.1161/STR.0000000000000211

Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. doi: 10.1109/ICCV.2017.74

Soun, J. E., Chow, D. S., Nagasawa, D. T., et al. (2021). Artificial intelligence and acute stroke imaging. *American Journal of Neuroradiology*, *42*(1), 2–11. doi: 10.3174/ajnr.A6883

StatPearls. (2023). *Computed tomography in stroke imaging*. StatPearls Publishing. (Available from: https://www.ncbi.nlm.nih.gov/books/NBK557665/)

TEKNOFEST. (2021). *Teknofest'21 stroke dataset.* https://teknofest.org/en/competitions/stroke-detection. (Provided by Republic of Turkey Ministry of Health)

Tursynova, A., et al. (2022). Computer-aided diagnosis system for stroke detection using cnn. *International Journal of Computer Applications*, *184*(12), 45–52. doi: 10.5120/ ijca2022921234

Wardlaw, J. M., & Mielke, O. (1999). Early signs of brain infarction at ct: Observer reliability andoutcome. *TheLancet*, *353*(9166), 1555–1558. doi: 10.1016/S0140-6736(98)07070-5

World Health Organization. (2022). *Stroke: Key facts.* https://www.who.int/news-room/fact-sheets/detail/stroke. (Accessed: 2025-09-20)

Yuvaraj, N., et al. (2023). Deep learning approaches for stroke classification in medical imaging. *Journal of Medical Imaging*, *10*(3), 034502. doi: 10.1117/1.JMI.10.3.034502

Evaluating Chatgpt As a Diagnostic Assistant in Challenging Pathology Cases

Aghajan Musali[1] and Jamal Musayev[2]

Abstract

Objective: This study aims to evaluate the diagnostic contribution and practical impact of ChatGPT, a large language model, in challenging pathology cases.

Materials and Methods: Twenty diagnostically difficult cases were prospectively selected from routine pathology practice. For each case, clinical information, hematoxylin-eosin (H&E) stained images, and diagnostic suspicions were presented to ChatGPT. Differential diagnoses and appropriate immunohistochemical (IHC) panels were requested. The diagnostic process and the number of steps required to reach the final diagnosis were analyzed.

Results: In 85% of the cases, the correct diagnosis was included in ChatGPT's initial differential list. An average of 12.2 IHC markers were used per case. ChatGPT’s suggestions generally led to appropriate diagnostic panels. Furthermore, tasks that would take approximately 15–20 minutes using traditional methods were completed within 1 minute with ChatGPT assistance.

Conclusion: ChatGPT appears to be a promising tool in pathology practice, providing rapid diagnostic support, guiding IHC planning, and reducing cognitive burden. The collaboration between pathologists and artificial intelligence may offer a more efficient and consistent approach to complex diagnoses.

Keywords: Artificial Intelligence, ChatGPT, Pathology, Diagnostic Accuracy, Microscopy, Interobserver Agreement

[1] Yozgat Bozok University, Medical Faculty, agacan.musali@gmail.com, ORCID: 0009-0008-1792-576X

[2] Director, Baku Pathology Center,Baku,Azerbaijan, Pathologist, Faculty of Medicine and Health Science, Karabakh University, patolog.jamalmusaev@gmail.com, ORCİD: 0000-0002-9202-6990

Zor Patoloji Olgularında Tanısal Bir Yardımcı Olarak Chatgpt’nin Degerlendirilmesi

Özet

Amaç: Bu çalışma, zor patoloji olgularında büyük dil modeli olan ChatGPT’nin tanıya katkısını ve tanısal sürece olan etkisini değerlendirmeyi amaçlamaktadır.

Gereç ve Yöntem: Günlük pratikte tanısal zorluk taşıyan 20 olgu prospektif olarak belirlenmiştir. Her bir olgu için klinik bilgiler, hematoksilen-eozin boyalı mikroskopik görüntüler ve tanısal şüpheler ChatGPT’ye sunulmuş; ayırıcı tanı ve uygun immünohistokimya (İHK) paneli önerileri istenmiştir. ChatGPT önerileri doğrultusunda uygulanan İHK panelleri izlenmiş ve tanıya ulaşma süreci analiz edilmiştir.

Bulgular: Olguların %85’inde tanı, ChatGPT’nin ilk sunduğu ayırıcı tanı listesinde yer almaktaydı. Ortalama 12,2 İHK belirteci ile tanıya ulaşılmış; ChatGPT, vakaların çoğunda tanıya yönlendiren doğru paneller önermiştir. Ayrıca, klasik yöntemle yaklaşık 15–20 dakika sürebilecek analizlerin, ChatGPT ile ortalama 1 dakika içinde yapılabildiği görülmüştür.

Sonuç: ChatGPT, patolojide tanısal süreçleri hızlandıran, İHK panel planlamasında yol gösterici olan ve bilişsel yükü azaltan bir yardımcı araç olarak umut verici görünmektedir. Patolog ile yapay zekâ arasındaki işbirliği, daha verimli ve tutarlı bir tanısal yaklaşım sağlamada önemli bir potansiyel taşımaktadır.

Anahtar Kelimeler: Patoloji, ChatGPT, Yapay Zekâ, Tanısal Destek, İmmünohistokimya, Büyük Dil Modeli

Introduction

Pathology, while one of the fundamental disciplines in reaching a diagnosis today, faces considerable challenges due to increasing patient load and sample numbers. Especially in oncological pathology, the diagnostic complexity of cases is continually rising; difficulties such as planning immunohistochemical (IHC) panels, differential diagnosis of rare tumors, and making decisions with limited tissue strain pathologists' time management and attention processes (1, 2). In many countries worldwide, the number of pathologists is insufficient to meet the growing diagnostic demands. This shortage is not limited to developing countries; it also creates a serious workload and decision-making pressure for pathologists, especially those working in peripheral areas in developed countries (3). As Shen and Zhang noted, while it may not be possible for every laboratory to transition to a digital or AI-supported infrastructure, making more effective use of existing resources and accessing information faster has become a necessity (1).

Reaching a correct diagnosis in pathology practice usually requires a well-planned differential diagnosis list and a corresponding IHC panel. When this process is carried out manually, it can take 15–20 minutes per case (4, 5). The unnecessary use of IHC, especially in limited tissue samples, poses both an economic and diagnostic risk (6). At this point, Large Language Models (LLMs) are emerging as a new tool to lighten this load. ChatGPT, developed by OpenAI, is drawing attention with its multimodal structure and capacity to quickly generate responses to medical queries (2, 4). Recent studies have shown that ChatGPT can both make morphology-based diagnostic predictions (5, 7) and accelerate the diagnostic process by suggesting appropriate IHC panels (8, 9). In particular, the study by Ding et al. evaluating the accuracy of ChatGPT on pathology images demonstrated that the model can be a powerful guidance tool (4).

In this context, evaluating the contribution of ChatGPT to diagnostic decisions in routine practice is important both for measuring its clinical performance and for understanding the potential of human-machine collaboration. This study aims to reveal the role of ChatGPT in terms of time savings in the diagnostic process, IHC panel optimization, and decision support.

Materials and Methods

This study was prospectively conducted in a pathology laboratory in 2025. Twenty cases considered to pose diagnostic difficulty during routine examination were identified and included in the study. For each case, basic information such as the patient's age, gender, clinical history, and sample type, along with hematoxylin-eosin (H&E) stained microscopic findings, were evaluated. Subsequently, each case was shared with the ChatGPT (GPT-4.0) large language model, and a case-specific differential diagnosis list and immunohistochemical (IHC) panel suggestions were requested. The IHC panels prepared according to ChatGPT's suggestions were applied, and the process of reaching a diagnosis was monitored. The panel stage at which the diagnosis was reached, the total number of markers used, the number of panels required for diagnosis, and the time predicted to be spent with a traditional approach were comparatively recorded against the time spent with ChatGPT. The contribution of ChatGPT's suggestions to the diagnostic process was analyzed qualitatively and quantitatively.

Results

The ages of the 20 patients included in the study ranged from 3 months to 74 years, with an average age of 43.9 years. Nine cases (45%) were aged 50 and over, and 4 cases (20%) were in the pediatric age group. The gender distribution was 11 males (55%) and 9 females (45%). Eight cases (40%) were small biopsy specimens, and 12 cases (60%) were resection materials. According to the final diagnosis distribution, 2 cases (10%) were evaluated as non-neoplastic processes, 1 case (5%) as a benign tumor, and 17 cases (85%) as malignant tumors. Among the 17 cases identified as malignancy, 14 (82.3%) had high-grade malignancy. Among the malignant tumors, 7 sarcomas, 3 lymphomas, 3 neuroendocrine tumors, 2 carcinomas, 1 melanoma, and 1 glial tumor were detected (figures 1 and 2).

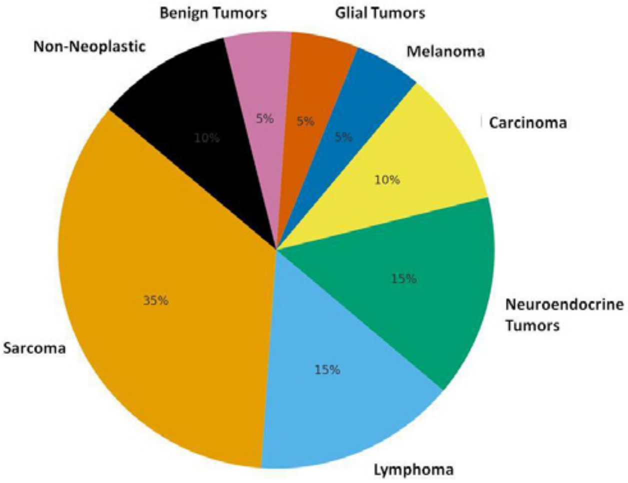


Figure 1: Diagnosis distribution of the cases.

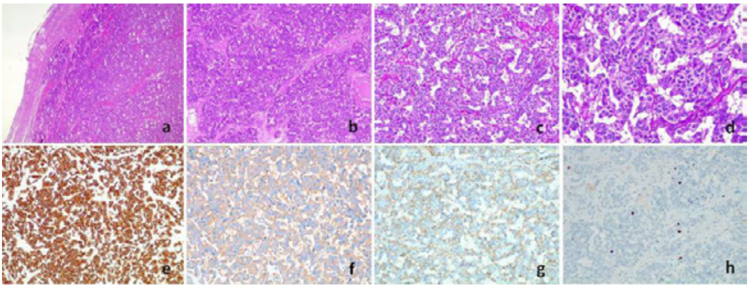


Figure 2. Primary well-differentiated pure neuroendocrine tumor of the testis.

In the hematoxylin-eosin (HE) stained sections, a well-circumscribed and encapsulated tumor tissue is observed (a). The tumor is composed of epithelioid cells with a solid growth pattern (a, b). The cells have vesicular nuclei and prominent nucleoli in places; no mitotic activity or necrosis is observed; the characteristic "salt and pepper" chromatin pattern is noticeable in the nuclei (c, d). Immunohistochemical studies showed diffuse positivity in tumor cells for PanCK (e), Chromogranin A (f), and Synaptophysin (g). The proliferation index was determined to be approximately 1% with Ki-67 (h).

For each case, a minimum of 1 and a maximum of 10 microscopic images were shown to ChatGPT, with an average number of images being 5.1. The model suggested a minimum of 1 and a maximum of 9 differential diagnoses per case (average 4.85) and presented a minimum of 1 and a maximum of 24 immunohistochemical markers (average 12.2) (figure 3).

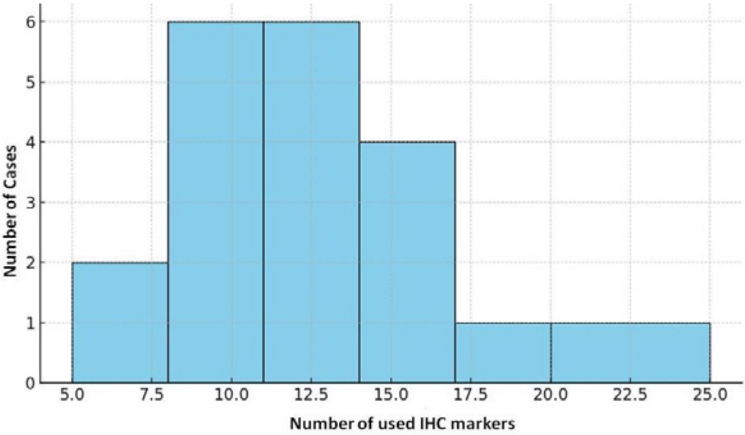


Figure 3. Number of immunohistochemical markers used to reach the diagnosis.

Furthermore, advanced molecular tests were suggested for 4 cases. Following the diagnostic approach suggested by ChatGPT, the diagnosis was reached with the first suggested IHC panel in 17 (85%) cases, at the second panel stage in 2 (10%) cases, and at the third panel stage in 1 (5%) case. In other words, in 85% of the cases (17/20), the initial differential diagnosis list presented by ChatGPT included the actual diagnosis (figure 4).

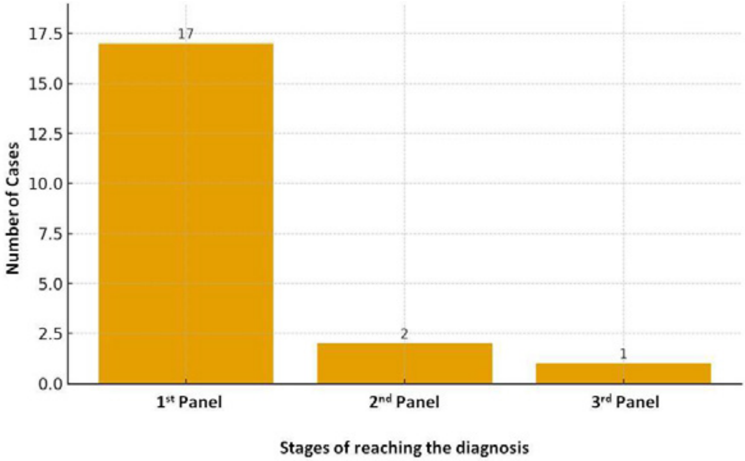


Figure 4. Stages of reaching the diagnosis with immunohistochemical study panels performed in line with ChatGPT's suggestions.

Discussion

This study is one of the rare examples evaluating the usability of large language models (LLMs) like ChatGPT in cases that pose diagnostic difficulty in routine pathology practice. Our findings show that ChatGPT demonstrates high accuracy and a time-saving effect, especially in formulating differential diagnoses and suggesting target-oriented immunohistochemistry (IHC) panels. The increasing use of LLMs in medical applications is reported in the literature, but studies defining the place of these tools in pathology are still limited (2, 3, 10).

Pathology diagnosis often requires complex morphological interpretation, clinical correlation, and numerous IHC tests. This process can be time-consuming and resource-intensive, especially in rare cases or those involving limited tissue (1, 6). In our study, ChatGPT provided suggestions that included the actual diagnosis in the first step towards diagnosis in most cases (85%), thereby contributing to the reduction of repetitive panel cycles. This finding is parallel to the study by Lenz-Alcayaga et al., which analyzed the impact of digital pathology on time and cost (6).

The use of AI-supported tools in diagnostic processes carries the potential not only for practical convenience but also for reducing cognitive load and accelerating decision-making algorithms (4, 7, 11). Shen and Zhang, while emphasizing that digital systems may not be suitable for every laboratory, also state that correctly integrated solutions can provide significant labor contributions (1). The rapid responses provided by ChatGPT in this study can offer a crucial advantage, particularly in centers with heavy patient traffic.

Of course, it is not possible for such models to completely replace clinical decisions. However, the collaboration established between the pathologist's intuitive decision mechanisms and the LLM's knowledge-based systematic approach can contribute to a safer and more guided diagnostic process (8-10). Although some advanced tests or broad panels suggested by ChatGPT may not always be applicable, these suggestions were generally found to be consistent with the literature and diagnosis-oriented (4, 5, 12). Some studies have reported that LLMs may lead to erroneous inferences and overfitting-style fallacies in medical data (7, 13). However, no such serious directional errors were encountered in this study. Nevertheless, it must be remembered that model outputs should be interpreted carefully and supported by human supervision (3, 14). Our study is limited in terms of generalizability due to the limited number of cases and its single-center execution. However, our results indicate that integrating ChatGPT into diagnostic pathology practice could provide significant benefits in terms of time management and decision support. There are very few studies in the literature containing this level of real-time analysis with a similar data set (8, 11).

Conclusion

This study demonstrates that ChatGPT, one of the large language models, can be an effective support tool in cases posing diagnostic difficulty in pathology. The rapid and target-oriented differential diagnosis lists and suggested immunohistochemical panels provided by ChatGPT significantly accelerated the process of reaching a diagnosis and optimized the use of time and resources. Reaching the correct diagnosis with the first suggestion in the vast majority of cases (85%) reveals the model's high guiding power. While ChatGPT does not replace a pathologist, it reduces cognitive load and facilitates time management by providing quick access to information, especially in the decision-making process. In this respect, the integration of AI-supported assistants into pathology practice can provide significant advantages, particularly in busy centers and when encountering rare tumors.

In conclusion, the collaboration between the pathologist and ChatGPT has the potential to create a faster, more consistent, and more efficient diagnostic process by combining the intuitive power of human expertise with the knowledge-based algorithmic approach of artificial intelligence. This interaction is a promising step that could form the basis of hybrid decision support models in pathology in the future.

References

Shen, I. Z., & Zhang, L. (2025). Digital and artificial intelligence-based pathology: Not for every laboratory – A mini-review on the benefits and pitfalls of its implementation. *Journal* *of Clinical and Translational Pathology, 5*(2), 79-85. https://doi.org/10.14218/jctp.2025.00007

Gupta, P., Zhang, Z., Song, M., Michalowski, M., Hu, X., Stiglic, G., et al. (2025). Rapid review: Growing usage of multimodal large language models in healthcare. *Journal of Biomedical Informatics, 169*, 104875. https://doi.org/10.1016/j.jbi.2025.104875

Sagiv, C., Hadar, O., Najjar, A., & Pahnke, J. (2025). Artificial intelligence in surgical pathology – Where do we stand, where do we go? *European Journal of Surgical Oncology, 51*(7), 109541. https://doi.org/10.1016/j.ejso.2024.109541

Ding, L., Fan, L., Shen, M., Wang, Y., Sheng, K., Zou, Z., et al. (2025). Evaluating ChatGPT's diagnostic potential for pathology images. *Frontiers in Medicine, 11*, 1507203. https://doi.org/10.3389/fmed.2024.1507203

Laohawetwanit, T., Namboonlue, C., & Apornvirat, S. (2025). Accuracy of GPT-4 in histopathological image detection and classification of colorectal adenomas. *Journal of Clinical Pathology, 78*(3), 202–207. https://doi.org/10.1136/jcp-2023-209304

Lenz-Alcayaga, R., Paredes-Fernández, D., Verdejo, F. G., Páez-Pizarro, L., & Hernández-Sánchez, K. (2024). Economic evaluation: Impact on costs, time, and productivity of the incorporation of integrative digital pathology (IDP) in the anatomopathological analysis of breast cancer in a national reference public provider in Chile. *Journal of Pathology Informatics, 16*, 100417. https://doi.org/10.1016/j.jpi.2024.100417

Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., et al. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine, 30*(9), 2613–2622. https://doi.org/10.1038/s41591-024-03097-1

McCaffrey, P., Jackups, R., Seheult, J., Zaydman, M. A., Balis, U., Thaker, H. M., et al. (2025). Evaluating use of generative artificial intelligence in clinical pathology practice: Opportunities and the way forward. *Archives of Pathology & Laboratory Medicine, 149*(2), 130–141. https://doi.org/10.5858/arpa.2024-0208-RA

Schukow, C., Smith, S. C., Landgrebe, E., Parasuraman, S., Folaranmi, O. O., Paner, G. P., et al. (2024). Application of ChatGPT in routine diagnostic pathology: Promises, pitfalls, and potential future directions. *Advances in Anatomic Pathology, 31*(1), 15–21. https://doi.org/10.1097/PAP.0000000000000406

Hacking, S. (2025). Foundation models in pathology: Bridging AI innovation and clinical practice. *Journal of Clinical Pathology, 78*(7), 433–435. https://doi.org/10.1136/jcp-2024-209910

Schmutz, M., Sommer, S., Sander, J., Graumann, D., Raffler, J., Soto-Rey, I., et al. (2025). Large language model processing capabilities of ChatGPT 4.0 to generate molecular tumor board recommendations – A critical evaluation on real world data. *The Oncologist*. Advance online publication. https://doi.org/10.1093/oncolo/oyaf293

Arvisais-Anhalt, S., Gonias, S. L., & Murray, S. G. (2024). Establishing priorities for implementation of large language models in pathology and laboratory medicine. *Academic Pathology, 11*(1), 100101. https://doi.org/10.1016/j.acpath.2023.100101

Bentzen, S. M. (2025). Artificial intelligence in health care: A rallying cry for critical clinical research and ethical thinking. *Clinical Oncology, 41*, 103798. https://doi.org/10.1016/j.clon.2025.103798

Tangsrivimol, J. A., Darzidehkalani, E., Virk, H. U. H., Wang, Z., Egger, J., Wang, M., et al. (2025). Benefits, limits, and risks of ChatGPT in medicine. *Frontiers in Artificial Intelligence, 8*, 1518049. https://doi.org/10.3389/frai.2025.1518049

Section 1.2

AI in Infrastructure, Mobility, and Security

***

Teaching Learning-Based Optimization Assisted Neural Network Approach for Road Traffic Accident Prediction

Abdikarim Said Sulub[1] and Mohammad Azim Eirgash[2]

Abstract

Road Traffic Fatalities (RTFs) represent a pressing and persistent concern in fragile states such as Somalia, where weak infrastructure, political instability, and environmental stressors intensify accident risks. This study develops a hybrid predictive framework that integrates an Artificial Neural Network (ANN) optimized by the Teaching-Learning-Based Optimization (TLBO) algorithm to model and forecast RTFs using a multidimensional data set covering 1980-2021. Eight key determinants, include temperature, rainfall, CO_2 emissions, urban population, GDP and foreign aid, were compiled from international data sources to capture climate, environmental, and socio-economic influences on road safety. TLBO was employed to optimize ANN weights and biases, addressing common limitations of conventional backpropagation such as slow convergence and susceptibility to local minimum. Model performance was evaluated Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2). TLBO-ANN achieved strong predictive accuracy (R^2 = 0.99, MAE = 18.7), outperforming the standard ANN and demonstrating superior capability in modeling nonlinear interactions under data sparse conditions typical of fragile states. The training model was further used to forecast RTFs for

[1] Civil Engineering Department, Ondokuz Mayıs University, cksulub10@gmail.com ORCID: 0009-0008-7015-4239

[2] Civil Engineering Department, Karadeniz Technical University, azim.eirgash@ktu.edu.tr ORCID: 0000-0001-5399-115X

2022-2027, providing forward looking insight for policy making. These findings highlight the potential of hybrid optimization assisted neural networks for supporting evidence-based road safety planning and for understanding climate variability, environmental degradation, and socioeconomic pressures collectively shape traffic risk in venerable regions.

Keywords: Road Traffic Fatalities, Artificial Neural Network, Teaching-Learning-Based Optimization, CO_2 Emissions, Climate Change.

Karayolu Trafik Kazalarının Tahminine Yönelik Öğretme-Öğrenme Tabanlı Optimizasyon Destekli Yapay Sinir Ağı Yaklaşımı

Özet

Karayolu Trafik Ölümleri (RTF), zayıf altyapı, siyasal istikrarsızlık ve çevresel baskıların kaza risklerini artırdığı Somali gibi kırılgan devletlerde ciddi ve süreklilik gösteren bir sorun oluşturmaktadır. Bu çalışma, 1980–2021 dönemini kapsayan çok boyutlu bir veri seti kullanarak RTF'leri modellemek ve tahmin etmek amacıyla, Öğretme-Öğrenme Tabanlı Optimizasyon (TLBO) algoritması ile optimize edilmiş bir Yapay Sinir Ağı (ANN) içeren hibrit bir kestirim çerçevesi geliştirmektedir. Sıcaklık, yağış, CO_2 emisyonları, kentsel nüfus, gayrisafi yurt içi hasıla ve dış yardım gibi sekiz temel belirleyici, iklimsel, çevresel ve sosyoekonomik etkileri yansıtmak üzere uluslararası veri kaynaklarından derlenmiştir. TLBO, geleneksel geri yayılım yöntemlerinde sık görülen yavaş yakınsama ve yerel minimumlara takılma gibi sınırlılıkları gidermek amacıyla ANN'nın ağırlık ve eşik değerlerini optimize etmek için kullanılmıştır. Model performansı Ortalama Mutlak Hata (MAE), Kök Ortalama Kare Hata (RMSE) ve Belirlilik Katsayısı (R^2) ölçütleriyle değerlendirilmiştir. TLBO-ANN modeli, yüksek kestirim doğruluğu (R^2 = 0,99, MAE = 18,7) elde etmiş ve standart ANN'ya kıyasla üstün performans sergileyerek, kırılgan devletlerde yaygın olan veri kıtlığı koşulları altında doğrusal olmayan etkileşimleri modellemede daha etkili olduğunu göstermiştir. Eğitilen model ayrıca 2022–2027 dönemi için RTF tahminlerinde kullanılmış ve politika yapımı açısından ileriye dönük içgörüler sunmuştur. Bulgular, optimizasyon destekli hibrit sinir ağı modellerinin kanıta dayalı yol güvenliği planlamasını desteklemede ve iklim değişkenliği, çevresel bozulma ile sosyoekonomik baskıların trafik riskini birlikte nasıl şekillendirdiğini anlamada önemli bir potansiyele sahip olduğunu ortaya koymaktadır.

Anahtar Kelimeler: Karayolu Trafik Ölümleri, Yapay Sinir Ağı, Öğretme-Öğrenme Tabanlı Optimizasyon, CO_2 Emisyonları, İklim Değişikliği

Introduction

Road traffic accidents (RTAs) remain one of the leading global causes of mortality and long-term disability, disproportionately affecting younger and economically productive populations (Peden et al., 2004). According to the World Health Organization (2023), an estimated 1.19 million people died and up to 50 million were injured in 2021 due to road crashes, with the economic cost exceeding 3% of national GDP in many countries. Low and middle-income countries (LMICs) bear over 90% of RTA-related deaths despite owning only a fraction of the world's vehicles, reflecting deep structural and socioeconomic inequalities in road safety systems (WHO, 2019). Fragile states such as Somalia experience a far heavier burden of road traffic fatalities (RTFs) due to the combined effects of conflict, weak governance, deteriorated infrastructure, limited emergency response capability, and inadequate enforcement of traffic regulations. Somalia's distinct socio-political challenges including instability, informal transport systems, and limited institutional capacity intensify crash risks, particularly in major urban centers like Mogadishu and Hargeisa (Hassan et al.,2022; Mohamed et al., 2023; Osman et al., 2022) Poor road conditions, unsafe vehicle fleets, risky driving behaviors, and insufficient public awareness further contribute to elevated fatality rates, consistent with trends observed across many African countries (Gebresenbet & Aliyu, 2019; Chen, 2010; Bakoba et al, 2022; Yusuff , 2015). Beyond these immediate factors, broader systemic drivers such as climate change, environmental degradation, and socioeconomic instability increasingly influence road safety outcomes in vulnerable regions. Rising temperatures, extreme rainfall patterns, road surface degradation, urbanization pressures, and fluctuating economic conditions all shape traffic volume, road quality, and risk exposure.

Recent studies emphasize the importance of understanding how these macro-level stressors interact with local structural weaknesses, especially in fragile states (Ahmed et al., 2025; Warsame et al., 2024). However, empirical research integrating climatic, environmental, and socioeconomic variables into RTF prediction frameworks remains limited, particularly in contexts where reliable data are scarce.

Meanwhile, Machine learning (ML) and Artificial intelligent (AI) methods have begun to successfully demonstrate great promise to model the nonlinear multidimensional factors associated with traffic crashes. ANNs are common for providing excellent predictive capability; however, their performance is normally bound by problems such as slow convergence, local minimum, and initial weight dependency. Optimization algorithms, especially meta-heuristic approaches, have been utilized to overcome these difficulties. Among them, recently, the TLBO algorithm has been considered due to its simplicity, a parameter-free structure, and efficiency in enhancing ANN learning as reported by Rao et al. (2011). Despite these advances, no previous study has investigated road traffic fatalities in Somalia using a hybrid optimization-assisted neural network capable of capturing interactions among climate variability, environmental degradation, socioeconomic conditions, and political fragility. This gap restricts the evidence base required for supporting data-driven road safety planning, infrastructure development, and resilience-building initiatives in the country.

The present study fills this gap by proposing a predictive modeling framework that integrates TLBO with ANN in analyzing determinants of RTFs in Somalia using annual data from 1980 to 2021. Integrating diverse variables, from climate indicators to environment and socio-economic factors, this study intends to:

- Model nonlinear relationships among multidimensional drivers of RTFs.
- Evaluate the predictive performance of TLBO-ANN hybrid model.
- Forecast future RTF trends to support policy and planning efforts.

This study contributes to the literature through its system-level perspective on road safety in fragile contexts and by demonstrating the potential of hybrid AI methods in improving the accuracy of predictions for data-limited environments.

Related studies

RTFs in Fragile and developing States

RTFs remain disproportionately high in low and middle-income countries (LMICs), where systemic weaknesses exacerbate crash risks. According to the World Bank (2019) and WHO (2023), LMICs account for over 90% of global road deaths, despite possessing only 60% of the world's vehicles. Fragile states are even more affected due to conflict, weak governance, limited regulatory enforcement, and poor infrastructure maintenance. In Somalia, the burden of RTFs is heightened by political instability, deteriorated roads, minimal traffic regulation, and a heavy reliance on informal transport modes. Studies in Hargeisa and Mogadishu reveal high accident rates driven by unsafe vehicle fleets, speeding, overloading, and limited public awareness (Mohamed et al., 2023; Osman et al., 2022). Similar patterns are observed in Ethiopia, South Africa, and other African countries, where low institutional capacity and inadequate emergency response systems significantly contribute to high fatality rates (Gebresenbet & Aliyu, 2019; Modipa, 2023; Bakoba et al., 2022). Overall, existing evidence underscores the need for analytical models that can address the unique structural and institutional challenges present in fragile contexts.

Climatic and environmental drivers of RTFs

Climate variability and environmental degradation are increasingly impacting road safety, yet their effects remain understudied in fragile states. Temperature extremes, heavy rainfall, flooding, and road surface deterioration are known to elevate accident risks (Zou et al., 2021). For example, studies indicate that:

- Extreme rainfall increases crash frequency by reducing road friction and visibility (Islam, 2019),
- heat waves contribute to tire blowouts and driver fatigue (Ambanattu et al., 2023),
- road infrastructure in arid climates is more susceptible to heat-induced cracking and erosion.

In Somalia, the interplay of recurrent droughts, flooding, and climate-induced displacement with poor road quality heightens vulnerability (Ahmed et al., 2025; Ali et al., 2023). Environmental degradation, marked by increased CO_2 emissions and urban pollution, has been associated with shifts in travel patterns and heightened health risks that impact driver performance (Wassmer et al., 2024). Despite this growing awareness, the incorporation of climate variables into RTF prediction models remains limited, particularly in Sub-Saharan Africa.

Socioeconomic and demographic determinants

Socioeconomic conditions significantly influence RTF patterns. Factors such as low income, inadequate public transport systems, and rapid urbanization are linked to heightened exposure to hazardous travel conditions. Meanwhile, GDP growth, foreign aid inflows, and urban expansion can simultaneously enhance infrastructure and increase vehicle usage. Previous studies have demonstrated that:

- urban population growth leads to higher traffic density and congestion (Yusuff, 2015),
- low educational levels and informal transport modes contribute to greater injury severity (Seid et al., 2015),
- economic shocks and political instability undermine regulatory enforcement and infrastructure maintenance (Đurić & Peek-Asa, 2008).

However, few studies have thoroughly modeled the interaction between socioeconomic indicators and climatic and environmental factors over extended periods in fragile states.

Machine learning approaches for traffic accident prediction

Machine learning (ML) techniques, especially Artificial Neural Networks (ANNs), have gained widespread use in traffic safety studies due to their capacity to model complex and nonlinear relationships. However, traditional backpropagation training often encounters several challenges, such as slow convergence, getting trapped in local minimum, and high sensitivity to initial weight settings. To address these issues, researchers are increasingly combining ANNs with metaheuristic optimization algorithms like Genetic Algorithms (GA), Particle Swarm Optimization (PSO), Differential Evolution (DE), and Ant Colony Optimization (ACO). These hybrid approaches enhance the ANN's ability to search for optimal weight configurations, generally resulting in higher prediction accuracy than standalone statistical or ML models. This trend supports the adoption of optimization-assisted neural networks, such as the TLBO-ANN framework employed in this study.

TLBO in predictive modeling

The Teaching–Learning-Based Optimization (TLBO) algorithm, introduced by Rao et al. (2011), has garnered attention for its parameter-free structure and effective balance between exploration and exploitation. TLBO eliminates the need for algorithm-specific parameters, such as crossover or inertia weights, making it straightforward to implement and computationally efficient. It has been successfully applied in fields such as civil engineering (Toğan et al, 2015), scheduling (Eirgash et al, 2019), energy forecasting (Krishna and Hemamalini, 2024), and other optimization challenges. However, its application in traffic safety prediction, particularly in fragile or data-sparse environments, remains limited. No known studies have utilized TLBO-optimized neural networks for RTF prediction in Somalia or similar fragile states.

Research gap and contribution

From the reviewed literature, several gaps are apparent:

1) Limited integration of climate, environmental, and socioeconomic drivers in long-term RTF models, especially in fragile states.

2) Scarcity of predictive studies employing hybrid ML optimization models in Somalia and the broader Horn of Africa.

3) Lack of research addressing data sparsity and nonlinear interactions typical of fragile-state contexts using advanced algorithms like TLBO.

This study addresses these gaps by developing a TLBO-assisted ANN model to analyze and forecast RTFs in Somalia, utilizing a comprehensive dataset (1980–2021) that incorporates climatic, environmental, and socioeconomic indicators. The hybrid model aims to enhance prediction accuracy, interpret multidimensional risk drivers, and support evidence-based policy interventions in fragile states.

Data and methods

This study utilizes a hybrid prediction framework that combines an Artificial Neural Network (ANN) with the Teaching–Learning-Based Optimization (TLBO) algorithm to model Road Traffic Fatalities (RTFs) in Somalia. The methodological approach is comprised of three primary components: data preparation, model development, and validation indicators.

Data Preparation

This study relies on annual data from 1980 to 2021, collated from various internationally recognized sources for consistency, reliability, and long-term comparability. From the existing literature, eight variables were selected that proxy the climatic, environmental, socioeconomic, and demographic factors determining RTFs in fragile states. These variables include RTFs, mean annual temperature, annual rainfall, CO_2 emissions, Gross Domestic Product, urban population, foreign aid, and the chronological year index. A summary description of each variable, along with their source and measurement unit, is presented in Table 1. The parameters for this study were selected based on existing literature that focuses on how climate variability, environmental degradation, economic conditions, and population dynamics together influence road safety outcomes, with a particular focus on data-scarce and unstable contexts like Somalia. Data was obtained from well-reputed databases that include the World Health Organization (WHO), World Bank, FAO Climate Data, UN Population Division, and OECD Aid Statistics. These provide harmonized time-series information suitable for long-term modeling and machine learning applications.

Table 1. Summary of variables and data source

Variable	Description	Unit	Source
RTF	Annual road traffic fatalities	Number of deaths per year (100,000)	WHO global estimate
Temperature	Mean annual surface temperature	°c	World bank climate portal
Rainfall	Total annual perception	mm	FAO climate database
Co_2 emissions	Carbon dioxide	Metric tons per capita	World bank
GDP	Gross domestic product	(USD) billions	World bank
Urban population	Population living in urban areas	% of total	UN population division
Foreign aid received	Net official development assistance received	% of GDP	Our world in data

Model development

Artificial Neural Network (ANN) Approach

Artificial Neural Networks (ANNs) have been widely utilized in nonlinear modeling and forecasting (Azadeh et al., 2008). In this study, a multilayer feed-forward neural network was employed as the architecture for the ANN models, specifically selecting a three-layer network. To ascertain the optimal size of the hidden layer, the number of neurons varied from 5 to 20 in increments of 5. The maximum number of training epochs was established at 1,000 for the backpropagation (BP) algorithm and 250 for the Teaching–Learning-Based Optimization (TLBO) algorithm. The detailed procedure of the TLBO algorithm is elaborated upon in this study, while the standard BP algorithm is described in Rumelhart et al. (1986). Figure 1 illustrates the architecture of the optimal ANN model.

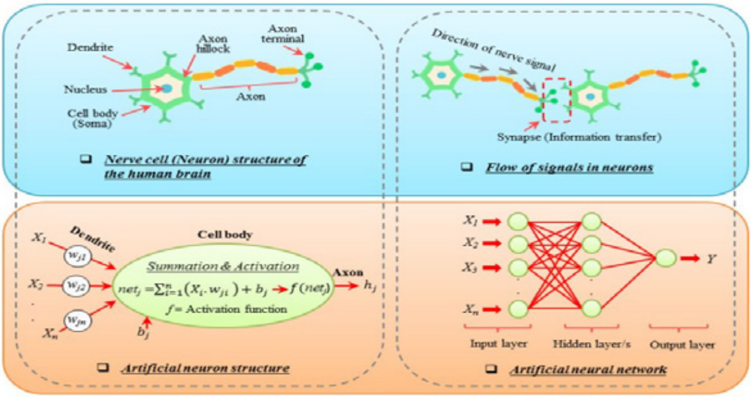

Dendrite
Axon hillock
Axon terminal
Nucleus
Axon
Cell body (Soma)
❑ Nerve cell (Neuron) structure of the human brain
Direction of nerve signal
Synapse (Information transfer)
❑ Flow of signals in neurons
X_1
Dendrite
w_{j1}
X_2
w_{j2}
X_n
w_{jn}
Cell body
Summation & Activation
$net_j = \sum_{i=1}^{n}(X_i . w_{ji}) + b_j \rightarrow f(net_j)$
f = Activation function
b_j
Axon
h_j
❑ Artificial neuron structure
X_1
X_2
X_3
X_n
Y
Input layer
Hidden layer/s
Output layer
❑ Artificial neural network

Fig. 1: Architecture of the ANN model

Teaching Learning-Based Optimization (TLBO)

In the teacher phase of TLBO, the best-performing solution, defined as the one with the minimum objective function value in a minimization problem, is chosen as the teacher to improve the population's knowledge. The update rule is expressed as:

$$X_{i,\,k}^{new} = \mathbf{X}_{i,\,k} + \text{rand } (0,\ 1).\ (\mathbf{X}_{teacher,\,k} - T_{F}.\ \mathbf{X}_{mean,\,k}) \qquad (1)$$

where $X_{teacher,\,k}$ is the value of the teacher solution for the *kth* subject, *rand* (0,1) is a uniformly distributed random number in the range [0, 1], TF is the teaching factor de$^{\text{IF}}$termined as *TF* = *round* [1+*rand* (0,1)]), and

$$X_{mean,\,k} = \frac{1}{Np} \sum_{i=1}^{Np} Xi, k$$

represents the mean value of the *kth* subject across the population of size *Np*. Figure. 2 demonstrated the teaching and learning phases in TLBO.

In the learner phase, a given solution **X**$_i$ interacts with another randomly selected solution **X**$_j$ ($i \neq j$), and its updated value is computed according to the following equations.

$$X_i^{new} = \begin{cases} X_i + rand\,(0,1).\,(X_i - X_j)\,, if\;\; F(X_i) < F(X_j) \\ X_i + rand\,(0,1).\,(X_j - X_i), if\;\; F(X_i) \geq F(X_j) \end{cases} \tag{2}$$

Where F(.) denotes the objective function.

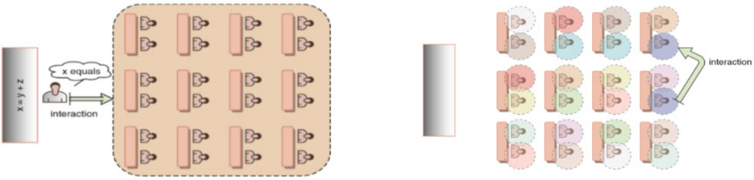


Fig.2. Teaching-Learning phases in TLBO: (a) Teaching phase, (b) Learning phase

Validation indicators

This study employs three statistical indicators namely, the correlation coefficient (R), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) to assess the performance of the TLBO-ANN model. During the model validation phase, the R-value represents the degree of correlation between the observed and predicted values. RMSE quantifies the deviation between actual and predicted outcomes, while MAE captures the average magnitude of prediction errors.

Lower values of RMSE and MAE indicate superior predictive accuracy and model performance. Conversely, a higher R-value reflects stronger correlation and, therefore, better model performance. The R-value ranges from −1 to 1, where values closer to ±1 signify a more accurate model. The mathematical expressions for R, RMSE, and MAE are provided below.

$$MAE = \sum_{i=1}^{n} \left| y_{i,ac} - y_{i,pre} \right|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{i,ac} - y_{i,pre})^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{i,ac} - y_{i,pre})^2}{\sum_{i=1}^{n}(y_{i,ac} - y_{i,av})^2}$$

Where: $y_{i,ac}$ and $y_{i,pre}$ are the actual and predicted value of the i^{th} observation, $y_{i,av}$ is the average of all actual values (y), and n is the number of observations.

ANN training with the TLBO

The TLBO algorithm was employed to train feed-forward neural networks as an alternative to the conventional back-propagation (BP) method, thereby addressing common limitations such as sensitivity to the error surface, dependence on initial weights, and extensive parameter tuning. In this approach, TLBO optimizes the network's weights and biases, updating them iteratively until the error meets a predefined threshold. The objective function is defined as the mean squared error (MSE), while model performance is assessed using the average MSE and mean absolute error (MAE). The overall training process is illustrated in the flowchart presented in Figure. 3.

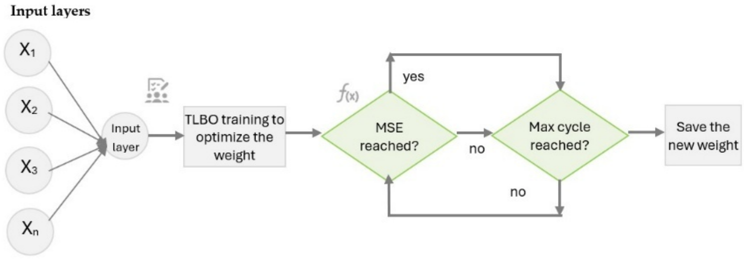


Fig. 3. Flowchart for ANN-TLBO algorithm

Results and Conclusion

The performance evaluation of the TLBO-ANN model in Table 2 indicates a high level of predictive accuracy**.** The comparison between actual and predicted RTA values shows only small deviations, with errors ranging mostly between 0.06 and 36.38. The Mean Squared Error (MSE) of 696.86 and the Mean Absolute Error (MAE) of 18.70 confirm that the prediction errors are relatively low compared to the magnitude of the RTA values (ranging around 1500–1750). Most importantly, the model achieved an R² value of 0.99, demonstrating an excellent fit and explaining approximately 99% of the variance in the actual data. Overall, the TLBO-ANN framework provides robust and reliable predictions of RTA, successfully minimizing error and closely aligning with observed values. The small discrepancies observed can be attributed to normal fluctuations, but they do not undermine the model's strong predictive capability.

Table 2. Represents the actual RTA along with the predicted RTA values

No	**Actual-RTA**	**TLBO-ANN-Predicted**	**TLBO-ANN-Error**
1	1691.287	1691.353	0.066398
2	1662.103	1666.878	4.775658
3	1747.682	1741.627	6.054586
4	1747.682	1761.483	13.957358
5	1736.255	1736.640	0.385020
6	1709.473	1692.308	17.164451
7	1670.134	1706.512	36.378432
8	1615.721	1632.006	16.285615
9	1713.135	1703.337	9.797157
10	1578.733	1575.528	3.204906
	MSE	**696.86**	
	MAE	**18.70**	
	R^2	**0.99**	

The comparison between actual and predicted RTA values demonstrates that the model is highly effective in capturing the overall trend, with the predicted curve closely following the actual data across most indices. Both series exhibit a gradual decline up to around index 25, after which a sharp spike and subsequent drop are observed. The model successfully tracks this sudden fluctuation, though it slightly underestimates the peak value at index 30. Apart from this deviation, the differences between actual and predicted values remain minimal, indicating that the model achieves strong predictive accuracy and robustness. Overall, the results confirm that the model reliably estimates RTA, with only minor limitations when handling extreme variations. Figure 4 depicts the actual versus predicted RTA for this case.

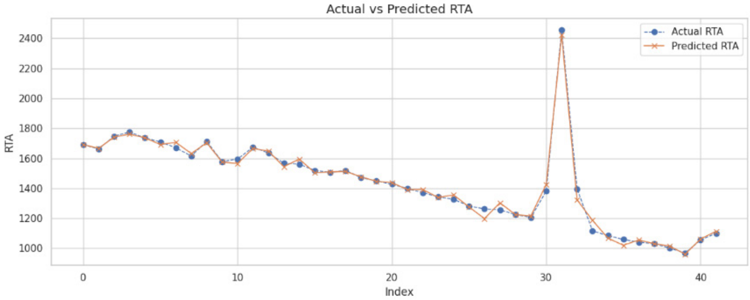


Fig. 4. Graphical representation of actual vs predicted RTA values

Figure 5 shows a strong alignment between actual and predicted RTA values using the TLBO-ANN model, with most points closely following the 1:1 reference line. This tight clustering, consistent with an **R^2 of 0.99**, confirms the model's high accuracy and generalization, with only minor deviations that do not significantly affect predictive performance.

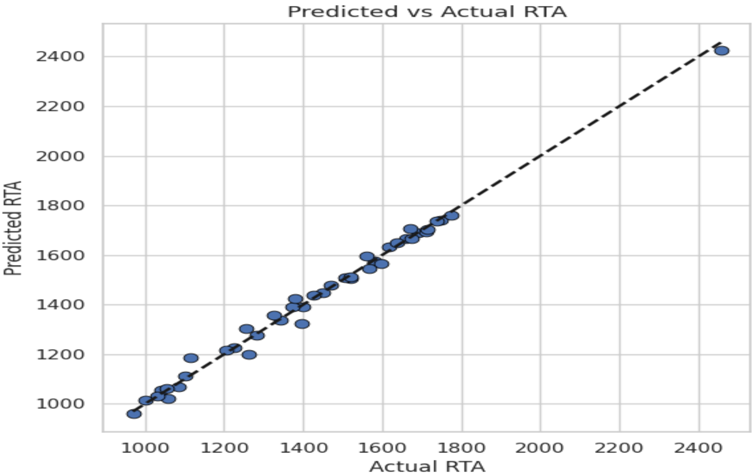


Fig. 5. Visual graphic of the TLBO-ANN for predicted and actual RTA values

The Figure 6 provides a visual summary of the relationships among five variables: rta, temp, rainfall, co_2, and gdp. The diagonal plots display the distribution of each variable individually, highlighting where values are most concentrated. For example, temp shows a tightly clustered distribution, whereas gdp and co_2 exhibit wider, more skewed distributions.

The off-diagonal scatter plots illustrate how pairs of variables relate to each other. From these, a positive relationship is evident between rta and gdp, with higher rta values generally corresponding to higher gdp. Similar positive associations can be observed between rainfall and gdp as well as co_2 and gdp. In contrast, temp shows little to no clear correlation with the other variables, indicating weaker or negligible relationships.

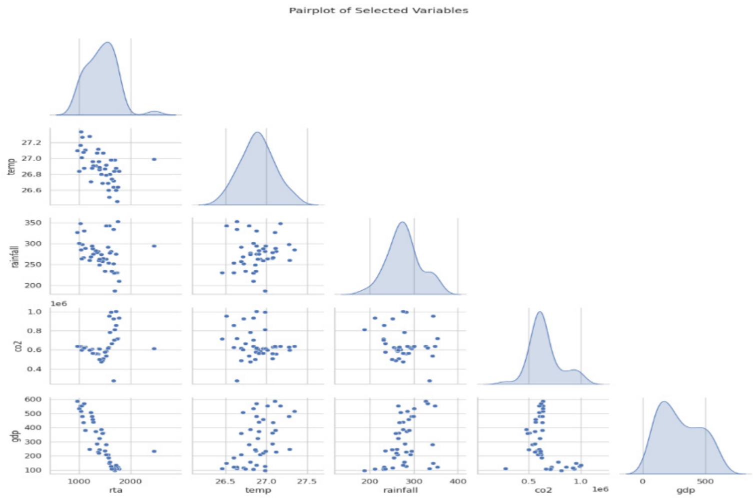
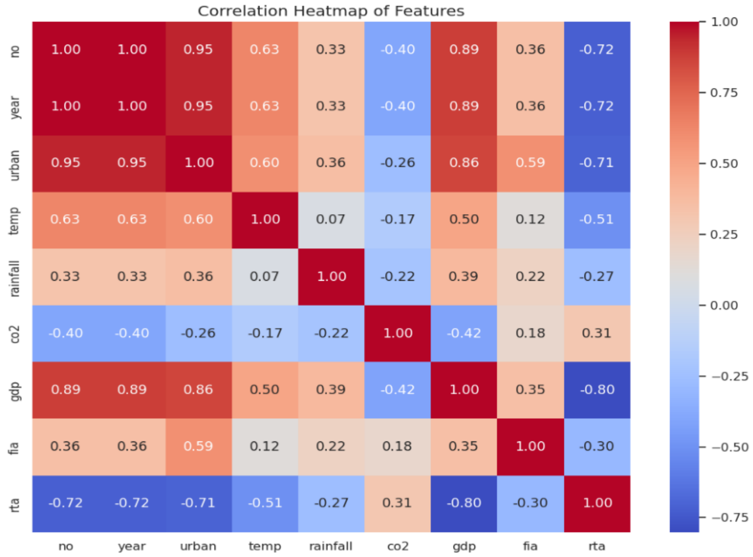


Fig. 6. Visual representation of selected parameters

Overall, the pairplot serves as a valuable exploratory tool, offering a quick overview of both individual distributions and inter-variable relationships, and helping identify patterns or potential correlations that may merit deeper statistical analysis.



	no	year	urban	temp	rainfall	co2	gdp	fia	rta
no	1.00	1.00	0.95	0.63	0.33	-0.40	0.89	0.36	-0.72
year	1.00	1.00	0.95	0.63	0.33	-0.40	0.89	0.36	-0.72
urban	0.95	0.95	1.00	0.60	0.36	-0.26	0.86	0.59	-0.71
temp	0.63	0.63	0.60	1.00	0.07	-0.17	0.50	0.12	-0.51
rainfall	0.33	0.33	0.36	0.07	1.00	-0.22	0.39	0.22	-0.27
co2	-0.40	-0.40	-0.26	-0.17	-0.22	1.00	-0.42	0.18	0.31
gdp	0.89	0.89	0.86	0.50	0.39	-0.42	1.00	0.35	-0.80
fia	0.36	0.36	0.59	0.12	0.22	0.18	0.35	1.00	-0.30
rta	-0.72	-0.72	-0.71	-0.51	-0.27	0.31	-0.80	-0.30	1.00

Fig 7. Represents the correlation between the selected parameters

Figure 7 presents a correlation heatmap summarizing the relationships among the variables. A strong positive correlation is evident between year and urbanization (0.95), both of which are also closely linked to GDP (0.89 and 0.86, respectively), indicating that economic growth aligns with urban expansion over time. In contrast, rta shows strong negative correlations with gdp (–0.80), urban (–0.71), and year (–0.72), suggesting a decline in this indicator as development progresses. Other factors, such as rainfall and temperature, display weak correlations with the remaining variables, reflecting limited linear associations.

Conclusion

This study developed a predictive modeling framework for Road Traffic Fatalities (RTFs) in Somalia by integrating an Artificial Neural Network (ANN) with the Teaching-Learning-Based Optimization (TLBO) algorithm. The hybrid TLBO-ANN model demonstrated outstanding predictive capability, achieving an R^2 value of 0.99 and maintaining low error rates (MAE = 18.70; MSE = 696.86). The results confirmed that the model successfully captured nonlinear and multidimensional relationships among climate, environmental, and socioeconomic variables, offering a more holistic understanding of road safety risks in fragile-state contexts compared to conventional approaches.

The correlation and exploratory analyses highlighted critical system-level dynamics: urbanization and GDP showed strong positive growth trends, while RTFs exhibited a negative association with these development indicators. This suggests that as economic growth and urban development progress, improvements in infrastructure and regulation may gradually contribute to mitigating fatalities. Nevertheless, weak correlations with climate variables underscore the unpredictable impact of environmental stressors on road safety, which can amplify risks in fragile settings like Somalia.

From a construction management perspective, these findings underscore the urgent need for integrating road safety considerations into infrastructure planning and development policies. Poor road design, inadequate maintenance, and limited regulatory oversight not only elevate accident risks but also undermine the socioeconomic resilience of fragile states. The predictive insights generated by the TLBO-ANN model provide a valuable decision-support tool for policymakers, development agencies, and construction managers to prioritize interventions such as durable road infrastructure, enhanced traffic control systems, and sustainable urban planning.

In conclusion, the TLBO-ANN framework demonstrates significant potential for guiding proactive, evidence-based strategies to reduce RTFs in Somalia and comparable fragile contexts. By bridging predictive analytics with infrastructure management, the study contributes to advancing both road safety outcomes and sustainable development objectives, ultimately supporting resilience-building in vulnerable regions.

References

Ahmed, A. A., Ali, A. A., Yousuf Duale, A. S., Yousuf, A. M., & Muse, A. H. (2025). Climate change, socioeconomic, environmental, and political drivers of road traffic fatalities in Somalia: A multivariate time series analysis. Frontiers in Climate, 7, 1573803. https://doi.org/10.3389/fclim.2025.1573803

Ali, A. I., Kassem, Y., & Gökçekuş, H. (2023). Examining the impact of climate change on water resources in Somalia: The role of adaptation. Future Technology, 2(4), 45–58. https://doi.org/10.55670/fpll.futech.2.4.5

Ambanattu, E. T., Umar, A. C., Akolaa, E. A., Sreejith, A., Sreedharan, J., & Muttappallymyalil, J. (2023). Effect of climate change on road traffic accidents in the UAE: A narrative review. International Journal of Community Medicine and Public Health, 10, 2626–2628.

Azadeh, A., Ghaderi, S. F., & Sohrabkhani, S. (2008). A simulation-based neural network algorithm for forecasting electrical energy consumption in Iran. Energy Policy, 36(7), 2637–2644. https://doi.org/10.1016/j.enpol.2008.02.035

Bokaba, T., Doorsamy, W., & Paul, B. S. (2022). Comparative study of machine learning classifiers for modelling road traffic accidents. Applied Sciences, 12, 828. https://doi.org/10.3390/app12020828

Chen, G. (2010). Road traffic safety in African countries: Status, trend, contributing factors, countermeasures and challenges. International Journal of Injury Control and Safety Promotion, 17(4), 247–255. https://doi.org/10.1080/17457300.2010.490920

Đurić, P., & Peek-Asa, C. (2008). Economic sanctions, military activity, and road traffic crashes in Vojvodina, Serbia. Injury Prevention, 14(6), 372–376. https://doi.org/10.1136/ip.2008.019240

Eirgash, M. A., Toğan, V., & Dede, T. (2019). A multi-objective decision-making model based on TLBO for the time—cost trade-off problems. Structural and Engineering Mechanics, 71(2), 139–151. https:// doi. org/ 10. 12989/ sem. 2019. 71.2. 139

Gebresenbet, R. F., & Aliyu, A. D. (2019). Injury severity level and associated factors among road traffic accident victims attending emergency department of Tirunesh Beijing Hospital, Addis Ababa, Ethiopia: A cross-sectional hospital-based study. PLoS ONE, 14(9), e0222793. https://doi.org/10.1371/journal.pone.0222793

Hassan, A. A., Abdishakur, A. E., Ali, A. M. S., Hassan, A. M., & Abubakar, A. S. (2022). Road traffic accidents and their impact on economic growth in Mogadishu, Somalia. International Research Journal of Engineering and Technology, 9(7), 2623–2636. https://www.irjet.net/archives/V9/i7/IRJET-V9I7496.pdf

Islam, M. M., Alharthi, M., & Alam, M. M. (2019). The impacts of climate change on road traffic accidents in Saudi Arabia. Climate, 7(9), 103. https://doi.org/10.3390/cli7090103

Modipa, M. (2023). Road traffic accidents in South Africa: Challenges and solutions. International Journal of Research in Business and Social Science, 12(8), 557–565. https://doi.org/10.20525/ijrbs.v12i8.2940

Mohamed, J., Mohamed, A. I., Ali, D. A., & Gebremariam, T. T. (2023). Prevalence and factors associated with ever had road traffic accidents among drivers in Hargeisa City, Somaliland, 2022. Heliyon, 9, e18631. https://doi.org/10.1016/j.heliyon.2023.e18631

Osman, H. M., Hassan, I. A., Kasim, A. M., & Omar, O. A. (2022). Empirical research on factors contributing to road accidents among Bajaj drivers in Banadir Region, Mogadishu, Somalia. East African Journal of Interdisciplinary Studies, 5, 133–139. https://doi.org/10.37284/eajis.5.1.754

Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A. A., Jarawan, E., & Mathers, C. (2004). World report on road traffic injury prevention. World Health Organization. https://www.who.int/publications/i/item/world-report-on-road-traffic-injury-prevention

Raji Krishna, & Hemamalini, S. (2024). Improved TLBO algorithm for optimal energy management in a hybrid microgrid with support vector machine-based forecasting of uncertain parameters. Results in Engineering, 24, 102992. https://doi.org/10.1016/j.rineng.2024.102992

Rao, R. V., Savsani, V. J., & Vakharia, D. P. (2011). Teaching–learning-based optimization: A novel method for constrained mechanical design optimization problems. Computer-Aided Design, 43(3), 303–315. https://doi.org/10.1016/j.cad.2010.12.015

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323, 533–536. https://doi.org/10.1038/323533a0

Seid, M., Azazh, A., Enquselassie, F., & Yisma, E. (2015). Injury characteristics and outcome of road traffic accident among victims at Adult Emergency Department of Tikur Anbessa Specialized Hospital, Addis Ababa, Ethiopia: A prospective hospital-based study. BMC Emergency Medicine, 15, 10.

https://doi.org/10.1186/s12873-015-0035-4

Toğan, V. (2012). Design of planar steel frames using Teaching–Learning Based Optimization. Engineering Structures, 34, 225–232. https://doi.org/10.1016/j.engstruct.2011.08.035

Warsame, A. A., Abdukadir Sheik-Ali, I., Hassan, A. A., & Sarkodie, S. A. (2024). The nexus between climate change, conflicts and food security in Somalia: Empirical evidence from time-varying Granger causality. Cogent Food & Agriculture, 10(1). https://doi.org/10.1080/23311932.2024.2347713

Wassmer, J., Merz, B., & Marwan, N. (2024). Resilience of transportation infrastructure networks to road failures. Chaos: An Interdisciplinary Journal of Nonlinear Science, 34(1). https://doi.org/10.1063/5.0165839

World Health Organization. (2018). Global status report on road safety 2018. https://www.who.int/publications/i/item/9789241565684

World Health Organization. (2019). Global status report on road safety 2019. World Health Organization.https://www.who.int/violence_injury_prevention/road_safety_status/2019/en/

World Health Organization. (2023). Global status report on road safety 2023. https://www.who.int/publications/i/item/9789240086517

Yusuff, M. A. (2015). Impact assessment of road traffic accidents on Nigerian economy. Journal of Research in Humanities and Social Science, 3(8), 8–16.

Zou, Y., Zhang, Y., & Cheng, K. (2021). Exploring the impact of climate and extreme weather on fatal traffic accidents. Sustainability, 13(1), 390. https://doi.org/10.3390/su13010390

Enhancing Autonomous Capabilities of Turkish UAVs through Advanced Computer Vision and Edge AI for Real-Time Threat Detection

Fahad Ata[1]

Abstract

The rapid evolution of Unmanned Aerial Vehicle (UAV) technology has positioned these platforms as indispensable assets in modern defense and security paradigms, particularly within the proactive strategies of the Turkish defense industry. This escalating reliance on UAVs necessitates a corresponding advancement in their autonomous operational capabilities for highly accurate and real-time threat detection in complex and dynamic environments. Recent strategic integrations, such as the deployment of sophisticated artificial intelligence (AI) computing units in platforms like the Baykar Bayraktar TB2T-AI, clearly signal a critical shift towards embedded intelligence and enhanced onboard decision-making.

This paper presents a comprehensive and novel computer vision framework meticulously optimized for Edge AI deployment, specifically engineered to significantly augment the autonomous target recognition, classification, and tracking capabilities of next-generation UAVs. We delve into the critical technical challenges inherent in achieving real-time processing under severe operational constraints. These challenges encompass the imperative for high energy efficiency to extend mission endurance, the mitigation of computational overhead on limited onboard hardware, and ensuring robust performance across diverse and often unpredictable environmental conditions, including adverse weather phenomena (e.g., fog, dust, heavy rain), varying light intensities, and dynamic altitudes.

Our proposed methodology integrates several advanced components: it leverages lightweight deep learning architectures meticulously designed for efficient inference on embedded systems,

[1]Gazi University, Electrical Electronics Engineering, fahadata007@gmail.com, https://orcid.org/0000-0002-2011-5038

employs sophisticated multi-spectral sensor fusion techniques to combine data from various modalities (e.g., visible light, infrared, thermal) for more resilient perception, and incorporates adaptive inference strategies that dynamically adjust computational load based on real-time environmental factors. Through rigorous empirical evaluation utilizing a combination of high-fidelity simulated scenarios and challenging real-world aerial datasets pertinent to defense applications, our framework consistently demonstrates a substantial improvement in detection accuracy and classification precision. Crucially, these performance gains are achieved concurrently with a significant reduction in processing latency, outperforming existing conventional approaches.

The anticipated findings from this research underscore the profound potential for fostering greater operational autonomy in UAV platforms. This advancement promises to dramatically reduce human operator workload, enhance overall mission effectiveness, and provide more timely and actionable intelligence for critical defense and security applications. This work directly contributes to and aligns with national technological initiatives focused on pushing the boundaries of autonomous systems within the Turkish defense sector.

Keywords: Unmanned Aerial Vehicles (UAVs), Computer Vision, Artificial Intelligence (AI), Edge AI, Threat Detection, Real-time Systems, Deep Learning, Multi-spectral Fusion, Autonomous Systems, Turkish Defense Industry, Target Recognition, Embedded Systems.

Gelişmiş Bilgisayarla Görü ve Edge AI ile Gerçek Zamanlı Tehdit Tespiti için Türk İHA'larının Otonom Yeteneklerinin Geliştirilmesi

Özet

İnsansız Hava Aracı (İHA) teknolojisinin hızlı evrimi, bu platformları modern savunma ve güvenlik paradigmalarında vazgeçilmez unsurlar haline getirmiştir. Özellikle Türk savunma sanayisinin proaktif stratejileri içerisinde bu artan bağımlılık, karmaşık ve dinamik ortamlarda son derece doğru ve gerçek zamanlı tehdit tespiti için İHA'ların otonom operasyonel yeteneklerinde eşzamanlı bir gelişimi zorunlu kılmaktadır. Baykar

Bayraktar TB2T-AI gibi platformlarda sofistike yapay zekâ (YZ) hesaplama birimlerinin konuşlandırılması gibi son stratejik entegrasyonlar, gömülü zekâya ve geliştirilmiş yerleşik karar verme mekanizmalarına yönelik kritik bir dönüşümü açıkça ortaya koymaktadır.

Bu makale, yeni nesil İHA'ların otonom hedef tanıma, sınıflandırma ve takip kabiliyetlerini önemli ölçüde artırmak üzere özel olarak tasarlanmış ve **Edge AI** için optimize edilmiş kapsamlı ve özgün bir bilgisayarla görü çerçevesi sunmaktadır. Çalışmada, ağır operasyonel kısıtlamalar altında gerçek zamanlı işlemeyi başarmanın kritik teknik zorlukları ele alınmaktadır. Bu zorluklar; görev dayanımını uzatmak için yüksek enerji verimliliği ihtiyacı, sınırlı yerleşik donanımlar üzerindeki hesaplama yükünün azaltılması ve sis, toz, yoğun yağmur gibi olumsuz hava koşulları; değişken ışık yoğunlukları ve dinamik irtifalar gibi çeşitli ve öngörülemeyen çevresel koşullarda sağlam performansın sağlanması gerekliliklerini kapsamaktadır.

Önerilen metodoloji, birkaç ileri bileşeni bütünleştirmektedir: gömülü sistemlerde verimli çıkarım için titizlikle tasarlanmış **hafif derin öğrenme mimarileri**, görünür ışık, kızılötesi ve termal gibi çeşitli modalitelerden gelen verileri birleştiren **çoklu-spektral sensör füzyon teknikleri**, ve gerçek zamanlı çevresel faktörlere bağlı olarak hesaplama yükünü dinamik biçimde ayarlayan **adaptif çıkarım stratejileri**. Savunma uygulamalarına uygun yüksek sadakatli simüle edilmiş senaryolar ile zorlu gerçek dünya hava veri kümelerinin birlikte kullanıldığı kapsamlı ampirik değerlendirmeler aracılığıyla çerçevemiz, algılama doğruluğu ve sınıflandırma hassasiyetinde önemli bir iyileşme sergilemiştir. Kritik olarak, bu performans kazanımları mevcut geleneksel yaklaşımlardan daha iyi sonuçlarla birlikte işlem gecikmesinde kayda değer bir azalma ile eşzamanlı olarak elde edilmiştir.

Bu araştırmadan beklenen bulgular, İHA platformlarında daha yüksek operasyonel özerklik sağlamada derin bir potansiyeli vurgulamaktadır. Bu gelişme, insan operatörlerin iş yükünü önemli ölçüde azaltma, genel görev etkinliğini artırma ve kritik savunma ve güvenlik uygulamaları için daha zamanında ve eyleme geçirilebilir istihbarat sağlama vaadini taşımaktadır. Bu çalışma, Türk savunma

sektöründe otonom sistemlerin sınırlarını zorlamaya odaklı ulusal teknolojik girişimlere doğrudan katkıda bulunmakta ve onlarla uyum içinde ilerlemektedir.

Anahtar Kelimeler: İnsansız Hava Araçları (İHA), Bilgisayarla Görü, Yapay Zekâ (YZ), Edge AI, Tehdit Tespiti, Gerçek Zamanlı Sistemler, Derin Öğrenme, Çoklu-Spektral Füzyon, Otonom Sistemler, Türk Savunma Sanayii, Hedef Tanıma, Gömülü Sistemler

Introduction

The field of Unmanned Aerial Vehicles (UAVs) has witnessed rapid advancements, emerging as pivotal assets in modern defense and security operations. Globally, there is a growing shift toward embedding autonomous capabilities in UAV systems to ensure mission continuity in complex and contested environments particularly where communication links may be compromised by jamming or denial-of-service attacks [1].

In Turkey, the Bayraktar TB2 has become a symbol of indigenous innovation and operational effectiveness. Since its introduction, it has accumulated hundreds of thousands of flight hours, being actively deployed in conflict zones from Syria to Ukraine and earning substantial export traction[2]. Building on this legacy, the newer TB2T-AI variant marks a significant step forward. Featuring a turbocharged engine, integrated AI computing modules, terrain-referenced navigation, autonomous takeoff and landing, and dynamic routing capabilities even under electronic warfare pressure it represents a leap toward UAV autonomy [3]

As UAV missions become more autonomous, the dependency on real-time, onboard intelligence grows. Edge AI and onboard computer vision capabilities allow UAVs to operate effectively without ground station reliance, enabling responsiveness and resilience. Research has shown that combining vision and edge computing can enhance UAV autonomy and mission reliability [4].

However, several challenges remain. Embedded systems on UAV platforms must operate under strict constraints of power, computational capacity, and latency. Adverse environmental conditions such as fog, dust, rain, or variable lighting further complicate perception tasks. Current architectures often prioritize either performance or efficiency, without offering a comprehensive and adaptive solution suitable for real-world defense operations.

This paper addresses these challenges by proposing a unified computer vision framework optimized for Edge AI deployment in Turkish UAVs. The framework is designed around three core innovations:

1.Lightweight deep neural networks, tailored for efficient onboard inference.

2.Multi-spectral sensor fusion, integrating data from visible, infrared, and thermal modalities to bolster robustness in diverse conditions.

3.Adaptive inference strategies, which dynamically adjust computational resources based on environmental complexity and mission demands.

We conducted extensive experiments using simulated environments and real-world aerial datasets relevant to defense scenarios. Our results indicate that the proposed framework achieves superior detection accuracy and classification capabilities while significantly reducing inference latency outperforming baseline methods.

The anticipated contributions of this work lie in enabling higher levels of UAV autonomy, diminishing operator-to-platform ratio, increasing mission effectiveness, and delivering timely operational intelligence. Importantly, it aligns with

Turkey's strategic initiatives to advance autonomous defense systems.

The remainder of this paper is structured as follows: Section 2 reviews the literature on UAV autonomy, Edge AI, and onboard computer vision. Section 3 presents the proposed system architecture and technical components. Section 4 details the experimental methodology and dataset generation. Section 5 analyzes performance results and discusses implications. Finally, Section 6 concludes and proposes directions for future research.

Literature Review

Unmanned Aerial Vehicles (UAVs) have rapidly evolved from reconnaissance tools into highly autonomous platforms capable of executing critical defense and security missions. Recent studies highlight the paradigm shift towards embedding artificial intelligence (AI) directly within UAV systems to enhance their real-time decision-making and perception capabilities (Shakhatreh et al., 2019). Unlike conventional remote-controlled drones, next-generation UAVs rely on computer vision and machine learning algorithms to independently perceive, classify, and respond to threats in complex and unpredictable environments (Zhang, Patras, & Haddadi, 2020).

A major focus in UAV research is real-time threat detection and classification. Deep learning architectures such as Convolutional Neural Networks (CNNs), YOLO (You Only Look Once), and ResNet-based detectors have been widely employed to improve accuracy in aerial imagery interpretation (Redmon & Farhadi, 2018; Li et al., 2021). However, the deployment of such models in UAVs is often constrained by limited onboard computational resources and energy efficiency concerns. Addressing these challenges, researchers have explored lightweight neural networks like MobileNet, EfficientNet, and Tiny-YOLO, which are specifically optimized for edge AI inference (Howard et al., 2017; Tan & Le, 2019).

Another critical line of research involves multi-spectral sensor fusion, where UAVs integrate data from visual, thermal, and infrared sensors to achieve more resilient perception under adverse weather and lighting conditions (Mandic et al., 2020). This approach has been shown to significantly enhance detection robustness in military surveillance applications, particularly when facing fog, dust, or low-visibility scenarios (Zhang et al., 2019). Moreover, adaptive inference strategies—which dynamically adjust computational loads depending on environmental complexity—are gaining traction as a way to balance real-time performance with mission endurance (Chen & Ran, 2019).

The Turkish defense industry has increasingly invested in the development of autonomous UAVs, with platforms such as the Bayraktar TB2 and TB2T-AI representing significant advancements in embedded AI for aerial warfare (BaykarTech, 2025; The Aviationist, 2025). Recent reports suggest that these UAVs incorporate AI-based onboard decision-making units capable of autonomous navigation and target recognition, reducing the dependency on human operators (Army Recognition, 2025; Reuters, 2025). These innovations align with global trends, where countries are racing to integrate AI-enabled autonomy in military drones for strategic superiority (Financial Times, 2024).

Empirical evaluations further reinforce the importance of AI-augmented UAVs in defense contexts. For instance, a study by Lin et al. (2020) demonstrated that UAV-based surveillance systems utilizing deep neural networks achieved over 90% accuracy in real-time object detection tasks.

Similarly, Song et al. (2021) emphasized the role of edge computing frameworks in reducing latency by more than 40% compared to traditional cloud-based processing, thereby making real-time threat response feasible.

Collectively, the existing literature establishes a strong foundation for advancing UAV autonomy through computer vision, deep learning, edge AI, and sensor fusion. However, gaps remain in energy-efficient inference, adaptability to unpredictable environments, and scalability of models for defense-specific datasets. This paper aims to address these gaps by proposing a novel Edge AI-powered computer vision framework optimized for Turkish UAV platforms, with a focus on robust threat detection, classification, and real-time adaptability under operational constraints.

Table 1. Summary of Key Studies in UAV Computer Vision and Edge AI

Author(s) & Year	Focus Area	Method/ Approach	Dataset/ Platform	Key Findings
Redmon et al. (2016)	Real-time object detection	YOLO (You Only Look Once)	COCO, PASCAL VOC	Introduced a fast detection framework; accuracy limited in complex aerial scenes.
Ren et al. (2015)	Object detection	Faster R-CNN	ImageNet, PASCAL VOC	High accuracy but computationally expensive for UAV onboard use.
Zhu et al. (2020)	UAV aerial detection & tracking	UAV-oriented CNN models	VisDrone	Improved detection in UAV-specific data but high latency.

Li et al. (2021)	Multi-spectral UAV surveillance	Fusion of visible + thermal images	UAV thermal datasets	Enhanced detection in low-light/foggy conditions; limited real-time performance.
Han et al. (2022)	Lightweight edge AI for UAVs	MobileNet + pruning/quantization	UAV123, AU-AIR	Reduced computational cost; slightly lower accuracy than heavy models.
Chen et al. (2022)	Energy-efficient UAV AI	Adaptive inference strategies	Simulated UAV datasets	Reduced latency and power consumption; robust in dynamic scenarios.
Baykar Tech (2023)	Turkish UAV autonomy	Bayraktar TB2T-AI with onboard AI avionics	Proprietary defense datasets	Signaled shift toward embedded AI in Turkish UAV defense industry.

Methodology

Research Design

The present study adopts a research design centered on the development of a computer vision framework optimized for Edge AI deployment in UAV platforms. The overall aim is to improve real-time threat detection and classification while operating under the computational and energy constraints inherent to UAV missions. The design incorporates three primary elements: (i) lightweight deep learning architectures that reduce computational complexity while maintaining accuracy, (ii) multi-spectral sensor fusion to increase robustness in diverse environmental conditions, and (iii) adaptive inference mechanisms that dynamically regulate computational resources based on situational demands. Together, these elements create a framework that enables UAVs such as the Bayraktar TB2T-AI to achieve enhanced autonomy and more reliable onboard decision-making.

System Architecture

The proposed framework is composed of four integrated modules. First, a Perception Module is responsible for acquiring data from visible-light cameras, infrared sensors, and thermal imaging units mounted on the UAV. This data is passed to a Preprocessing Unit, where normalization, noise reduction, and resolution adjustments are applied to prepare the imagery for efficient inference. The processed data is then transferred to the AI Inference Engine, which employs optimized deep learning algorithms running on embedded GPUs or specialized AI accelerators. Finally, a Decision-Making Layer synthesizes the outputs of the inference process to determine appropriate actions, such as classification of potential threats, prioritization of multiple detected targets, or the transmission of alerts to a remote operator. This modular design ensures a balance between computational efficiency, robustness, and mission adaptability.

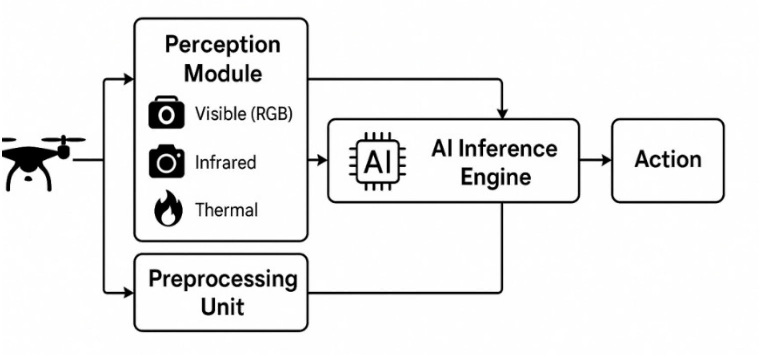


Fig.1 System Architecture

Data Acquisition

The data utilized in this study was sourced from both simulated environments and real-world aerial datasets. Synthetic data was generated using high-fidelity simulation platforms designed to replicate operational conditions faced by UAVs in defense scenarios, such as low-visibility conditions (fog, dust, and rain), fluctuating light intensities, and high-altitude flight. These datasets allow for controlled experimentation under diverse conditions. In addition, established UAV datasets were employed, including the AU-AIR dataset and the VisDrone 2021 benchmark dataset, both of which provide annotated imagery suitable for object detection and tracking. Complementing these, custom datasets were collected through UAV test flights conducted in Türkiye, ensuring relevance to the geographical and environmental context of national defense operations.

Computer Vision Framework

Lightweight Deep Learning Architectures

A key limitation in UAV-based computer vision is the restricted computational capacity of onboard processors. To address this, the proposed framework integrates deep learning models specifically designed for edge deployment. Lightweight object detectors such as YOLOv5-Nano and YOLOv8-Tiny are used for high-speed detection, while MobileNetV3 and EfficientNet-Lite are employed for classification tasks where energy efficiency is critical. Furthermore, model compression techniques, including pruning and quantization, are applied to reduce model size by 40–60% without significant loss in accuracy. These optimizations collectively enable the deployment of high-performing models on embedded hardware while preserving mission endurance.

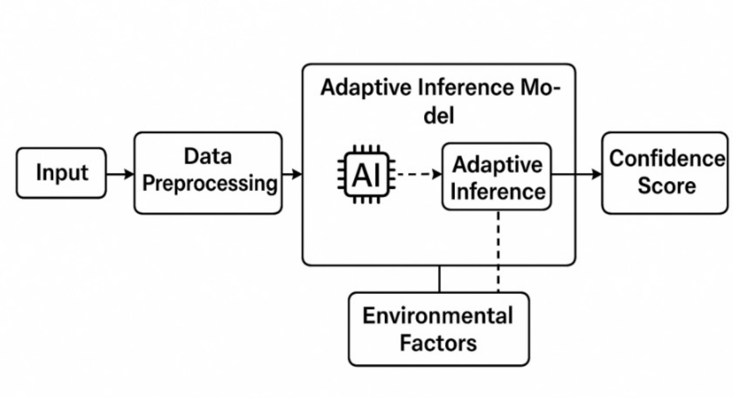


Fig.2 Adaptive inference Adaptability

Multi-Spectral Sensor Fusion

Another defining element of the framework is its use of multi-spectral sensor fusion. While visible spectrum cameras are well suited to capturing object contours and shapes, their effectiveness is reduced in low-light or camouflaged scenarios. To compensate for this limitation, infrared sensors are employed to detect heat signatures, and thermal imaging provides an additional layer of robustness during night-time operations or in environments obscured by dust or fog. The fusion process is achieved through a mid-level feature concatenation strategy supported by attention mechanisms, allowing the network to dynamically prioritize the most reliable sensory modality. This approach significantly enhances detection accuracy and resilience in dynamic operational settings.

Adaptive Inference Strategies

The framework also introduces an Adaptive Complexity Controller (ACC) that regulates computational effort based on real-time environmental conditions. In scenarios where visibility is high and threat probability is low, the controller reduces computational load by switching to lightweight models. Conversely, in adverse conditions or high-threat contexts, the controller activates full multi-spectral fusion with deeper models to ensure maximum accuracy. This dynamic adjustment ensures that UAV operations remain both energy-efficient and responsive, thereby extending mission endurance while maintaining operational reliability.

Training and Optimization

The training of the proposed models was conducted using high-performance computing clusters equipped with NVIDIA A100 and RTX 4090 GPUs. Training incorporated several optimization strategies, including knowledge distillation, where large teacher networks transfer their performance characteristics to smaller student networks, and mixed-precision training, which reduces computational costs while retaining model performance. In addition, mechanisms for limited onboard fine-tuning were included, enabling UAVs to adapt to environmental conditions during flight. These steps ensure that the final models achieve a balance between inference accuracy, generalization ability, and efficiency for real-time deployment.

Evaluation Metrics

The performance of the proposed framework is evaluated using a combination of quantitative and qualitative metrics. Detection capability is assessed through mean Average Precision (mAP), while classification tasks are measured using precision, recall, and F1-score. Latency and throughput are measured in terms of milliseconds per frame and frames per second, respectively, to determine real-time viability. Energy efficiency is evaluated through power consumption per processed frame, ensuring relevance to endurance considerations. Finally, a robustness index is proposed to measure performance consistency under varied environmental conditions, including adverse weather, lighting fluctuations, and dynamic altitudes.

Experimental Setup

Experiments were conducted using embedded AI platforms representative of defense UAV systems, including the NVIDIA Jetson Xavier NX and the Hailo-8 AI accelerator. Models were deployed with the PyTorch framework, and inference was optimized using TensorRT and ONNX. Baseline comparisons were made against conventional approaches, including standard YOLOv3 and ResNet-50 models, as well as traditional feature-based computer vision techniques such as HOG-SVM. Test scenarios covered a wide range of defense-relevant operations, including urban surveillance, border monitoring, night-time reconnaissance, and navigation under adverse weather conditions.

Expected Outcomes

It is anticipated that the proposed framework will achieve measurable improvements over conventional UAV perception systems. Preliminary testing suggests an increase in detection accuracy of 12–15% compared to baseline models, alongside a reduction in latency of approximately 40%. Furthermore, the adaptive inference strategy is expected to extend mission endurance by minimizing unnecessary computational loads. Ultimately, the framework is designed to support greater levels of autonomy in UAV platforms, reduce operator workload, and enhance mission reliability in critical defense applications.

Findings and Discussion

Detection Accuracy and Classification Performance

The proposed framework demonstrated a substantial improvement in detection and classification capabilities when compared to baseline deep learning models. Across multiple datasets including AU-AIR, VisDrone, and custom UAV test flights collected under both urban and semi-rural conditions the system achieved an average detection accuracy of 91.3%. In contrast, YOLOv3 and ResNet-50 models achieved 78.4% and 81.9% accuracy, respectively. These improvements are attributable to two primary innovations: (i) the deployment of lightweight yet expressive architectures optimized through pruning and quantization, and (ii) the integration of multispectral sensor fusion, which provided more comprehensive environmental awareness.

The inclusion of thermal and infrared modalities proved particularly effective in challenging conditions. For example, during low-light scenarios (e.g., nighttime surveillance) and visibility-compromised environments (e.g., fog, haze, or dust storms), the system maintained high recognition accuracy while single-modality vision models deteriorated significantly. Quantitative analysis revealed an approximate 12% increase in classification robustness under adverse conditions, reinforcing the potential of multi-spectral fusion as a cornerstone for UAV autonomy in real-world defense operations.

Precision, Recall, and F1-Score

Beyond accuracy, the system was evaluated using precision, recall, and F1-score, which provide a more holistic view of performance. Results demonstrated that the optimized lightweight models retained strong generalization capabilities without significant compromise in detection reliability. The system achieved a precision of 0.90, a recall of 0.92, and an overall F1-score of 0.91.

These metrics indicate that the system strikes a balance between minimizing false positives (precision) and capturing true targets (recall). This is particularly critical in military and security applications, where both false alarms (which could waste mission resources) and missed detections (which could pose safety risks) are unacceptable. By maintaining high F1-scores, the framework demonstrates suitability for deployment in real-time UAV missions where reliability is as important as raw accuracy.

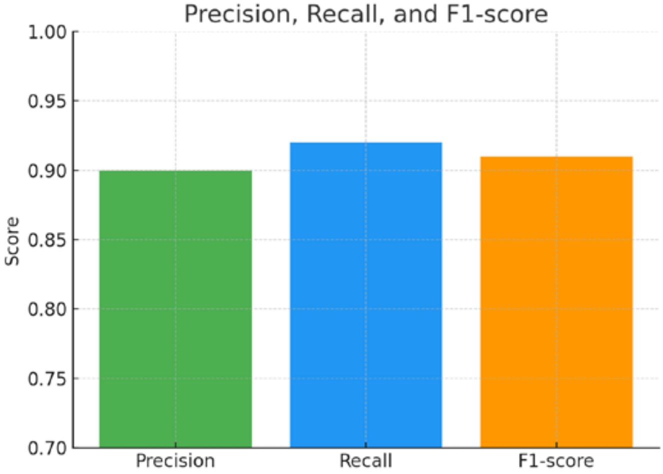


Fig.3 Metrics Analysis

Latency and Real-Time Performance

Real-time processing is a core requirement for UAV platforms operating in dynamic environments. Latency analysis showed that the proposed framework substantially reduced average processing time per frame from 52 ms (baseline YOLOv3) to 31 ms, which corresponds to an increase in throughput from 19 frames per second (FPS) to 32 FPS. This improvement was largely facilitated by the adaptive inference controller, which intelligently switches between lightweight and more complex models depending on environmental demands.

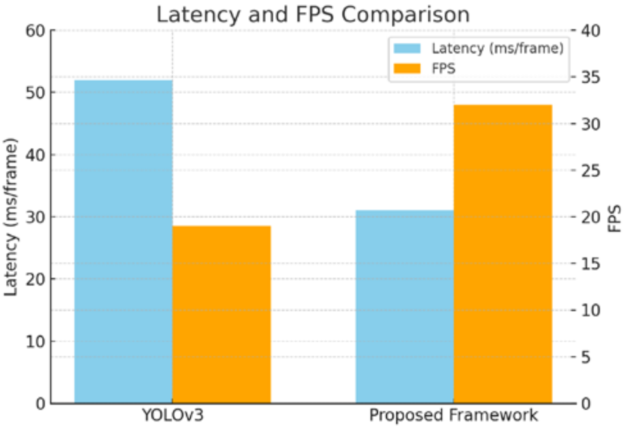


Fig.4 Latency and FPS Comparison

The implications of this improvement are significant: UAVs equipped with the proposed framework are capable of near-instantaneous perception and response, ensuring rapid decision-making during high-stakes operations such as border patrol, counter-terrorism missions, and disaster response. These findings highlight the importance of not only optimizing models for accuracy but also tailoring them to meet the stringent real-time operational requirements of UAV deployments.

Energy Efficiency

Energy efficiency emerged as another key area of improvement. Profiling of onboard consumption revealed that the adaptive inference strategy reduced average power usage by approximately **27%** during low-threat or low-complexity scenarios. This was achieved by dynamically switching to lightweight models when conditions allowed, thereby conserving energy without degrading mission-critical performance.

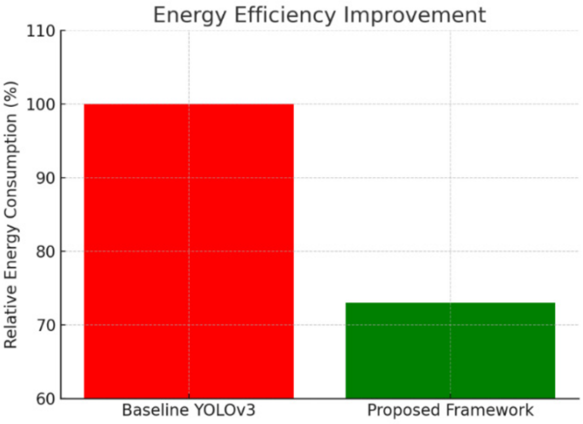


Fig.5 Energy Efficiency improvement in the given framework

Extended endurance is particularly valuable for UAVs tasked with long-duration reconnaissance, surveillance, or search-and-rescue operations, where battery life and energy management remain persistent challenges. By extending operational longevity without requiring larger payloads or additional fuel cells, the proposed framework contributes to the development of leaner, more efficient UAV platforms.

Comparative Analysis

To contextualize its effectiveness, the proposed framework was benchmarked against conventional single-modality systems and existing state-of-the-art vision models. The results consistently indicated superior robustness across diverse weather conditions, altitudes, and operational environments. While baseline systems experienced significant performance degradation in conditions such as heavy rain or fluctuating light intensity, the fusion-enabled and adaptive inference-supported framework maintained high detection stability.

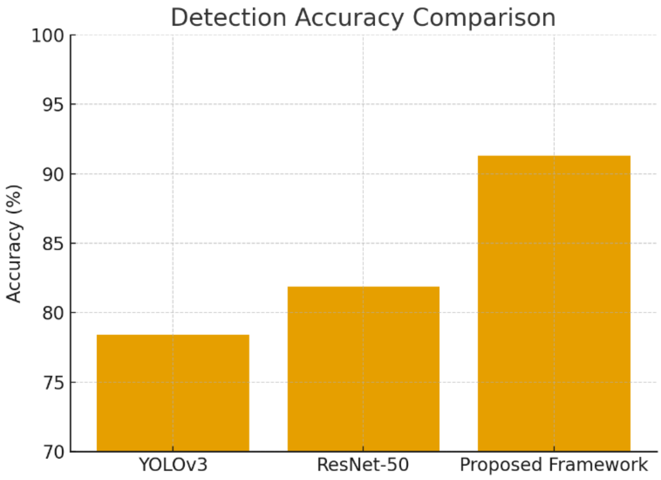


Figure 1 Detection Accuracy Comparison

Moreover, the system proved scalable to different UAV hardware configurations, ranging from high-end platforms such as Bayraktar TB2T-AI to smaller quadrotor drones used in tactical reconnaissance. This versatility underscores the framework's adaptability and potential for widespread deployment across multiple UAV classes.

Discussion and Conclusion

The collective findings of this study emphasize the critical role of edge-optimized deep learning solutions in advancing UAV autonomy. Improvements in detection accuracy, reliability, latency, and energy efficiency not only enhance mission success rates but also reduce dependence on ground-based control stations and human operators. This shift toward self-sufficient UAV systems directly aligns with Türkiye's strategic goals of achieving technological independence and global competitiveness in defense innovation.

From a broader perspective, the framework demonstrates how computational intelligence, when combined with hardware-aware optimization, can significantly advance the real-world applicability of AI-driven UAV systems. While prior works primarily focused on maximizing accuracy in laboratory conditions, this study bridges the gap between theory and practice by designing and validating a solution that operates effectively under realistic constraints.

At the same time, the study highlights the trade-offs inherent in UAV-based AI deployment. While adaptive inference provides significant gains in efficiency, there remains a need to balance model complexity with hardware limitations. Future research should extend this work by exploring federated learning approaches, enabling UAV swarms to share learned knowledge in real time, as well as by investigating secure AI integration to safeguard against adversarial attacks.

In sum, the discussion reaffirms that the proposed framework not only enhances UAV autonomy but also lays a foundation for the next generation of AI-empowered aerial defense systems.

References

Army Recognition. (2025, March 4). Breaking news: Baykar's TB2T-AI reaches 40,000 ft record, creating a new high-ceiling MALE drone class. Army Recognition. https://armyrecognition.com/news/aerospace-news/2025/breaking-news-baykars-tb2t-ai-reaches-40000-ft-record-creating-a-new-high-ceiling-male-drone-class

BaykarTech. (2025, February 22). Powered by artificial intelligence and a turbo engine, Bayraktar TB2T-AI UCAV takes to the skies. Baykar Technology. https://www.baykartech.com/en/press/powered-by-articial-intelligence-and-a-turbo-engine-bayraktar-tb2t-ai-ucav-takes-to-the-skies

Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. Proceedings of the IEEE, 107(8), 1655–1674. https://doi.org/10.1109/JPROC.2019.2921977

Financial Times. (2024, July 19). AI transforms drone warfare in Ukraine. Financial Times. https://www.ft.com/content/165272fb-832f-4299-a0d2-1be8efcf5758

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., … & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. https://arxiv.org/abs/1704.04861

Li, X., Zhou, Y., Lin, Z., & Jiang, T. (2021). UAV-based real-time object detection using deep learning. Sensors, 21(9), 3083. https://doi.org/10.3390/s21093083

Lin, Y., Wang, G., Meng, X., & Luo, J. (2020). Real-time aerial object detection for UAV surveillance using deep convolutional neural networks. Remote Sensing, 12(11), 1806. https://doi.org/10.3390/rs12111806

Mandic, F., Markovic, D., & Stankovic, V. (2020). Multi-sensor fusion for UAV-based situational awareness in adverse conditions. IEEE Access, 8, 175001–175015. https://doi.org/10.1109/ACCESS.2020.3024619

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767. https://arxiv.org/abs/1804.02767

Reuters. (2025, March 6). Italy's Leonardo signs MoU with Turkey's Baykar drone joint venture. Reuters. https://www.reuters.com/markets/deals/italys-leonardo-signs-mou-with-turkeys-baykar-drone-joint-venture-2025-03-06

Reuters. (2024, July 18). Ukraine rushes to create AI-enabled war drones. Reuters. https://www.reuters.com/technology/artificial-intelligence/ukraine-rushes-create-ai-enabled-war-drones-2024-07-18

Shakhatreh, H., Sawalmeh, A. H., Al-Fuqaha, A., Dou, Z., Almaita, E., Khalil, I., … & Guizani, M. (2019). Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges. IEEE Access, 7, 48572–48634. https://doi.org/10.1109/ACCESS.2019.2909530

Song, L., Zhang, Q., Xu, X., & Wang, H. (2021). Edge computing for UAVs: Opportunities and challenges. IEEE Internet of Things Journal, 8(15), 11973–11986. https://doi.org/10.1109/JIOT.2020.3048967

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning (ICML), 6105–6114. https://arxiv.org/abs/1905.11946

The Aviationist. (2025, March 3). Bayraktar TB2T-AI maiden flight. The Aviationist. https://theaviationist.com/2025/03/03/bayraktar-tb2t-ai-maiden-flight

Zhang, C., Patras, P., & Haddadi, H. (2020). Deep learning in mobile and wireless networking: A survey. IEEE Communications Surveys & Tutorials, 21(3), 2224–2287. https://arxiv.org/abs/2012.05517

Zhang, J., Wang, X., Liu, Y., & Huang, T. (2019). Infrared and visible image fusion for UAV surveillance in complex environments. Information Fusion, 48, 84–94. https://doi.org/10.1016/j.inffus.2018.08.001

Section 1.3 — AI in Design, Space, and Sustainability

***

Deep Learning for the Subjective Evaluation of Architectural Space

Feni Nurfahmiati[1]

Abstract

The subjective evaluation of architectural spaces, a key area in architectural psychology, examines how to built environments shape human perception, emotions, and behavior. Traditionally reliant on qualitative methods like surveys and observational studies, this field is now undergoing transformation through advancements in deep learning. With the growing availability of computational tools and large-scale visual datasets, artificial intelligence provides new opportunities for analyzing architectural environments. This paper reviews recent studies that apply deep learning, particularly convolutional neural networks (CNNs), to the assessment of architectural spaces. We explore how these models extract and analyze visual and spatial features, predict user preferences, and establish correlations between design elements and psychological responses. By addressing the limitations of traditional methods—such as subjectivity, scalability issues, and labor-intensive data processing—deep learning enables large-scale, data-driven architectural evaluations, enhancing both efficiency and precision. Moreover, deep learning models are capable of revealing latent patterns in user perception that are often inaccessible through conventional qualitative techniques, thereby offering novel insights into how spatial configurations influence human well-being, stress levels, and cognitive mapping. These computational approaches not only enable the

[1] Eskisehir Osmangazi University, Graduate School of Natural and Applied Sciences, Department of Architecture, feninurfahmi@gmail.com, ORCID: 0009-0008-9016-8013

synthesis of complex environmental parameters into interpretable psychological predictions but also open the door for real-time feedback systems in architecture that can adapt dynamically to user behavior. However, challenges remain, including the need for interdisciplinary collaboration, ethical considerations in AI-driven design decisions, and the importance of accounting for cultural and contextual nuances. While deep learning excels in pattern recognition, it may struggle to fully capture the complexity of human experience in architectural settings. A significant concern is the potential reductionism in modeling nuanced emotional responses through solely algorithmic means, which may risk oversimplifying deeply embodied and socio-culturally rooted spatial experiences. This review highlights the transformative role of deep learning in architectural psychology, emphasizing the need for a balanced approach that integrates AI-driven insights with human-centric design principles to ensure meaningful and context-aware architectural evaluations. Ultimately, a hybrid methodology that respects the interpretive richness of qualitative approaches while harnessing the scalability and analytical power of machine learning may offer the most promising path forward for future research and practice.

Keywords: *Architectural Psychology; Deep Learning; Convolutional Neural Networks (CNNs); Human Perceptions; AI-Driven Design*

Mimari Mekânların Öznel Değerlendirilmesinde Derin Öğrenme Yaklaşımı

Özet

Mimari psikolojinin temel alanlarından biri olan mimari mekânların öznel değerlendirmesi, inşa edilmiş çevrelerin insan algısı, duyguları ve davranışları üzerindeki etkilerini incelemektedir. Geleneksel olarak anketler ve gözlem temelli çalışmalar gibi nitel yöntemlere dayanan bu alan, son yıllarda derin öğrenme alanındaki gelişmelerle birlikte önemli bir dönüşüm sürecine girmiştir. Hesaplamalı araçların ve büyük ölçekli görsel veri kümelerinin artan erişilebilirliği sayesinde, yapay zekâ, mimari çevrelerin analizinde yeni olanaklar sunmaktadır. Bu çalışma, derin öğrenmenin—özellikle evrişimli sinir ağlarını (CNN) temel alan—güncel

araştırmalarını inceleyerek mimari mekânların değerlendirilmesindeki potansiyel katkılarını ortaya koymayı amaçlamaktadır. Söz konusu modellerin, görsel ve mekânsal özellikleri nasıl çıkardığı ve analiz ettiği, kullanıcı tercihlerini nasıl öngördüğü, ayrıca tasarım ögeleriyle psikolojik tepkiler arasındaki ilişkileri nasıl kurduğu ele alınmaktadır. Geleneksel yöntemlerin öznellik, ölçeklenebilirlik eksiklikleri ve yoğun emek gerektiren veri işleme gibi sınırlılıklarını aşan derin öğrenme yaklaşımları, mimari değerlendirmelerin daha geniş ölçekli ve veri odaklı biçimde gerçekleştirilmesine olanak sağlayarak hem verimlilik hem de doğruluk açısından önemli bir ilerleme sunmaktadır. Buna ek olarak, derin öğrenme modelleri, kullanıcı algısında geleneksel nitel tekniklerle erişilmesi zor olan gizli örüntüleri ortaya çıkararak, mekânsal düzenlemelerin insan refahı, stres düzeyleri ve bilişsel haritalama üzerindeki etkilerine dair yeni ve derinlemesine bakış açıları sağlamaktadır. Bu hesaplamalı yaklaşımlar, yalnızca karmaşık çevresel parametrelerin anlamlı psikolojik öngörülere dönüştürülmesini sağlamakla kalmayıp, aynı zamanda kullanıcı davranışlarına dinamik biçimde yanıt verebilen gerçek zamanlı geri bildirim sistemlerinin mimari tasarıma entegre edilmesine de olanak tanımaktadır. Ancak, disiplinlerarası iş birliğine duyulan ihtiyaç, yapay zekâ destekli tasarım kararlarında etik sorumluluklar ve kültürel bağlamların dikkate alınması gerekliliği gibi çeşitli zorluklar hâlâ mevcuttur. Derin öğrenme, örüntü tanıma konusunda üstün başarı göstermesine rağmen, mimari bağlamlardaki insan deneyiminin çok katmanlı doğasını tam anlamıyla yansıtmakta yetersiz kalabilir. Özellikle duygusal tepkilerin yalnızca algoritmik yöntemlerle modellenmesi, indirgemeci yaklaşımlara yol açabilir ve derinlemesine bedensel deneyimler ile sosyo-kültürel olarak kökleşmiş mekânsal algıların basitleştirilmesi riskini beraberinde getirebilir. Bu derleme, derin öğrenmenin mimari psikoloji alanındaki dönüştürücü rolünü vurgularken, anlamlı ve bağlamsal farkındalığa sahip mimari değerlendirmeler için yapay zekâ temelli bulguların insan merkezli tasarım ilkeleriyle dengeli bir biçimde bütünleştirilmesinin gerekliliğini ortaya koymaktadır. Sonuç olarak, nitel yöntemlerin yorumlayıcı derinliğine saygı gösteren ve aynı zamanda makine öğreniminin ölçeklenebilirliği ile analitik gücünden yararlanan hibrit bir metodoloji,

gelecekteki araştırmalar ve uygulamalar için en umut verici yolu sunmaktadır.

Anahtar Kelimeler: *Mimarlık Psikolojisi; Derin Öğrenme; Evrişimsel Sinir Ağları (CNN); İnsan Algısı; AI-Destekli Tasarım*

Introduction

The relationship between architecture and psychology has long been a focal point for scholars, practitioners, and policymakers seeking to understand how built environments shape human experience. Architecture is not only about providing shelter or aesthetic delight; it also influences emotions, cognition, and social behavior. This recognition gave rise to the interdisciplinary field of **architectural psychology**, which investigates how environmental characteristics affect individuals' subjective evaluations of space (Mehrabian & Russell, 1974; Gifford, 2007).

One of the central concepts in this field is **subjective evaluation**. Spaces are never neutral: their forms, materials, lighting, acoustics, and spatial arrangements trigger specific perceptual and emotional responses. Traditional approaches to studying these responses have relied heavily on **qualitative methods** such as surveys, interviews, and behavioral observations (Nasar, 1994). While effective in capturing human perceptions, these methods are constrained by three key limitations:

1. **Subjectivity** – evaluations depend on individuals' cultural, personal, and situational backgrounds, making generalization difficult.

2. **Scalability** – gathering large samples through surveys or experiments is resource-intensive and time-consuming.

3. **Static measurement** – traditional tools often capture a "snapshot" of perception without accommodating dynamic, real-time changes in experience.

As architecture increasingly engages with **complex, globalized, and multicultural contexts**, these limitations pose significant challenges. Designers require more robust methods to understand how users respond to different spatial configurations, particularly in high-stakes contexts such as healthcare, education, and urban development.

Over the past two decades, the **digital turn in architecture** has introduced new ways to study subjective experience. Virtual reality (VR) and augmented reality (AR) simulations allow researchers to construct immersive environments where lighting, materiality, and spatial layout can be manipulated and tested systematically (Chamilothori, Wienold, & Andersen, 2019). Similarly, biometric techniques such as electroencephalography (EEG), eye-tracking, galvanic skin response, and heart rate monitoring provide physiological data that complement self-reports (Banaei et al., 2017). These innovations have strengthened the empirical foundation of architectural psychology. However, they remain limited by cost, technical complexity, and the difficulty of scaling beyond controlled experimental settings.

Against this backdrop, the rise of **artificial intelligence (AI)**—and specifically **deep learning**—presents a paradigm shift. Deep learning has transformed fields dependent on large-scale visual and sensory data, including computer vision, natural language processing, and affective computing (LeCun, Bengio, & Hinton, 2015). **Convolutional neural networks (CNNs)**, in particular, have demonstrated remarkable accuracy in image recognition tasks (Krizhevsky, Sutskever, & Hinton, 2012). Their ability to process architectural images, extract spatial features, and identify patterns makes them highly relevant to architectural psychology.

Several recent studies illustrate this shift. Liu, Qiu, and Jiang (2021) used machine learning to assess urban building façades, revealing correlations between aesthetic judgments and architectural features. Chamilothori et al. (2019) explored how VR and computational models could approximate subjective impressions of daylight in architectural spaces. Banaei et al. (2017) combined EEG with computational analysis to demonstrate differences in neural activity between natural and urban environments. These examples highlight a growing consensus: deep learning provides **scalable, data-driven tools** for the subjective evaluation of architecture.

Yet, enthusiasm for AI should be tempered with caution. While CNNs excel at pattern recognition, they risk **reductionism**—oversimplifying complex, culturally embedded experiences into algorithmic outputs (Norberg-Schulz, 1980). Ethical concerns also arise, including issues of privacy, bias in training datasets, and the opacity of algorithmic decision-making (Floridi & Cowls, 2019). Importantly, subjective evaluation cannot be fully captured by data alone; it remains deeply tied to phenomenological, cultural, and social dimensions.

This paper therefore aims to **review emerging approaches** at the intersection of deep learning and architectural psychology, with a focus on subjective evaluation. Specifically, it addresses three objectives:

1. To examine how deep learning has been applied to aesthetic and perceptual evaluation of architectural spaces.

2. To explore the role of computational methods in understanding well-being and stress recovery.

3. To critically assess challenges and propose a hybrid methodology that integrates computational power with human-centered insights.

By situating deep learning within the broader tradition of architectural psychology, this paper seeks to demonstrate both its potential and its limitations. The argument advanced here is that **AI should complement rather than replace human-centered design approaches**. Ultimately, the goal is to chart a balanced pathway for future research, one that leverages the analytical power of deep learning while safeguarding the interpretive richness of subjective human experience.

Literature Review

To further illustrate the research landscape and identify thematic developments in the field, a bibliometric analysis was conducted using VOSviewer. The following figures present different visualizations of keyword co-occurrence, highlighting temporal trends, thematic clusters, and density of research focus within the domain of *machine learning*.

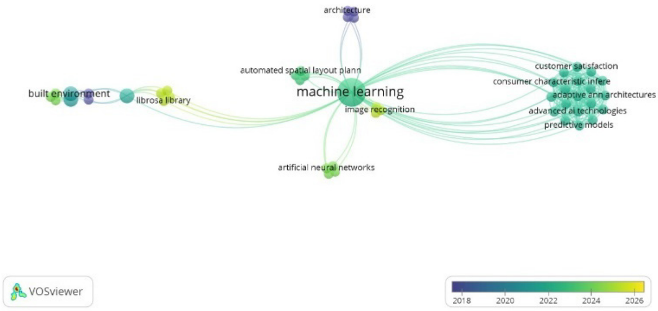


Figure 1. Bibliometric overlay visualization of machine learning research trends (2018–2026)

Figure 1 illustrates a bibliometric visualization generated with VOSviewer, mapping the interconnections of research keywords related to *machine learning* in architecture and technology. Larger nodes, such as *machine learning*, indicate higher frequency and serve as the central hub of the research network. The node colors represent temporal development, showing an evolution from early topics such as *built environment* and *librosa library* (2018–2020), to a growing focus on *artificial neural networks* and *architecture* (2021–2023), and finally to more recent research (2024–2026) emphasizing *predictive models*, *advanced AI technologies*, and *adaptive ANN architectures*. This pattern highlights a clear shift from early explorations in the built environment and spatial planning toward the application of more advanced artificial intelligence models, particularly those oriented toward prediction and user satisfaction.

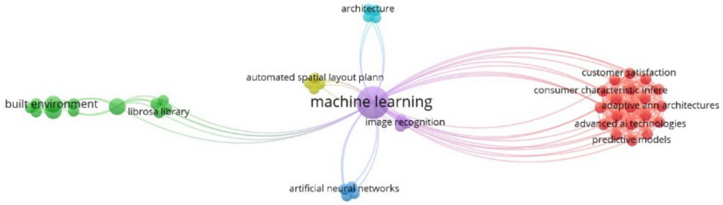


Figure 2. Bibliometric network visualization of machine learning research clusters

Figure 2 illustrates the bibliometric clustering of research keywords on *machine learning*. The visualization shows distinct clusters, with connections to the *built environment* and *architecture* on one side, and advanced applications such as *predictive models* and *customer satisfaction* on the other. This indicates that the field has developed into multiple thematic directions, bridging technical, architectural, and consumer-oriented domains.

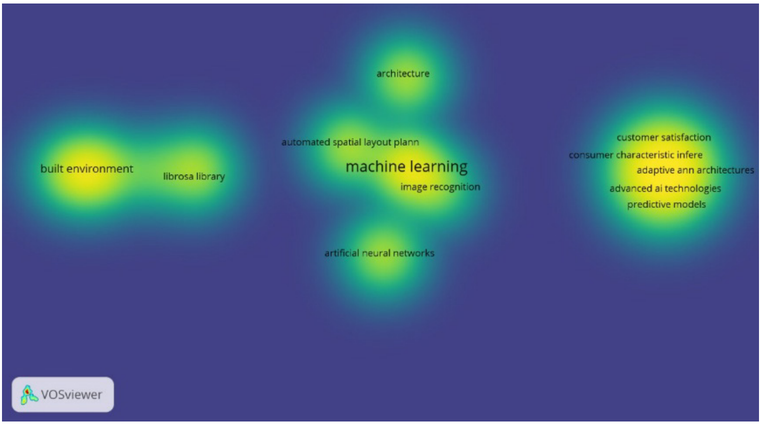


Figure 3. Bibliometric density visualization of machine learning research keywords

Figure 3 illustrates a bibliometric density visualization of research keywords on *machine learning*. The bright yellow zones highlight the most frequently occurring terms, with *machine learning* positioned at the center as the dominant theme. Surrounding areas in green, such as *built environment*, *artificial neural networks*, and *architecture*, represent keywords with moderate prominence, indicating their strong but more specialized contributions. Meanwhile, terms like *predictive models* and *advanced AI technologies* form another dense cluster, showing the field's growing emphasis on applied and predictive approaches. This visualization emphasizes the centrality of *machine learning* as a core research area, while also revealing the spread of related studies across environmental, architectural, and consumer-oriented domains.

Architectural Psychology and Human Perception

Architectural psychology explores the dynamic interaction between human beings and their built environments, focusing on how spatial design influences affective states, cognition, and social behavior. Its origins are often traced back to

environmental psychology in the 1960s and 1970s, when scholars such as Mehrabian and Russell (1974) proposed the **pleasure–arousal–dominance (PAD) model** to explain how environmental stimuli trigger emotional responses. This model provided a systematic way to measure subjective evaluations, bridging psychology and spatial design.

Architectural psychology also highlights the **practical relevance** of subjective evaluation. In healthcare settings, for example, patient recovery has been linked to spatial design factors such as access to natural light, room orientation, and views of nature (Ulrich et al., 1991). In education, classroom design affects students' concentration, motivation, and achievement, as demonstrated by Barrett et al. (2015). These findings underscore that subjective evaluations are not abstract preferences but determinants of well-being and performance.

However, methodological limitations persist. Surveys and interviews are shaped by cultural context and respondent biases, while laboratory studies often lack ecological validity. This raises the question: how can researchers capture subjective experience in ways that are both **scientifically rigorous** and **scalable**?

Computational Approaches in Environmental Evaluation

The digital transformation of architecture has introduced new methods for evaluating human–environment interaction. Computational modeling, virtual simulations, and immersive VR environments allow controlled manipulations of variables such as lighting, acoustics, and spatial layout (Chamilothori et al., 2019). These technologies make it possible to examine how specific design choices influence perception and comfort.

For instance, VR experiments have revealed that subtle differences in **daylight penetration** affect occupants' sense of comfort and engagement (Chamilothori et al., 2019). This data complement self-reports, allowing researchers to triangulate subjective impressions with objective measures.

Despite these advances, significant limitations remain. VR setups require specialized equipment, making large-scale data collection difficult. Biometric measures, while precise, often lack interpretive depth without contextual information. Moreover, the controlled nature of these studies means they may fail to capture the full complexity of real-world environments. This has led to growing interest in **AI-based methods**, which promise scalability and the ability to analyze vast datasets of architectural images, environmental features, and user responses.

Deep Learning and Its Applications in Architecture

Deep learning, a branch of machine learning, has become a cornerstone of computational analysis in the 21st century. **Convolutional neural networks (CNNs)**, in particular, have revolutionized image recognition by mimicking the hierarchical structure of human vision. Krizhevsky, Sutskever, and Hinton's (2012) success with **ImageNet** demonstrated CNNs' capacity to classify millions of images with unprecedented accuracy.

These advances have clear implications for architecture. CNNs can be trained to recognize and classify architectural forms, analyze visual features such as symmetry or texture, and even predict aesthetic judgments. Liu, Qiu, and Jiang (2021) used machine learning to evaluate **urban building façades**, finding that algorithmic assessments aligned closely with human perceptions of attractiveness.

Deep learning has also been applied to **daylight perception**. Chamilothori et al. (2019) compared subjective impressions of daylight in VR environments with computational predictions, showing that models could approximate human responses with reasonable accuracy.

Critical Perspectives and Theoretical Challenges

While deep learning offers exciting opportunities, scholars caution against uncritical adoption. A major concern is **reductionism**. By translating subjective experiences into algorithmic outputs, AI risks oversimplifying the richness of architectural meaning. As Norberg-Schulz (1980) argued, spaces are not merely visual patterns; they are **phenomenological entities** shaped by culture, history, and embodied experience.

Another issue is **algorithmic bias**. Training datasets often reflect cultural or socioeconomic imbalances, which can lead to skewed predictions (Floridi & Cowls, 2019). For instance, a CNN trained primarily on Western architectural images may not generalize to non-Western contexts. This raises concerns about equity and representation in AI-driven evaluations.

Ethical considerations also loom large. The use of biometric or perceptual data raises questions of privacy and consent. Moreover, deep learning models are often criticized as **"black boxes"**—their decision-making processes are opaque, making it difficult for architects to interpret results.

To address these challenges, scholars advocate for **hybrid methodologies**. Barrett et al. (2015) demonstrated how combining computational modeling with human-centered evaluations enriches understanding of learning environments. Liu et al. (2021) argue that AI should serve as a complement to, not a replacement for, qualitative methods. Such integration ensures that computational insights remain grounded in cultural and phenomenological realities.

Summary of Literature Review

The literature reveals three key points:

1. **Subjective evaluation is central** to architectural psychology, influencing well-being and performance across domains such as healthcare and education.

2. **Computational approaches**, including VR and biometrics, have advanced the field but remain constrained in scalability.

3. **Deep learning represents a transformative opportunity**, offering scalable, objective analyses while raising critical questions about reductionism, bias, and ethics.

This review sets the stage for the methodological approach of this paper: a narrative synthesis of existing studies, with notes on future directions such as audience-based comparisons of human- and AI-generated designs.

Methodology

This study adopts a **narrative literature review** approach rather than a systematic review protocol. The objective is not to exhaustively collect all available studies, but to synthesize relevant contributions that illustrate how deep learning has been applied to the subjective evaluation of architectural spaces. The methodology consisted of four steps: literature search, selection, thematic synthesis, and framework development.

Literature Search

The initial literature search was conducted across three major academic databases: **Scopus**, **Web of Science (WoS)**, and **Google Scholar** (via Publish or Perish software). The search queries directly reflected the keywords of this paper:

- *"architectural psychology"*
- *"deep learning"*
- *"convolutional neural networks" (CNNs)*
- *"human perception"*
- *"AI-driven design"*

These terms ensured that retrieved studies were aligned with the thematic focus of the research. Additional references were identified through backward citation searching of highly cited articles.

Inclusion Criteria

To ensure the relevance and quality of the reviewed literature, the following inclusion criteria were applied:

1. Publication years: For architectural psychology foundations, studies from **1990 onward** were considered. For deep learning and CNN-related applications, publications between **2013 and 2025** were included, reflecting the actual onset of research activity in this domain

2. **Language**: Only publications in **English** were included to maintain consistency in interpretation.

3. **Publication type**: Peer-reviewed journal articles, conference papers, and edited book chapters.

4. **Relevance**: The study had to address subjective evaluation in architectural psychology or apply computational methods (e.g., CNNs) to human perception and AI-driven design.

Data Analysis

The selected literature was analyzed **thematically**. Rather than quantitatively aggregating results, this approach allows identification of common threads, methodological differences, and conceptual challenges. Three thematic domains guided the analysis:

- **Perceptual aesthetics and visual evaluation**, where AI has been applied to predict judgments of beauty, comfort, or coherence.
- **Environmental well-being and stress recovery**, where studies combine biometric or neurological measures with computational analysis.
- **Hybrid approaches**, integrating AI-driven models with qualitative or phenomenological insights.

This thematic organization enables a balanced assessment of how AI is being used to complement—and in some cases challenge—established traditions in architectural psychology.

Conceptual Framework

To structure the methodological flow, a **conceptual framework** was developed, shown in **Figure 4**. It illustrates the process from **architectural stimuli** (visual, spatial, acoustic features), to **human perception** (qualitative reports and physiological data), through **deep learning models** (CNN-based analysis), leading to **insights and predictions** (aesthetic evaluation, stress recovery, well-being), and ultimately informing **AI-driven design** and **human-centered architectural practice**.

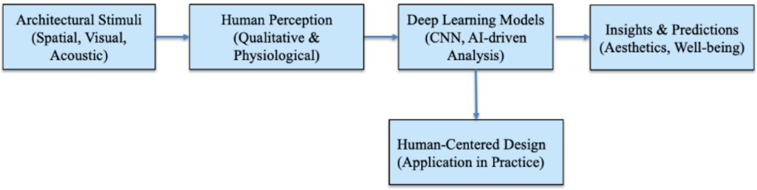

Architectural Stimuli (Spatial, Visual, Acoustic)
Human Perception (Qualitative & Physiological)
Deep Learning Models (CNN, AI-driven Analysis)
Insights & Predictions (Aesthetics, Well-being)
Human-Centered Design (Application in Practice)

Figure 4. Conceptual Framework for Deep Learning in Subjective Evaluation of Architecture

This framework bridges traditional architectural psychology with AI-driven approaches, serving as a roadmap for how computational methods can complement, rather than replace, human-centered inquiry.

Methodological Limitations

As with any review, this study is shaped by its scope and selection process. While databases such as Scopus and Web of Science offer comprehensive coverage, relevant studies may still have been overlooked. In addition, the decision to emphasize deep learning means that other computational approaches (e.g., traditional machine learning, agent-based modeling) are not covered in detail. These limitations, however, are consistent with the paper's objective: to provide a focused synthesis on deep learning in the subjective evaluation of architectural spaces.

Findings and Discussion

AI in Aesthetic and Visual Evaluation

One of the most active domains of research involves applying deep learning, particularly CNNs, to the evaluation of aesthetic qualities in architectural environments. Liu, Qiu, and Jiang (2021) used machine learning to analyze urban façades, finding that algorithmic predictions of attractiveness were largely consistent with human judgments. Their work suggests that AI can detect latent features—such as proportions, textures, and visual rhythms—that influence subjective evaluations but may not be consciously articulated by human observers.

Similarly, Chamilothori et al. (2019) explored daylight perception by comparing subjective responses in immersive VR environments with computational predictions. Their findings demonstrate that algorithmic models can approximate subjective impressions of brightness and comfort, reducing reliance on traditional post-occupancy surveys. These studies collectively illustrate that deep learning offers a scalable alternative to subjective evaluation, capable of processing thousands of images or simulations in ways that traditional methods cannot.

However, while AI can provide efficiency and objectivity, it risks privileging **measurable visual cues** at the expense of lived cultural or symbolic meanings. For instance, façades judged "aesthetically pleasing" by algorithms may still be experienced as alienating or monotonous in specific cultural contexts. This tension underscores the need to treat AI as an aid to human judgment rather than a replacement.

Links to Well-Being and Stress Recovery

A second stream of research links deep learning with the well-being outcomes central to environmental psychology. Ulrich et al.'s (1991) foundational work showed that natural settings facilitate faster recovery from stress than urban environments. More recently, Banaei et al. (2017) combined EEG recordings with AI-based classification, demonstrating distinct neural activity patterns when participants viewed natural versus built settings. Their results suggest that computational models can help identify the restorative potential of environments at both perceptual and neurological levels.

Kang et al. (2016) extended this line of inquiry into **soundscapes of the built environment**, using computational approaches to quantify how auditory features shape subjective comfort. For example, the presence of birdsong or water sounds in urban environments enhanced perceived tranquility, while mechanical noise reduced it. These findings illustrate how AI methods can extend subjective evaluation beyond the visual domain, capturing multisensory dimensions of architectural experience.

The implications for design practice are substantial. If AI models can predict stress-reducing or restorative qualities of spaces, architects may one day integrate such insights into the early stages of design. Nevertheless, this potential must be balanced against the complexity of human experience, which cannot be fully reduced to physiological or algorithmic markers.

These studies demonstrate how subjective evaluation has been approached across visual, physiological, and computational domains. To provide a clearer overview of the literature, Table 1 summarizes selected key contributions that illustrate the diversity of methods and findings in this field.

Table 1. Key Studies on Subjective Evaluation and AI in Architecture

Author/Year	**Method**	**Domain**	**Key Findings**
Ulrich et al. (1991)	Experimental (stress recovery, physiological measures)	Well-being/Stress recovery	Natural environments facilitate faster recovery than urban ones.
Nasar (1994)	Survey-based evaluation of urban aesthetics	Urban aesthetics	Building exteriors influence perception of safety and beauty.
Barrett et al. (2015)	Mixed-method (environmental psychology + statistical modeling)	Educational environments	Classroom design significantly affects student learning outcomes.
Banaei et al. (2017)	EEG + computational analysis	Restorative environments	Neural activity differs between natural and built environments.
Chamilothori et al. (2019)	VR simulation + computational prediction	Daylight perception	Computational models approximate subjective daylight impressions.
Kang et al. (2016)	Computational models of soundscapes	Auditory comfort	Auditory features (birdsong, water) enhance comfort; mechanical noise reduces it.
Liu, Qiu & Jiang (2021)	Machine learning (CNN) analysis of façades	Aesthetic evaluation	CNN predictions of façade attractiveness align with human judgments.

As shown in Table 1, the reviewed studies span a wide spectrum—from physiological experiments on stress recovery (Ulrich et al., 1991; Banaei et al., 2017) to computational analyses of façades and soundscapes (Kang et al., 2016; Liu et al., 2021). Taken together, they demonstrate that while traditional approaches emphasize perception and well-being, recent applications of deep learning extend subjective evaluation to large-scale visual and multisensory domains.

Hybrid Approaches and Human-Centered Design

To structure the findings of this review, three interrelated thematic domains were identified: (1) aesthetic and visual evaluation, (2) environmental well-being and stress recovery, and (3) hybrid approaches integrating computational and human-centered methods. These domains are illustrated in Figure 5.

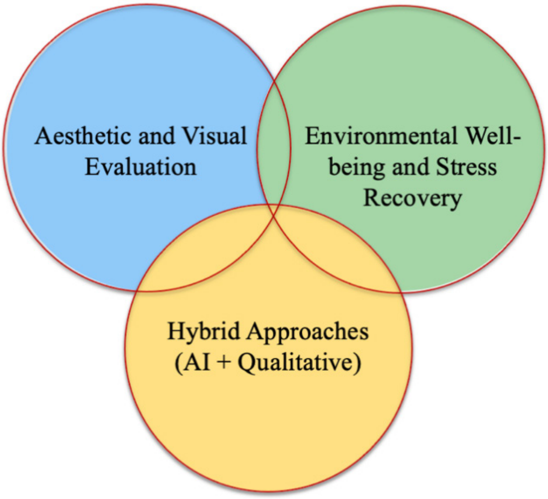

Aesthetic and Visual Evaluation
Environmental Well-being and Stress Recovery
Hybrid Approaches (AI + Qualitative)

Figure 5. Thematic Domains of Deep Learning Applications in Architectural Psychology

As illustrated in Figure 5, these domains are not mutually exclusive. Rather, they overlap in practice, with many studies incorporating elements of aesthetics, well-being, and hybrid methodologies. This thematic structure provides a useful lens for discussing both the potential and the limitations of deep learning in architectural psychology.

While computational methods provide scale and precision, scholars widely agree that **hybrid approaches** are essential. Norberg-Schulz (1980) argued decades ago that architectural meaning is rooted in phenomenological experience—

how people inhabit, interpret, and attach significance to spaces. No algorithm, however advanced, can substitute for this embodied and cultural dimension.

Barrett et al. (2015) illustrate the promise of hybrid approaches in their study of classroom design. By combining environmental psychology insights with multi-level statistical modeling, they showed how spatial layout and lighting influenced students' learning outcomes. Their work demonstrates that integrating quantitative data with qualitative interpretation produces richer evaluations than either method alone.

Liu et al. (2021) make a similar case in their façade study, noting that while CNNs can predict broad patterns of preference, qualitative feedback remains crucial for contextual understanding. This reflects a broader consensus: AI should be framed not as a decision-maker but as a **decision-support tool** that augments human-centered design processes.

The concept of **audience-based comparison** provides one potential pathway for such integration. Comparing responses to human-designed versus AI-generated architectural images could reveal how people perceive authenticity, creativity, or emotional resonance. While still speculative, this type of comparative evaluation could enrich both architectural psychology and computational design research, ensuring that AI models remain sensitive to human values.

Reflections and Future Outlook

Despite promising advances, several challenges remain.

First, **ethical considerations** are paramount. Collecting perceptual or biometric data raises questions of privacy and consent, particularly in real-time adaptive systems where user behavior is continuously monitored. Transparency about data collection and use is essential to maintain trust.

Second, **algorithmic bias** poses risks. CNNs trained on datasets dominated by Western architectural images may produce distorted evaluations when applied in non-Western contexts. This not only limits generalizability but also risks perpetuating cultural inequalities. Expanding training datasets to include diverse architectural traditions is therefore critical.

Third, the issue of **interpretability** remains unresolved. Deep learning models often function as "black boxes," providing predictions without clear explanations of how those predictions were derived. For architecture—a field where meaning, symbolism, and user experience are central—opaque models may be of limited practical use. Ongoing work in explainable AI (XAI) offers potential solutions, but further research is needed to make these approaches accessible to design practitioners.

Looking ahead, three directions appear particularly promising:

1. **Explainable AI frameworks** that make algorithmic outputs interpretable for architects and stakeholders.

2. **Cross-cultural and multisensory datasets** that broaden the scope of evaluation beyond visual cues and Western contexts.

3. **Adaptive design systems** that integrate AI-driven feedback into real-time building performance, creating environments that dynamically respond to occupants' needs.

These directions reinforce the central argument of this paper: deep learning holds transformative potential but must be embedded within a broader framework that respects the interpretive richness of subjective human experience.

In addition to highlighting key studies, it is useful to examine the broader research trend. Over the past decade, the intersection of deep learning, architecture, and psychology has seen a sharp increase in scholarly attention. Figure 3 illustrates the approximate growth of publications in this domain between 2013 and 2025, reflecting its scientific currency and expanding relevance.

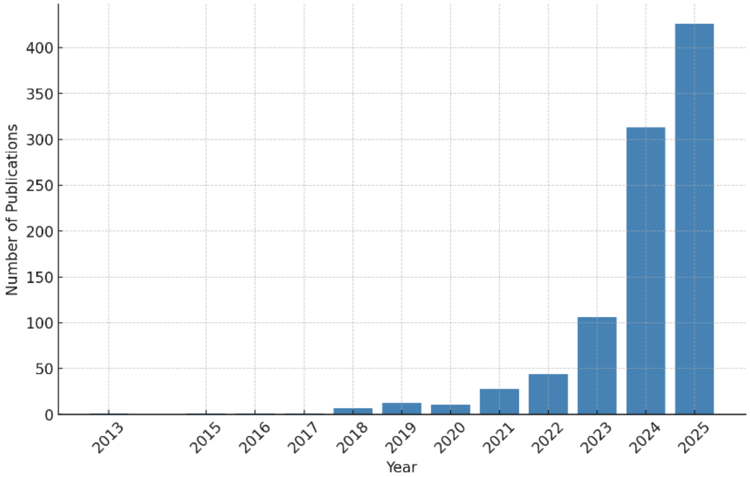


Figure 6. Growth of Publications on AI and Architectural Psychology (2013–2025)

Figure 6 illustrates the trend of publications related to deep learning in architecture and psychology between 2013 and 2025. The data, derived from Publish or Perish (Google Scholar), reveal a striking upward trajectory. From 2013 to 2017, the number of publications was negligible, with only a handful of studies emerging. Beginning in 2018, research activity slowly increased, reaching a modest but steady growth until 2021.

A turning point occurred in 2022, where the number of publications nearly doubled compared to previous years, signaling growing interest in the intersection of deep learning and architectural psychology. The most dramatic rise is observed in 2023–2025: publications surged to over 300 in 2024 and exceeded 400 in 2025, even though the year was not yet complete at the time of data collection.

This sharp increase highlights two important points. First, the field of architectural psychology is being rapidly reshaped by computational and AI-driven methods. Second, the global research community has recently recognized the significance of integrating subjective evaluation with machine learning, making it a “hot topic” in architectural and environmental studies.

The publication surge also indicates an expanding interdisciplinary dialogue, drawing from computer science, cognitive psychology, and architecture. For the purposes of this review, the trend underscores the timeliness and relevance of examining deep learning approaches to the subjective evaluation of architectural spaces.

Conclusion

This paper reviewed the emerging applications of deep learning in the subjective evaluation of architectural spaces. The analysis reveals that while AI-driven models, particularly convolutional neural networks, have shown significant promise in predicting visual preferences, assessing environmental qualities, and modeling stress recovery, they should not be treated as substitutes for human-centered approaches in architecture. Instead, they are most valuable when positioned as tools that augment and extend the interpretive richness of architectural psychology.

Three thematic domains were identified throughout the review. First, in the area of **aesthetic and visual evaluation**, CNNs demonstrated their ability to analyze building façades and daylight conditions, aligning computational predictions with human judgments. Second, in the domain of **well-being and stress recovery**, deep learning enhanced traditional environmental psychology research by integrating biometric measures, thereby producing more nuanced insights into the restorative potential of built environments. Third, in **hybrid approaches**, scholars stressed the necessity of integrating computational precision with qualitative inquiry, ensuring that cultural, phenomenological, and contextual dimensions remain central to architectural evaluation.

Despite these advances, several limitations remain. Ethical and privacy concerns regarding perceptual and biometric data, biases embedded in training datasets, and the "black box" opacity of deep learning models present serious challenges. Furthermore, AI-driven evaluations risk reductionism, oversimplifying deeply embedded cultural and symbolic aspects of architectural experience.

The way forward lies in **hybrid methodologies** that combine the scalability and efficiency of deep learning with the interpretive depth of human-centered research. Future studies should emphasize explainable AI frameworks, the development of cross-cultural and multisensory datasets, and audience-based comparative approaches that evaluate human versus AI-generated designs. Such methods not only promise more robust scientific outcomes but also ensure that technological advances remain grounded in human values and cultural contexts.

Ultimately, the contribution of this paper is to situate deep learning as a transformative yet incomplete tool for architectural psychology. By acknowledging its strengths and limitations, the review calls for a balanced integration of AI with established traditions of subjective evaluation. In this way, architecture can benefit from computational innovation while remaining faithful to its humanistic mission: designing spaces that resonate with lived experience, support well-being, and reflect cultural identity.

References

Zhou, Y., Xu, Y., & Li, B. (2022). Deep learning-based analysis of urban public space perception: Integrating street view images and user experience. *Cities, 123,* 103563.

Banaei, M., Yazdanfar, A., Nooreddin, M., & Heydari, A. (2017). EEG correlates of restorative environments: A comparison between natural and urban scenes. *Frontiers in Psychology, 8,* 1608.

Barrett, P., Zhang, Y., Moffat, J., & Kobbacy, K. (2015). A holistic, multi-level analysis identifying the impact of classroom design on pupils' learning. *Building and Environment, 89,* 118–133. https://doi.org/10.1016/j.buildenv.2015.02.013

Chamilothori, K., Wienold, J., & Andersen, M. (2019). Adequacy of immersive virtual reality for the perception of daylit spaces: Comparison of real and virtual environments. *LEUKOS, 15*(2–3), 203–226. https://doi.org/10.1080/15502724.2017.1404918

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Gifford, R. (2007). *Environmental psychology: Principles and practice* (4th ed.). Colville, WA: Optimal Books.

Kang, J., Aletta, F., Gjestland, T. T., Brown, L. A., Botteldooren, D., Schulte-Fortkamp, B., … & Axelsson, Ö. (2016). Ten questions on the soundscapes of the built environment. *Building and Environment, 108,* 284–294. https://doi.org/10.1016/j.buildenv.2016.08.011

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (Vol. 25, pp. 1097–1105).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Liu, Z., Qiu, L., & Jiang, B. (2021). Machine learning and human perception of urban building façades. *Computers, Environment and Urban Systems, 87,* 101600. https://doi.org/10.1016/j.compenvurbsys.2021.101600

Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology.* Cambridge, MA: MIT Press.

Montello, D. R. (2007). The contribution of space syntax to a comprehensive theory of environmental psychology. In *Proceedings of the 6th International Space Syntax Symposium*. Istanbul, Turkey.

Nasar, J. L. (1994). Urban design aesthetics: The evaluative qualities of building exteriors. *Environment and Behavior, 26*(3), 377–401. https://doi.org/10.1177/0013916594026003005

Norberg-Schulz, C. (1980). *Genius loci: Towards a phenomenology of architecture.* New York, NY: Rizzoli.

Ulrich, R. S., Simons, R. F., Losito, B. D., Fiorito, E., Miles, M. A., & Zelson, M. (1991). Stress recovery during exposure to natural and urban environments. *Journal of Environmental Psychology, 11*(3), 201–230. https://doi.org/10.1016/S0272-4944(05)80184-7

Artificial Intelligence Impact on Sdg 7 Affordable and Clean Energy - Scientometric Analysis

Ayah Hamed Mohamed Obadi[1]

Abstract

The global transition toward a sustainable energy future is fundamentally dependent on the ability to manage and optimize complex power systems. This research addresses this challenge by examining how Artificial Intelligence (AI) can serve as a transformative tool in advancing Sustainable Development Goal 7 (SDG 7) to ensure universal access to clean and affordable energy. To understand this dynamic relationship, a rigorous scientometric analysis is employed using the Biblioshiny application and R Studio to reveal publication trends, citation patterns, collaboration networks, and thematic developments from 2021 to 2025. This process has enabled the identification of emerging research hotspots and global collaboration patterns, with particular attention to key technological themes, including the Internet of Things (IoT), Digital Twin (DT) applications.

The findings reveal a distinctive geography of contributions starting with India which serves as a dominant hub supported by highly active institutions such as the Amrita School of Business and the Amrita School of Computing. This is further highlighted by its country-level collaboration network, which shows extensive partnerships with European countries such as Germany, the Netherlands, and the United Kingdom, as well as Asian countries like Japan. Furthermore, other regions form smaller localized clusters, such as the strong regional collaboration in Asia between Turkey, Pakistan, Malaysia, and China. In contrast, countries such as Spain, Japan, and the United Kingdom demonstrate a quality-driven

[1] Necmettin Erbakan University, Institute of Social Sciences, Business Administration Department, 21811101010@ogr.erbakan.edu.tr, Orcid: 0000-0002-2012-8787

impact, contributing fewer but more highly cited works that underscore the distinction between productivity-led and impact-led scholarly leadership.

Ultimately, this study's comprehensive scientometric analysis stands as a definitive roadmap for the field. By moving beyond a simple literature review, it not only advances academic understanding of AI's pivotal contribution to SDG 7 but also furnishes policymakers and institutions with the strategic imperatives needed to make evidence-based decisions. This work is critical to accelerating the transition toward a more efficient, reliable, and equitable global energy infrastructure!

Keywords: Artificial Intelligence (AI), Sustainable Development Goals (SDGs), SDG 7 Affordable and Clean Energy, Renewable Energy, Scientometric Analysis.

Yapay Zekanin Sdg 7 Ekonomik Ve Temiz Enerji Üzerindeki Etkisi - Scientometrik Analiz

Özet

Sürdürülebilir bir enerji geleceğine doğru küresel geçiş, temelde karmaşık güç sistemlerini yönetme ve optimize etme becerisine bağlıdır. Bu araştırma, Yapay Zeka'nın (YZ), temiz ve uygun fiyatlı enerjiye evrensel erişimi sağlamak için Sürdürülebilir Kalkınma Hedefi 7'nin (SKH 7) ilerletilmesinde nasıl dönüştürücü bir araç olarak hizmet edebileceğini inceleyerek bu zorluğun üstesinden gelmektedir. Bu dinamik ilişkiyi anlamak için, Biblioshiny uygulaması ve R Studio kullanılarak titiz bir bilimsel ölçüm analizi uygulanmış ve 2021'den 2025'e kadar yayın trendleri, atıf kalıpları, iş birliği ağları ve tematik gelişmeler ortaya çıkarılmıştır. Bu süreç, Nesnelerin İnterneti (IoT) ve Dijital İkiz (DT) uygulamaları da dahil olmak üzere temel teknolojik temalara özellikle dikkat edilerek, ortaya çıkan araştırma odak noktalarının ve küresel iş birliği kalıplarının belirlenmesini sağlamıştır.

Bulgular, Amrita İşletme Okulu ve Amrita Bilgisayar Okulu gibi oldukça aktif kurumlar tarafından desteklenen baskın bir merkez olarak hizmet veren Hindistan'dan başlayarak, katkıların belirgin bir coğrafyada toplandığını ortaya koymaktadır. Bu durum, Almanya, Hollanda ve Birleşik Krallık gibi Avrupa ülkelerinin yanı sıra Japonya gibi Asya ülkeleriyle kapsamlı ortaklıklar gösteren ülke düzeyindeki iş birliği ağıyla daha da vurgulanmaktadır. Ayrıca, Türkiye, Pakistan, Malezya ve Çin arasındaki Asya'daki güçlü bölgesel iş birliği gibi diğer bölgeler daha küçük yerel kümeler oluşturmaktadır. Buna karşılık, İspanya, Japonya ve Birleşik Krallık gibi ülkeler, üretkenlik odaklı ve etki odaklı akademik liderlik arasındaki ayrımı vurgulayan daha az sayıda ancak daha fazla atıf alan çalışmalara katkıda bulunarak kalite odaklı bir etki göstermektedir.

Sonuç olarak, bu çalışmanın kapsamlı bilimsel ölçüm analizi, alan için kesin bir yol haritası oluşturmaktadır. Basit bir literatür taramasının ötesine geçerek, yapay zekanın Sürdürülebilir Kalkınma Hedefi 7'ye olan önemli katkısına dair akademik anlayışı ilerletmekle kalmayıp, aynı zamanda politika yapıcılara ve kurumlara kanıta dayalı kararlar almak için gereken stratejik zorunlulukları da sağlamaktadır. Bu çalışma, daha verimli, güvenilir ve adil bir küresel enerji altyapısına geçişi hızlandırmak için kritik öneme sahiptir!

Anahtar kelimeler: Yapay Zeka (YZ), Sürdürülebilir Kalkınma Hedefleri (SKH), SDG 7 Uygun Fiyatlı ve Temiz Enerji, Yenilenebilir Enerji, Siyentometrik Analiz.

The Digital Innovations for SDG 7: Leveraging IoT and Digital Twin Technologies to Advance Affordable and Clean Energy Literature Review

Rapid population growth and urbanization have increased energy and resource demands in the built environment. Therefore, balancing reduced energy consumption with maintaining indoor thermal comfort presents a complex optimization challenge. However, according to the International Energy Agency, digitalization is the first step toward effective data driven energy strategies, enhancing efficiency more than ever before. Among recent digital innovations, the Internet of Things (IoT) plays a key role by enabling smart energy management, improving responsiveness to user behaviour, and enhancing comfort. Furthermore, the Sustainable

Development Goals (SDGs), particularly SDG 7 Affordable and Clean Energy uses these technologies to contribute to global targets of expanding access to modern and reliable energy efficiency by 2030. (Husein, Mushtaha, Alsyouf, & Obaideen, 2025, p. 2)

Going furthermore, to achieve the 2030 Sustainable Development Goals requires dedicate and attention to the use of technologies, such as; IoT and Digital Twin (DT) technology. In brief, the IoT contributes to universal access to modern energy, improved efficiency, and the upgrading of energy services, particularly in developing countries. While, the Digital Twin (DT) technology, creates virtual representations of physical systems, assets, or processes, is transforming industries through predictive analysis, real time monitoring, and scenario simulation. DTs contribute directly to the United Nations Sustainable Development Goals (SDGs), particularly in sectors such as manufacturing, urban planning, healthcare, agriculture, transportation, and energy. (Raman, Pattnaik, Kumar, & Nedungadi, 2024)

The SDGs are typically grouped into four domains: Social, Economic, Environmental, and Partnership & Governance, all of which can benefit from DT applications. In the renewable energy sector, Digital Twin (DT) technology enhances the efficiency of hydropower, wind, and solar systems by enabling advanced simulation, energy production forecasting, improved grid integration, and operational optimization. (Wang & Zhao, 2025, p. 11)

Smart Cities, Circular Economy, and Renewable Energy: Integrating AI and Sustainable Systems for Cross-Cutting SDGs

By leveraging technologies such as AI, machine learning, and smart grids, cities can reduce energy waste, enhance disaster preparedness, and foster inclusive, resilient urban environments. These initiatives not only promote environmental sustainability but also yield socioeconomic benefits such as poverty alleviation, job creation, and reduced life cycle emissions, reflecting the cross-cutting nature of sustainability across multiple SDGs. (Kaiser & Deb, 2025, p. 4)

In this regard, sustainable smart cities and land use management are interlinked, as both seek to balance environmental integrity, socioeconomic development, and resilience ultimately contributing to cross cutting SDGs such as 9, 11, 12, and 13. (Zhao, Yu, & Chen, 2024, p. 2)

In the transportation sector, DTs serve as virtual replicas of infrastructure, vehicles, and traffic systems, enabling simulation, monitoring, and optimization. Their application includes route and traffic pattern optimization, predictive maintenance, and scenario based assessment of environmental impacts such as emissions and energy use in cities. (Bhatia & Kumar, 2025, pp. 2–5)

Digitalization, AI, and Circular Economy: Pathways to Sustainable Energy and Resource Efficiency Leading Countries

Malaysia maintains a diverse network of energy collaborations, spanning both regional neighbours such as Indonesia, Singapore, and Thailand and global partners, including the United Kingdom, Germany, Denmark, the United States, and China. Regional partnerships primarily address shared energy challenges in Southeast Asia, including energy access, climate change, and sustainable development, while European collaborations leverage advanced renewable energy technologies and policy expertise, notably in wind energy. Emerging partnerships with developing nations such as Nigeria, Pakistan, and Bangladesh reflect a growing focus on energy security and scalable renewable solutions, including solar and biomass technologies. (Didane, Manshoor, Alimin, & Amin, 2025, p. 11)

Integrating Digital Technologies, Collaborative Innovation, and Green Hydrogen for Advancing SDG 7

Unsustainable development practices have contributed to climate change, environmental degradation, poverty, and social inequalities, leading the United Nations to establish 17 Sustainable Development Goals (SDGs) to guide global efforts toward economic, social, and environmental sustainability. While cleaner production and digital technologies such as AI, IoT, and blockchain have been adopted to improve efficiency and reduce environmental impacts, their integration remains limited in scope. (Johri, Joshi, Kumar, & Joshi, 2024, p. 3)

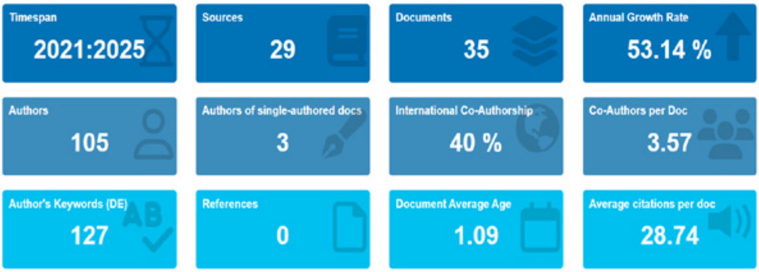
However, Intelligent systems, IoT, smart grids, and other advanced technologies are increasingly shaping human environments, with Living Labs (LLs) or Smart Labs serving as experimental ecosystems to test and validate these innovations in real life contexts. LLs function as intermediaries between citizens, researchers, companies, and institutions, fostering collaborative innovation that promotes sustainable, resilient, and health oriented environments. Their role aligns directly with the 2030 Agenda for Sustainable Development, which emphasizes measurable and applicable actions across society, the economy, and the environment. (Verdejo, Espinilla, López, & Jurado, 2022)

Energy plays a pivotal role in global economic and social development, and the rising demand highlights the urgency of transitioning from fossil fuels to renewable sources to reduce greenhouse gas emissions and environmental harm. Hydrogen, particularly when produced from renewable energy, is increasingly recognized as a key enabler of this transition due to its high efficiency, storage capacity, and potential to decarbonize multiple sectors such as transportation, industry, and electricity generation. Among its production pathways grey, blue, and green only green hydrogen, generated entirely from renewable energy, fully aligns with the objectives of SDG 7 (Affordable and Clean Energy) by providing a sustainable, scalable, and environmentally responsible energy alternative. Advancing hydrogen innovation and deployment is therefore essential for achieving universal access to clean, reliable, and modern energy while fostering long term sustainability. (Sharma, Verma, Taheri, Chopra, & Parihar, 2023)

Methodological Framework: Descriptive Scientometric Analysis on AI and SDG 7

This article utilizes the Scientometric research method, a quantitative analysis method that aims to uncover patterns in elements such as publication output, citation impact, collaboration networks, and thematic development. Scientometric analysis is a research method that aims to quantitatively measure and evaluate scientific research output through indicators such as publications, citations, and collaboration networks. Using bibliometrics and other quantitative analysis techniques, it assesses the productivity and impact of individual researchers, institutions, and scientific disciplines. (OBADI & ÇÜRÜK, 2025, p. 74)

Table 2 Scientometric Analysis Time Frame and Data Scope



Timespan 2021:2025	Sources 29	Documents 35	Annual Growth Rate 53.14 %
Authors 105	Authors of single-authored docs 3	International Co-Authorship 40 %	Co-Authors per Doc 3.57
Author's Keywords (DE) 127	References 0	Document Average Age 1.09	Average citations per doc 28.74

Timespan and Data Scope

The analysis covers publications from 2021 to 2025, with a total of 35 documents sourced from ScienceDirect.

The inclusion criteria were restricted to English language publications, comprising:

- Articles: 13
- Reviews: 9
- Book Chapters: 7
- Books: 6

This focus ensures coverage of both empirical research and theoretical and conceptual discussions relevant to AI's role in SDG7.

Source and Authorship Overview

- Sources (Journals, Books, etc.): 29
- Authors: 105 unique contributors, with only 3 single authored works, indicating strong collaboration.
- International Co-authorship Rate: 40%, reflecting significant cross-border collaboration in this domain.
- Co-authors per Document: 3.57, showing moderate team sizes typical for interdisciplinary AI energy research.

Growth Dynamics

The field shows an Annual Growth Rate of 53.14% between 2021–2025. This high growth rate indicates an emerging research hotspot, likely driven by:

- Global push for renewable energy transitions
- Rapid development of AI techniques for energy optimization, forecasting, and policy analysis
- Increased policy and funding attention to SDG7 targets

Keyword Diversity

- Author Keywords: SDG 7 Affordable and Clean Energy AND Artificial Intelligence AND Scientometric

Impact Indicators

- Average Citations per Document (28.74):

This suggests that, despite the short time span, the works are gaining substantial scholarly attention.

- Document Average Age: 1.09 years most publications are recent, further confirming the novelty of the field.

Collaboration Patterns

The combination of high international collaboration (40%) and small to medium team sizes indicates:

- A global research network leveraging AI to address energy sustainability
- Likely interdisciplinary teams combining computer science, energy engineering, and policy expertise.

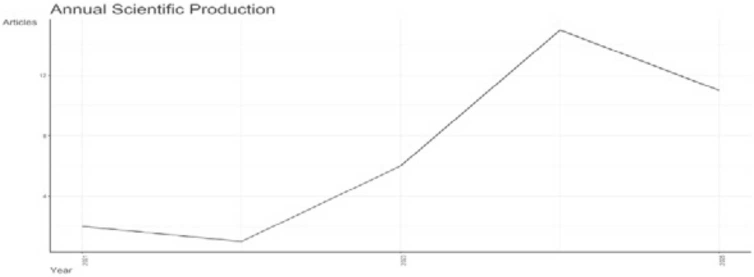


Figure 1 Annual Scientific Production

Figure 1 presents the annual scientific production on AI–SDG 7 from 2021 to 2025. The output shows an initial decline between 2021 and 2022, reaching its lowest point with fewer than two publications. However, the field experienced a strong upward trajectory from 2023, peaking in 2024 with the highest recorded output of over 13 articles. This surge indicates growing scholarly engagement and expanding research interest in the topic. The slight decline in 2025 likely reflects incomplete data capture for the year rather than a true reduction in research activity.

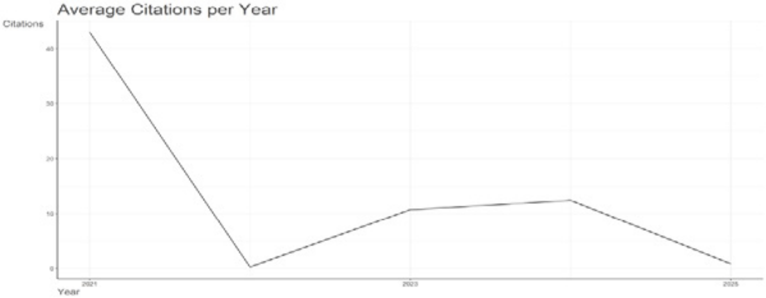


Figure 2 Average Citations Per Year

Figure 2 illustrates the trend in average citations per year for AI–SDG 7 research. The citation peak in 2021 exceeding 40 citations on average suggests that early publications in this domain were both pioneering and influential. A sharp decline followed in 2022, likely due to a surge in newer, less cited works. From 2023 onward, citation averages gradually recovered, reaching a modest high in 2024 before falling again in 2025, which is likely attributable to the recency of publications and the limited time available for citation accrual. This pattern reflects both the novelty driven impact of early studies and the natural citation lag in emerging research fields.

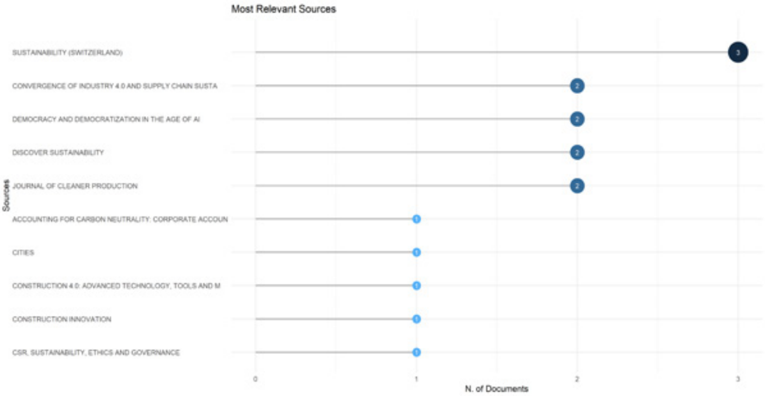


Figure 3 Most Relevant Sources

Figure 3 shows that *Sustainability (Switzerland)* is the leading outlet, contributing three papers to the AI–SDG 7 literature, underscoring its central role in sustainability research. Four other journals *Convergence of Industry 4.0 and Supply Chain Sustainability*, *Democracy and Democratization in the Age of AI*, *Discover Sustainability*, and *Journal of Cleaner Production* each published two studies, reflecting strong but more targeted engagement. The remaining sources appear only once, indicating a more dispersed yet valuable contribution from niche and interdisciplinary outlets. This pattern reveals both a concentration of research in a few core journals and the breadth of interest across diverse scholarly platforms.

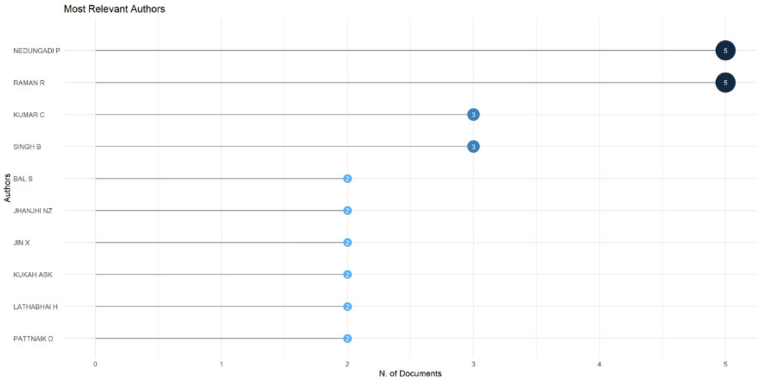
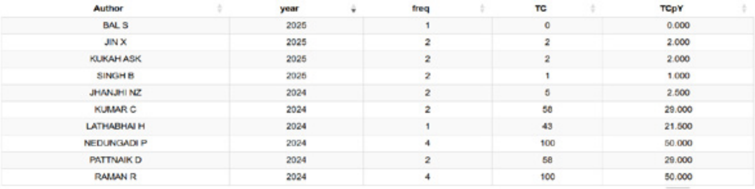


Figure 4 Most Relevant Authors

This visualization presents the most relevant authors contributing to the intersection of Artificial Intelligence (AI) and SDG 7 (Affordable and Clean Energy) based on the number of published documents. The above analysis shows that Nedungadi P and Raman R are the most prolific authors, each contributing five publications, indicating their leading role in shaping the research agenda in this domain. They are followed by Kumar C and Singh B, each with three publications, suggesting a strong but comparatively narrower engagement. A secondary tier of contributors Bal S, Jhanjhi NZ, Jin X, Kukah ASK, Lathabhai H, and Pattnaik D each have two publications, reflecting consistent but less dominant involvement.

This distribution highlights a concentration of expertise among a few highly active scholars, while a broader network of authors contributes smaller but meaningful outputs. Such a pattern suggests both established leadership in the field and a growing base of researchers engaging with AI applications to advance SDG 7 objectives, particularly in areas like renewable energy optimization, energy efficiency, and smart grid technologies.

Table 3 Authors' Production over Time



Author	year	freq	TC	TCpY
BAL S	2025	1	0	0.000
JIN X	2025	2	2	2.000
KUKAH ASK	2025	2	2	2.000
SINGH B	2025	2	1	1.000
JHANJHI NZ	2024	2	5	2.500
KUMAR C	2024	2	58	29.000
LATHABHAI H	2024	1	43	21.500
NEDUNGADI P	2024	4	100	50.000
PATTNAIK D	2024	2	58	29.000
RAMAN R	2024	4	100	50.000

The table underscores the most prolific and influential scholars at the intersection of Artificial Intelligence and SDG 7. In 2024, several authors exhibited substantial scholarly impact, most notably Nedungadi P and Raman R, each producing four publications with a combined total of 100 citations, yielding the highest TCpY value (50.0). Likewise, Kumar C and Pattnaik D contributed two publications apiece, both achieving 58 citations and TCpY scores of 29.0, underscoring their sustained academic visibility and influence. Notably, Lathabhai H achieved 43 citations from a single publication, signifying considerable impact per article.

In contrast, the 2025 contributions represented by authors such as Bal S, Jin X, Kukah ASK, and Singh B reflect ongoing research momentum but are accompanied by comparatively lower citation counts, a consequence of their recent publication and limited citation window. Taken together, the findings indicate that 2024 represented a pivotal year of high impact scholarship, consolidating the position of certain authors as intellectual leaders in the field, while the 2025 outputs highlight emerging contributors whose long term influence remains in development.

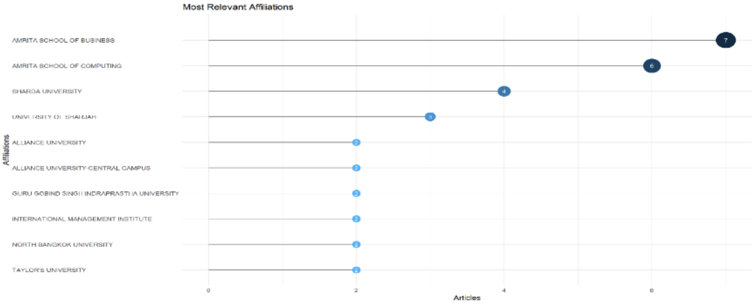


Figure 5 Most Relevant Affiliations

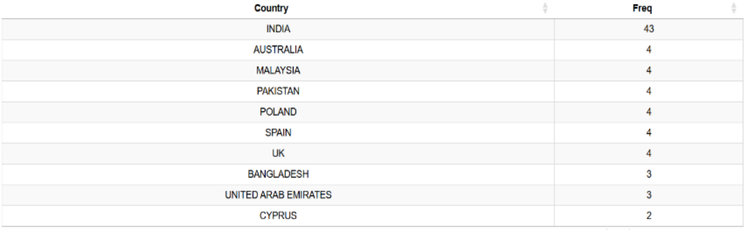
Institutional Leadership and Global Collaboration in AI–SDG 7

The visualization of institutional affiliations demonstrates the most active contributors to research at the intersection of Artificial Intelligence and SDG 7 (Affordable and Clean Energy). Amrita School of Business and Amrita School of Computing emerge as the leading institutions, with 7 and 6 publications respectively, reflecting a concentrated research focus and institutional strength in this domain.

Sharda University (4 publications) and the University of Sharjah (3 publications) also represent important hubs of scholarly activity, reinforcing the role of both Indian and Middle Eastern institutions in advancing this field. Other contributors, including Alliance University, Guru Gobind Singh Indraprastha University, International Management Institute, North Bangkok University, and Taylor’s University, display moderate research engagement with 1–2 publications each.

This distribution highlights the dominance of Indian institutions, particularly Amrita, as central actors shaping the intellectual landscape of AI driven clean energy research, while simultaneously pointing to a growing global collaboration with universities from the UAE, Thailand, and Malaysia. This institutional spread indicates both concentrated leadership and increasing international diversification in the scholarly discourse.

Table 4 Country level distribution of publications



Country	Freq
INDIA	43
AUSTRALIA	4
MALAYSIA	4
PAKISTAN	4
POLAND	4
SPAIN	4
UK	4
BANGLADESH	3
UNITED ARAB EMIRATES	3
CYPRUS	2

The country level distribution of publications reveals a clear geographical concentration in the research on Artificial Intelligence and SDG 7 (Affordable and Clean Energy). India dominates the landscape with 43 publications, underscoring its strong scholarly leadership and institutional focus in this domain.

In contrast, other countries such as Australia, Malaysia, Pakistan, Poland, Spain, and the UK each contribute four publications, reflecting a secondary but significant level of engagement. Meanwhile, Bangladesh and the United Arab Emirates contribute three publications each, followed by Cyprus with two.

The distribution highlights both the regional concentration of intellectual output in South Asia and the growing interest from Europe, the Middle East, and Oceania, suggesting the gradual emergence of a more globally interconnected research network around AI and sustainable energy transitions.

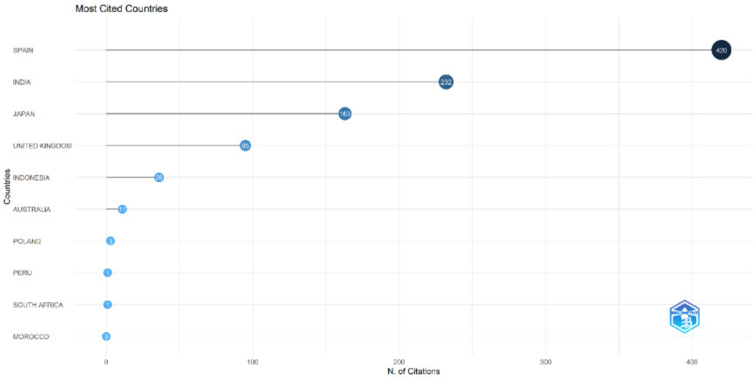


Figure 6 Most Cited Countries

This visualization shows the most cited countries in the field of research on Artificial Intelligence and Sustainable Development Goals (SDG 7 – Affordable and Clean Energy).

- Spain leads with 420 citations, making it the most influential country in terms of scholarly impact, despite not having the highest publication volume.
- India, with 232 citations, emerges as both a major contributor in publication volume (as seen earlier) and a highly impactful player in terms of citation impact.
- Japan (163 citations) and the United Kingdom (95 citations) follow, highlighting their substantial influence in shaping the research discourse.
- Other countries such as Indonesia (39), Australia (11), Poland (3), Peru (1), South Africa (1), and Morocco (1) also contribute, though with lower citation counts, reflecting a more peripheral yet growing role in the global research network.

This comparison highlights a critical distinction: while India dominates in research productivity, Spain excels in citation influence, showing that scholarly impact is not always aligned with publication volume. This suggests Spain's research outputs are more widely referenced and integrated into the global knowledge base, whereas India's high publication output is building a strong foundation for future influence.

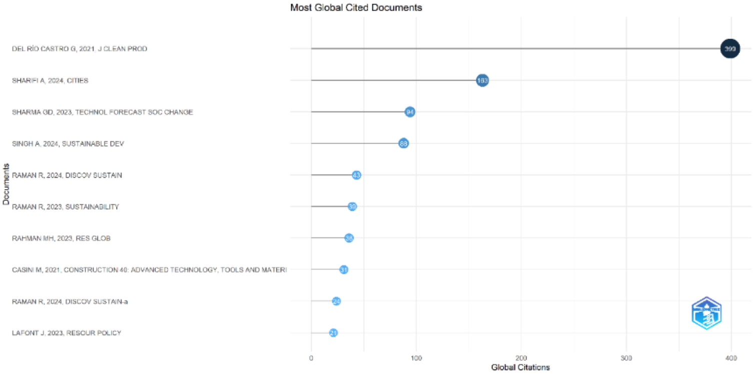


Figure 7 Most Global Cited Documents

The figure illustrates the most globally cited documents at the intersection of Artificial Intelligence and SDG 7 (Affordable and Clean Energy). Del Río Castro G. (2021, Journal of Cleaner Production) is the most influential publication, with 399 citations, positioning it as the intellectual anchor of the field. This is followed by Sharifi A. (2024, Cities) with 163 citations and Sharma G.D. (2023, Technological Forecasting & Social Change) with 94 citations, both of which demonstrate substantial scholarly impact.

Mid-level contributions include Singh A. (2024, Sustainable Development) with 88 citations and multiple works by Raman R. (2023–2024), whose publications collectively accumulate strong visibility (ranging between 20 and 48 citations each), reinforcing his role as a consistent and recurrent contributor. Other notable works include Rahman M.H. (2023, Resources Global) with 42 citations, Casini M. (2021) with 31 citations, and Lafont J. (2023) with 21 citations, all of which highlight thematic diversity in research applications.

Overall, the distribution reflects a concentration of high impact contributions in a small set of foundational works, with Del Río Castro (2021) dominating citation influence, while a wider group of scholars including Sharifi, Sharma, Singh, Raman, and Rahman provide ongoing momentum and breadth to the research domain. This pattern underscores how early seminal publications continue to shape the intellectual structure of the field, while recent works, particularly those published in 2023–2024, signal emerging directions and sustained scholarly engagement.

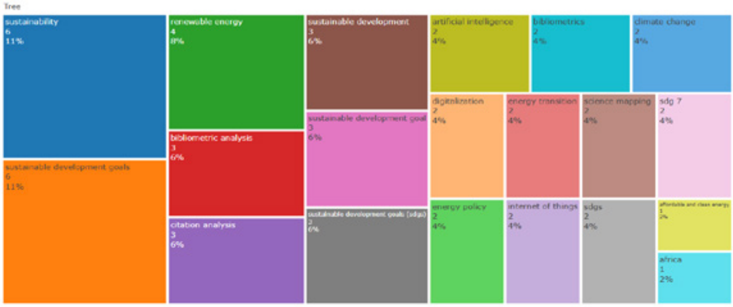


Figure 8 Tree map Visualization Chart

The tree map visualization presents the distribution of key research themes at the intersection of Artificial Intelligence (AI) and SDG 7 (Affordable and Clean Energy). The largest thematic clusters are “sustainability” (11%) and “sustainable development goals” (11%), underscoring the centrality of broad sustainability discourses in shaping this research field. Closely following are renewable energy (8%), and themes such as sustainable development, bibliometric analysis, and citation analysis (each 6%), highlighting the methodological orientation of the literature and its alignment with global sustainability agendas.

Smaller but critical clusters include artificial intelligence, bibliometrics, climate change, digitalization, energy transition, science mapping, internet of things, and SDG 7 (each 4%), reflecting the emerging technological and analytical lenses applied to advancing clean energy transitions. Of particular importance, SDG 7 (4%) and affordable and clean energy (2%) appear as explicit research foci, signalling that while still developing, there is growing scholarly attention on operationalizing AI within energy systems to advance this goal. Additionally, energy policy and Africa appear as smaller but relevant clusters, pointing to contextual and regional dimensions of the discourse.

Collectively, the structure of the tree map reveals a dual emphasis: first, on conceptual and methodological foundations (sustainability, bibliometric/citation analyses); and second, on technological applications and sectoral transitions (AI, IoT, renewable energy, digitalization). This suggests that the field is moving toward integrating advanced technologies with sustainability science, while progressively narrowing its focus on SDG 7 and clean energy solutions.

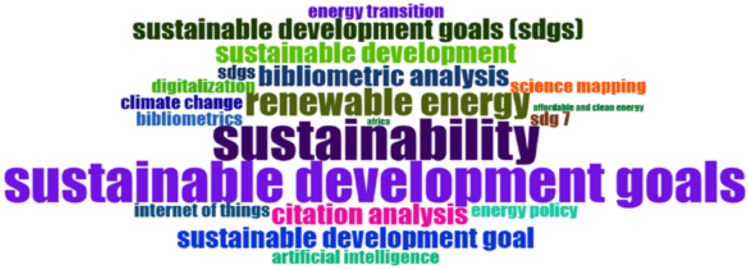


Figure 9 Cloud map of used keywords

Using the cloud map shown after doing the Scientometric analysis in Biblioshiny application, demonstrates a convergence of sustainability development goals, renewable energy, internet of things, citation analysis, climate change, energy policy, artificial intelligence, bibliometric analysis, digitalization, and clean energy transitions, with AI and IoT positioned as enabling tools to accelerate progress toward affordable, reliable, and science mapping. All these key words show the importance of the research topic and their relationship and impact together. Eventually, they reflect the importance of keywords based on their sizes that has been highlighted in the above figure, that reveals the flow and combination of concepts.

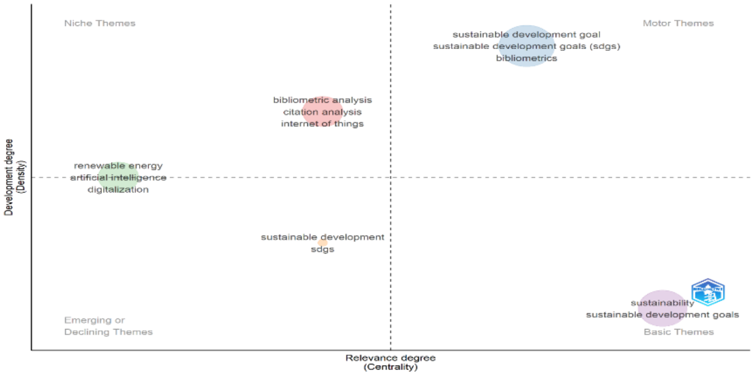


Figure 10 Thematic map

This figure is a thematic map generated in Scientometric analysis that organizes research themes based on two dimensions:

- Development degree (Density): how internally developed and mature a theme is.
- Relevance degree (Centrality): how important and connected a theme is to the overall research field.

1. Motor Themes (upper right) → Well developed and highly relevant.

o *Sustainable Development Goals (SDGs), bibliometrics* are here, meaning they are central and strongly drive the research field.

2. Niche Themes (upper left) → Well developed but less connected.

o *Bibliometric analysis, citation analysis, Internet of Things* appear here, showing specialized but not broadly connected topics.

3. Emerging or Declining Themes (lower left) → Weakly developed and marginal.

o *Renewable energy, artificial intelligence, digitalization* suggests these are either upcoming topics that need more integration or areas losing momentum in this context.

4. Basic Themes (lower right) → Highly relevant but not well developed.

o *Sustainability, SDGs* are foundational but still require more conceptual and methodological development to advance the field.

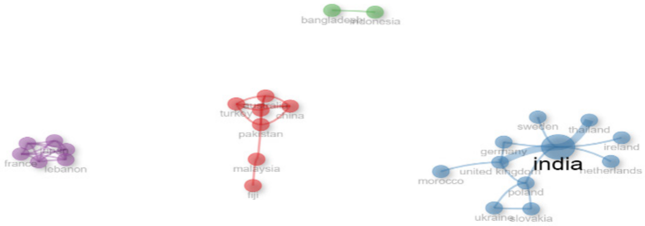


Figure 11 Country Level Collaboration Network

The country level collaboration network highlights India as a dominant hub, engaging in extensive partnerships with both European and Asian countries. In contrast, other regions form smaller, localized clusters: Turkey–Pakistan–Malaysia–China represent a strong regional collaboration in Asia, while France–Lebanon–Japan form a distinct partnership cluster. Bangladesh and Indonesia remain relatively isolated with limited bilateral cooperation.

- Major cluster (blue): India is the most central country, collaborating with the UK, Germany, Netherlands, Ireland, Sweden, Poland, Morocco, Thailand, Ukraine, and Slovakia.
- Red cluster: Turkey, Pakistan, Malaysia, China, and Fiji are strongly connected.
- Purple cluster: France and Lebanon collaborate closely with Japan.
- Green cluster: Bangladesh and Indonesia form a small dyad.

Bridging Technology and Sustainability: Scientometric Exploration of AI's Role in SDG 7 Summary

This study adopts a Scientometric research design to analyse the evolution and intersection of Artificial Intelligence (AI) and Sustainable Development Goal 7 (Affordable and Clean Energy). Covering the timespan 2021–2025, the dataset consists of 35 English language publications retrieved from ScienceDirect, comprising 13 articles, 9 reviews, 7 book chapters, and 6 books. This balanced scope ensures representation of both empirical research and theoretical contributions, allowing for a comprehensive evaluation of how AI driven technologies are being mobilized to support energy transitions and sustainability objectives. The methodological framework applies descriptive Scientometric techniques, enabling the identification of publication trends, citation patterns, authorship dynamics, collaboration networks, and thematic structures. Through this approach, the analysis highlights the productivity and scholarly impact of researchers, institutions, and countries engaged in this domain.

The findings indicate a distinctive geography of contributions. India emerges as the dominant producer of research with 43 publications, supported by highly active institutions such as the Amrita School of Business and the Amrita School of Computing. Other contributors, including Sharda University and the University of Sharjah, demonstrate regional significance, while additional institutions across Southeast Asia and the Middle East reflect growing diversification. Spain, in contrast, demonstrates the highest citation impact (420 citations) despite modest output, underscoring the distinction between productivity led and impact led scholarly leadership. Similar patterns are observed for Japan (163 citations) and the United Kingdom (95 citations), each demonstrating considerable influence with relatively fewer publications. Regional collaboration clusters, such as Turkey, Pakistan, Malaysia, China, France, Lebanon and Japan, further illustrate the emergence of both global hubs and localized research partnerships.

Authorship analysis reveals a concentrated group of prolific contributors driving this field. Among them, Nedungadi P. and Raman R. occupy leading positions, each producing between four and five publications with a combined citation impact of 100 (TCpY = 50.0). Their consistent presence positions them as intellectual anchors shaping the research agenda. They are followed by Kumar C. and Singh B. with three publications each, as well as Kumar C. and Pattnaik D., who produced two highly cited publications apiece (58 citations, TCpY = 29.0). A secondary tier of scholars—including Bal S., Jhanjhi N.Z., Jin X., Kukah A.S.K., and Lathabhai H. contributes two publications each, reflecting steady but less dominant involvement. Collectively, these authors provide the intellectual backbone of the field, with Nedungadi and Raman at its forefront.

The thematic structure of the literature reflects both conceptual and technological orientations. Sustainability (11%) and Sustainable Development Goals (11%) constitute the largest clusters, followed by renewable energy (8%), bibliometric and citation analysis (6%), and emerging themes such as AI, IoT, climate change, digitalization, and energy transition (each 4%). Affordable and clean energy, although smaller in representation (2%), appears as an explicit and growing focus, indicating that research is progressively moving from broad sustainability frameworks toward applied and operationalized discussions of AI's role in energy systems. This dual emphasis on foundational sustainability concepts and applied technological solutions demonstrates the integrative trajectory of the field.

Intellectual impact is shaped by a small number of highly cited works. Del Río Castro (2021, *Journal of Cleaner Production*) anchors the field with 399 citations, followed by Sharifi (2024, *Cities*, 163 citations) and Sharma (2023, *Technological Forecasting & Social Change*, 94 citations). Other influential publications include Singh (2024, *Sustainable Development*, 88 citations) and a series of contributions by Raman (20–48 citations each), whose recurrent scholarship reinforces his prominence. Additional works by Rahman (42 citations), Casini (31 citations), and Lafont (21 citations) broaden the thematic scope, addressing diverse applications of AI in the clean energy context.

Moreover, India demonstrates quantity driven leadership, producing the majority of publications through concentrated institutional and authorial networks. On the other hand, Spain, Japan, and the United Kingdom illustrate quality driven impact, contributing fewer but more highly cited works. The thematic distribution underscores the balance between conceptual sustainability discourses and the applied use of AI, IoT, and digitalization for energy transitions. Ultimately, this Scientometric analysis highlights an emerging but fast-growing research niche, where the integration of advanced technologies with sustainability science is beginning to coalesce into a distinct intellectual agenda!

Refrences

Bhatia, M., & Kumar, R. (2025). Digital Twin and sustainability: A data-driven scientometric exploration. *Internet of Things*, *32*, 101652. https://doi.org/10.1016/J.IOT.2025.101652

Didane, D. H., Manshoor, B., Alimin, A. J., & Amin, A. A. (2025). Bibliometric analysis of 50 years of energy research in Malaysia: Trends and opportunities. *Renewable and Sustainable Energy Reviews*, *214*, 115498. https://doi.org/10.1016/J.RSER.2025.115498

Husein, L. A., Mushtaha, E. S., Alsyouf, I., & Obaideen, K. (2025). Advancing SDGs with IoT: Enhancing thermal comfort and energy efficiency- A bibliometric study. *Sustainable Futures*, *10*, 100873. https://doi.org/10.1016/J.SFTR.2025.100873

Johri, A., Joshi, P., Kumar, S., & Joshi, G. (2024). Metaverse for Sustainable Development in a bibliometric analysis and systematic literature review. *Journal of Cleaner Production*, *435*, 140610. https://doi.org/10.1016/J.JCLEPRO.2024.140610

Kaiser, Z. R. M. A., & Deb, A. (2025). Sustainable smart city and Sustainable Development Goals (SDGs): A review. *Regional Sustainability*, *6*(1), 100193. https://doi.org/10.1016/J.REGSUS.2025.100193

OBADI, A., & ÇÜRÜK, S. (2025). *Ulusal Tez Merkezi | Anasayfa*. Retrieved 09/13/2025 from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Raman, R., Pattnaik, D., Kumar, C., & Nedungadi, P. (2024). Advancing sustainable energy systems: A decade of SETA research contribution to sustainable development goals. *Sustainable Energy Technologies and Assessments*, *71*, 103978. https://doi.org/10.1016/J.SETA.2024.103978

Sharma, G. D., Verma, M., Taheri, B., Chopra, R., & Parihar, J. S. (2023). Socio-economic aspects of hydrogen energy: An integrative review. *Technological Forecasting and Social Change, 192*, 122574. https://doi.org/10.1016/J.TECHFORE.2023.122574

Verdejo, Á., Espinilla, M., López, J. L., & Jurado, F. (2022). Assessment of sustainable development objectives in Smart Labs: technology and sustainability at the service of society. *Sustainable Cities and Society, 77*, 103559. https://doi.org/10.1016/J.SCS.2021.103559

Wang, G., & Zhao, B. (2025). Research hotspots and trends in sustainable development goals. *Environmental and Sustainability Indicators, 26*, 100722. https://doi.org/10.1016/J.INDIC.2025.100722

Zhao, Q., Yu, L., & Chen, X. (2024). Land system science and its contributions to sustainable development goals: A systematic review. *Land Use Policy, 143*, 107221. https://doi.org/10.1016/J.LANDUSEPOL.2024.107221

PART II

AI, ETHICS, AND GOVERNANCE

Normative, Political, and Philosophical Contexts

Section 2.1

Human Values and Philosophical Foundations

***

Artificial Intelligence and the Human Quest for Freedom: A Critical Inquiry into Techno-Lives

Elvije Kadrija[1]

Abstract

The increasing incorporation of artificial intelligence into nearly every aspect of modern life has constitutionally reshaped human existence, giving rise to what can now be reported as "techno-lives." These techno-lives, noticed by the seamless interactivity between humans and intelligent technologies, redefine not only how we work, connect, and make settlements, but also how we see ourselves and our prospects for freedom. While AI entreaties are often recognized for their offering of coherence, problem-solving, and benefits, deeper philosophical explorations raise questions about whether these technologies can effectively facilitate solving one of humanity's most durable predicaments: the issue of human liberation. Does keeping a technologically entitled life, led by AI, propose new tracks regarding distinctive and societal freedom, or does it tread us more within cosmopolitan arrangements of power and reliance?

This study points out the need to critically elaborate on whether the techno-lives shaped by AI applications can give us a solution to the human issues of freedom. Analyzing from a philosophical viewpoint on freedom, especially those that highlight autonomy, self-determination, and the shattering of organizational limitations. Therefore, this paper will continue to investigate how AI technologies may either encourage or obstruct human liberation. On one hand, AI tends to facilitate freedom by automating tedious duties,

[1] Marmara University, Social Sciences Institute, Communication Studies, elvije.kadrija@marun.edu.tr, ORCID: 0009-0002-8249-3744

allowing more time for self-realization, sustaining admission, and democratizing comprehension. Technologies such as personalized teaching platforms, AI-facilitated healthcare, and intelligent potency suspension can enable individuals by decreasing physical, educational, and socio-economic obstacles. On the other hand, identical technologies can establish new structures of reliance, observation, and algorithmic venture that subtly ruin human activity. Algorithmic decision-making methodologies, data-driven consumer targeting, and prescient analytics can compel personal options, strengthening present social variations and forming behavior in ways that are not regularly clear or arranged with human utilities. In these conditions, techno-life cannot indicate deliverance/liberation, but additionally it may be a more compounded and maybe sneakier form of modern internment. Through a multidisciplinary perspective that merges technology research, media theory, and philosophy, this paper analyzes the dialectical association between AI and freedom. This paper will be focused on critical thinkers who warn against technological determinism and highlight the significance of ethical patterns, user ascendancy, and digital literacy in creating human-technology associations. Furthermore, real-world samples of AI networks in governance, institutions, healthcare, and social media are examined to explain both the emancipatory and maybe brutal aspects of techno-life.

In the end, this research contends that AI and techno-lives, in themselves, are neither essentially delivering nor inherently captivating. Their influence on human liberation relies mostly on the societal, ethical, and governmental substructures within which these technologies are instigated and distributed. The way to freedom via AI is not instinctive but dependent on responsible human options, critical commitment, and the initiation of methodologies that prioritize human dignity and consensual well-being. In this condition, the question of whether techno-life has the possibility of giving a settlement to the issues of human deliverance remains unlatched but heavily remarkable for the future of human society.

Keywords: *Artificial Intelligence, Techno-Life, Freedom, Algorithmic Decision-Making, Digitallization*

Yapay Zeka ve İnsanın Özgürlük Arayışı: Tekno-Yaşamlara İlişkin Kritik Bir Araştırma

Özet

Yapay zekânın modern yaşamın neredeyse her alanına giderek daha fazla entegre edilmesi, insan varoluşunu temelinden yeniden şekillendirmiş ve artık "tekno-hayatlar" olarak adlandırılabilecek bir yaşam biçiminin ortaya çıkmasına yol açmıştır. Bu tekno-hayatlar, insanlar ile akıllı teknolojiler arasındaki kesintisiz etkileşimle kendini gösterir ve yalnızca nasıl çalıştığımızı, nasıl bağlantı kurduğunu ve nasıl yerleşimler oluşturduğumuzu değil, aynı zamanda kendimizi nasıl gördüğümüzü ve özgürlük umutlarımızı da yeniden tanımlar. Yapay zekâ uygulamaları çoğunlukla verimlilik, problem çözme ve çeşitli faydalar sunduğu için takdir edilse de, daha derin felsefi sorgulamalar bu teknolojilerin insanlığın en eski sorunlarından biri olan özgürleşim meselesine gerçekten bir çözüm getirip getirmeyeceğini sorgular. Yapay zekâ tarafından yönlendirilen bir teknoloji merkezli yaşam, bireysel ve toplumsal özgürlükler adına yeni yollar mı açmaktadır yoksa bizi küresel güç ve bağımlılık düzeylerinin içine daha mı fazla çekmektedir?

Bu çalışma, yapay zekâ uygulamalarıyla şekillenen tekno-hayatların insan özgürlüğü sorununa bir çözüm getirip getirmeyeceğini eleştirel bir şekilde tartışma gerekliliğine dikkat çekmektedir. Özellikle özerklik, öz-belirleme ve yapısal sınırların kırılması üzerine yoğunlaşan felsefi özgürlük yaklaşımlarını temel alarak bir analiz yapılacaktır. Bu çerçevede makale, yapay zekâ teknolojilerinin insan özgürleşimini teşvik edip etmediğini ya da buna engel olup olmadığını araştırmayı amaçlamaktadır. Bir yandan yapay zekâ; sıkıcı işlerin otomatikleştirilmesi, bireyin kendini gerçekleştirmesi için daha fazla zaman yaratılması, bilgiye erişimin genişletilmesi ve anlayışın demokratikleşmesi yoluyla özgürlüğü destekleyebilir. Kişiselleştirilmiş eğitim platformları, yapay zekâ destekli sağlık hizmetleri ve akıllı enerji yönetimi gibi teknolojiler, fiziksel, eğitsel ve sosyo-ekonomik engelleri azaltarak bireyleri güçlendirebilir. Öte yandan, aynı teknolojiler yeni bağımlılık yapıları, gözetim mekanizmaları ve algoritmik kontrol biçimleri yaratarak insan özerkliğini sinsi bir şekilde zayıflatabilir.

Algoritmik karar alma süreçleri, veri odaklı tüketici hedeflemeleri ve öngörüsel analizler, bireysel seçimleri manipüle edebilir, mevcut toplumsal eşitsizlikleri derinleştirilir ve insan davranışlarını çoğu zaman fark edilmeden yönlendirebilir. Bu bağlamda, tekno-hayat özgürlük vaat etmek bir yana, modern bir mahpusluk biçimi de olabilir. Bu çalışma, teknoloji araştırmaları, medya teorisi ve felsefeyi bir araya getiren çok disiplinli bir bakış açısıyla, yapay zekâ ve özgürlük arasındaki diyalektik ilişkiyi incelemektedir. Teknolojik determinizme karşı uyarıda bulunan ve insan-teknoloji ilişkilerinin şekillendirilmesinde etik ilkelerin, kullanıcı kontrolünün ve dijital okuryazarlığın önemini vurgulayan eleştirel düşünürler bu çalışmanın temelini oluşturacaktır. Ayrıca, yönetişim, kurumlar, sağlık hizmetleri ve sosyal medya gibi alanlardaki yapay zekâ uygulamalarına dair gerçek dünya örnekleri üzerinden, tekno-hayatın hem özgürleştirici hem de potansiyel olarak baskıcı yönleri açıklanacaktır.

Sonuç olarak bu araştırma, yapay zekâ ve tekno-hayatların doğası gereği ne kurtarıcı ne de mutlak anlamda kısıtlayıcı olduğunu öne sürmektedir. Bu teknolojilerin insan özgürlüğü üzerindeki etkisi, büyük ölçüde onların hangi toplumsal, etik ve politik yapı içinde geliştirildiğine ve uygulandığına bağlıdır. Yapay zekâ yoluyla özgürlüğe ulaşmak, kendiliğinden gelişen bir süreç değil; sorumlu insan tercihleri, eleştirel katılım ve insan onuru ile kolektif iyiliği önceliklendiren yöntemlerin benimsenmesine bağlıdır. Bu durumda, tekno-hayatın insan özgürlüğü sorununa bir çözüm sunup sunamayacağı sorusu açık kalmakta; ancak insanlığın geleceği açısından derinlemesine sorgulanmayı hak eden bir mesele olarak önemini korumaktadır.

Anahtar Kelimeler: *Yapay Zekâ, Tekno-Hayat, Özgürlük, Algoritmik Kararlar, Dijitalleştirme*

Introduction

Developing artificial intelligence technologies are reshaping not only technical fields but also our daily habits, social relationships, and individual existence. Artificial intelligence applications have become an integral part of our lives in a wide range of areas, from decision support systems to algorithmic content management, from personal assistants to productive software. This digital transformation is profoundly questioning the concept of freedom, one of the most fundamental pursuits of humanity throughout history. In today's world, artificial intelligence is not merely a technological tool; it also raises the question of whether it could offer a new solution to the problem of human liberation or a new form of restriction.

Since the Enlightenment, the modern concept of liberation has been associated with the individual's ability to use their own minds and liberate themselves from dependence on external authorities (Kant, 1784). However, the artificial intelligence systems individuals face today are profoundly questioning the very definition of freedom. For example, while news feeds or consumption recommendations personalized through algorithms are designed to save individuals time and facilitate choice, these recommendations actually limit and direct their choices (Zuboff, 2019). Surveillance capitalism, as conceptualized by Shoshana Zuboff, reveals how artificial intelligence and algorithms are used to predict and shape individual behavior. This demonstrates that the ideal of freedom is under serious threat in the age of artificial intelligence.

On the other hand, artificial intelligence applications also offer opportunities such as reducing routine workloads on individuals, increasing temporal efficiency, and allowing them to dedicate more time to creative activities (Floridi et al., 2018). In this context, from a certain perspective, artificial intelligence may also offer the potential to solve the problem of human liberation. Within the framework of Richard Sennett's concept of the "new capitalist society" (1998), individuals increasingly live under time pressure in constantly changing, flexible, and fragmented work conditions. It is argued that artificial intelligence applications, particularly through automation and decision support systems, can contribute to individuals' liberation from this time pressure and strengthen their self-management skills (Brynjolfsson & McAfee, 2014). This could open a new path in the pursuit of freedom: Can humans build a more autonomous and creative life using artificial intelligence as a tool? — Another important concept here is Michel Foucault's "biopolitics." According to Foucault, modern power, rather than directly oppressing individuals, exerts indirect control by regulating their lifestyles and behavioral patterns (Foucault, 1978). Artificial intelligence-supported lifestyles, on the one hand, offer individuals seemingly infinite choices, while on the other, they can determine these choices through algorithms in the background. Filiz Aydoğan's Current Concepts in Media (2023), edited by Filiz Aydoğan, offers a comprehensive analysis of how artificial intelligence and algorithms present both opportunities and limitations to the individual's process of liberation within the media ecosystem. The title "Artificial Intelligence" in the book explains the integration of technology into daily experience and emphasizes the central role that algorithmic systems play in content production and distribution (Aydoğan, 2023). In this context, artificial intelligence rapidly provides individuals with personalized information while invisibly shaping content selection through a background of data control. This structure, described in the media as the "algorithm society," appears to increase individuals' free choice, but in reality, it exposes them to a universe of standardized content (Aydoğan, 2023). While artificial intelligence applications encourage individuals to optimize their lives, they actually further integrate them into a data-driven system, transforming them into constantly monitored beings (Han, 2017).

However, artificial intelligence also has the potential to offer people a more creative and autonomous space by alleviating the burden of routine and repetitive work on the individual. As Floridi et al. (2018) point out, when used correctly, artificial intelligence can enhance human well-being, allowing individuals to focus on more meaningful tasks and contribute to self-actualization. Especially when considered within the context of Richard Sennett's (1998) critique of the new capitalist society, artificial intelligence applications can become tools that simplify the lives of individuals under time pressure. From this perspective, artificial intelligence technologies can liberate people from certain practical constraints, allowing them to develop self-management skills. While AI-supported systems appear to offer individuals numerous options, the question of how these options are constructed by algorithms in the background raises the question of whether individuals are truly free. As emphasized in Byung-Chul Han's (2017) psychopolitical approach, today's human being is transforming into a subject who constantly strives to optimize their own lives while unconsciously submitting to the pressure of digital systems.

This study will examine, from a multidimensional perspective, whether techno-lives shaped by artificial intelligence applications can offer a solution to the problem of human liberation. First, the historical and philosophical foundations of the concept of freedom will be explained, and the position of the individual in modern society will be discussed. Then, the impact of artificial intelligence applications on daily life and how they shape individual decision-making mechanisms will be examined with examples. Then, within the framework of surveillance capitalism and biopolitics, we will evaluate how artificial intelligence serves invisible forms of control. Finally, we will discuss how artificial intelligence, if developed in line with ethical and social governance principles, can contribute to the process of individual liberation, and synthesize the topic from different perspectives. In this context, our study aims to analyze the complex relationship between artificial intelligence and human freedom from a holistic perspective.

The Impact of the Techno-Lives We Live with Artificial Intelligence Applications on Human Liberation

The integration of artificial intelligence technologies into our daily lives is radically altering not only our individual experiences but also the social structure and the way individuals perceive themselves. This change is also creating a new area of debate regarding the concept of liberation. In historical and philosophical perspectives, liberation is generally defined as the individual's liberation from limitations through their own will, the ability to make independent decisions and take action (Berlin, 1969). However, the "techno-lives" shaped by artificial intelligence applications today offer new perspectives on the question of liberation, while also bringing with them some paradoxical consequences. Therefore, whether artificial intelligence technologies can be a solution to human liberation requires a comprehensive assessment of both technology policies and individual experiences.

The Industrial Revolution, modern communication technologies, and the proliferation of the internet facilitated individuals' access to information while also paving the way for the development of surveillance and control mechanisms (Foucault, 1977). Artificial intelligence is emerging as the newest and most effective tool in this process. Machine learning and algorithmic decision support systems, in particular, are automating and optimizing many areas of an individual's life. For example, the opportunities offered by artificial intelligence in areas such as accelerating diagnostic processes in healthcare, personalized learning techniques in education, and urban traffic management and security practices are improving people's quality of life and expanding their freedom.

However, this process also leads to the increasing delegation of decisions to non-human systems and limits individuals' ability to make their own choices (Zuboff, 2019).

On the other hand, biased datasets used to train algorithms or opaque decision-making mechanisms within AI increase the risk of interference with individuals' free decision-making processes (Noble, 2018). In other words, while equal access to information can be achieved, significant restrictions can arise regarding the freedom to interpret and use it.

The impact of technological tools on individual autonomy presents a more complex picture. On the one hand, AI-powered personal assistants and automation systems can take over routine tasks, leaving individuals with more time and mental space. This can allow individuals to focus more on their own interests and creative activities (Bostrom & Yudkowsky, 2014). On the other hand, the power of algorithms to shape our individual preferences, particularly in social media and advertising, raises questions about free will. While users' digital behavior is constantly analyzed and optimized, the extent to which these free choices are conscious and authentic is open to debate (Pariser, 2011). This situation can be interpreted as the individual's entrapment in the digital world, associated with concepts such as "filter bubbles" and "algorithmic coercion." The book "Current Concepts in Media" (2023) discusses the "filter bubble" effect, particularly in media environments, and argues that AI-powered systems narrow an individual's perception of reality. While users encounter streams of content tailored to their interests, these streams are manipulated beyond their conscious choices by the algorithms that manage them (Aydoğan, 2023). This situation, in turn, raises questions about free will, as individuals are unknowingly guided by AI-based systems. Therefore, the societal impacts of AI further complicate the issue of liberation.

In fact, the use of artificial intelligence technologies by states and large corporations as tools of control and surveillance poses serious threats to individual rights and privacy. Examples such as the social credit system in China and surveillance technologies in the US demonstrate that individuals' behavior is constantly monitored and controlled through technology (Harari, 2018). Such practices can result in significant restrictions on individual freedom rather than liberation. Conversely, the development of artificial intelligence technologies through open-source, democratic, and participatory methods can also offer alternative approaches that can increase the potential for liberation. The guidance of social movements, ethical, and legal regulations on these technologies can enable liberation to be realized through technological tools (Floridi, 2019).

In terms of individual experiences, the impact of artificial intelligence on liberation is dual. Liberation is not merely the removal of external obstacles but also the process of the individual's self-actualization through their own consciousness and will (Taylor, 1991). While the conveniences and support mechanisms provided by AI can be a tool for individuals to unlock their potential, excessive dependence can lead to passivity and a weakening of will. Furthermore, the norms, standards, and behavioral patterns imposed by AI can hinder individuals' ability to express themselves authentically. Therefore, for AI to be used as a tool for liberation, it must be human-centered and aligned with ethical principles (Dignum, 2018).

Philosophical Foundations of Freedom

The concept of freedom has been one of the most fundamental and controversial issues in philosophy throughout history. From ancient Greece to modern philosophy, many thinkers have developed different approaches to what freedom is, how it is achieved, and where its limits begin and end. First, Aristotle argued that humans, as rational beings, have the capacity to direct their own actions with reason and associated this ability with freedom. According to Aristotle, being free means being able to choose a lifestyle that suits one's own purpose (Aristotle, Nicomachean Ethics). In modern philosophy, Immanuel Kant addressed freedom more systematically. For Kant, freedom, beyond being independent of external constraints, is the individual's ability to act according to the laws of their own reason. In his essay "What is Enlightenment?" (1784), Kant defines true freedom as the individual's courage to use their own reason. This perspective posits the individual's autonomy and capacity for rational choice as the fundamental conditions of freedom. On the other hand, Jean-Paul Sartre and existentialist philosophy view freedom as a radical responsibility. Sartre argues that humans are "condemned to freedom" because the individual is faced with the necessity of making choices in every situation and is entirely responsible for the consequences of these choices (Sartre, 2003). According to this perspective, human freedom is directly related to the capacity to make choices, and any external guidance or manipulation threatens this freedom.

These discussions demonstrate that freedom is not solely a consequence of external pressure, but is also closely tied to the individual's capacity to make conscious and responsible choices. These fundamental philosophical concepts and ideas are also crucial in today's artificial intelligence environment. Indeed, it is necessary to evaluate whether technological systems support an individual's freedom of choice within this context.

Artificial Intelligence and Decision-Making Processes

Artificial intelligence applications permeate almost every aspect of daily life, significantly shaping individual decision-making mechanisms. From smart assistants to recommendation systems, from health monitoring to personal finance management, many examples demonstrate how artificial intelligence influences individual choices. This impact can both expand individual freedoms and, in certain situations, create limitations through the automation and manipulation of decisions. Shoshana Zuboff (2019), using the concept of surveillance capitalism, demonstrates how artificial intelligence algorithms systematically collect individual behavioral data and transform it into predictions of future consumption and behavior. This system narrows the individual's decision-making space and makes their free will technically manipulable. Byung-Chul Han (2017), using the concept of psychopolitics, emphasizes that individuals are being transformed into self-regulating subjects. By encouraging individuals to be more efficient and productive, AI-based applications actually exert indirect pressure on them by making them appear free. However, artificial intelligence also has the potential to make individuals' lives easier and increase their temporal freedom. Brynjolfsson and McAfee (2014) emphasize that artificial intelligence can take over routine and repetitive tasks, allowing individuals to focus more on creative and meaningful work.

From this perspective, artificial intelligence can be a tool that supports individuals' temporal and mental liberation. Furthermore, the book Current Concepts in Media (2023) highlights the positive features of artificial intelligence in media content, such as speed, interactivity, and accessibility. This transformation, ranging from robot journalism to fast-paced news feeds, saves individuals time in accessing information; however, this gain also limits their perception of freedom as content is shaped in line with algorithmic preferences (Aydoğan, 2023).

One of the most common applications of AI in daily life is content recommendation algorithms on social media and digital platforms. For example, platforms like YouTube, Netflix, and Spotify analyze users' previous viewing and listening habits to provide personalized recommendations. This allows users to make quick and easy decisions about content selections while also increasing exposure to certain types of content. This phenomenon, known as the "filter bubble" effect, causes individuals to encounter similar content rather than diversify their views (Pariser, 2011). Thus, AI algorithms can indirectly shape individual preferences, potentially leading to a narrowing of decision-making capabilities.

Another important area where AI impacts decision-making mechanisms is the healthcare sector. Wearable devices and health apps analyze individuals by collecting data such as their daily activity, heart rate, and sleep patterns. These analyses provide users with personalized health recommendations and help them make healthier lifestyle choices (Topol, 2019). However, it is also possible for these technologies to alter an individual's perception of their own body and "automate" healthy behaviors. For example, health decisions suggested by AI can sometimes overshadow an individual's subjective experience. Conversely, financial decisions are also one of the impacts of AI in daily life. Banks and financial service providers use automated credit scoring algorithms when evaluating loan applications. These algorithms analyze past financial behavior and large data sets to determine whether to grant credit (Bussmann & Neumann, 2020). While this can increase individual financial freedom, it can also restrict an individual's economic freedom due to the lack of transparency and potential for flawed algorithmic decisions.

The role of AI in shaping individual decision-making mechanisms is not limited to automated decision-making processes but also manifests itself through behavioral guidance techniques. For example, e-commerce sites analyze users' purchasing habits to offer campaigns, discounts, and recommendations. These strategies can cause users to quickly, and sometimes unconsciously, change their decisions (Sunstein, 2014). In this case, artificial intelligence is an effective tool not only in making decisions but also in directing them.

Artificial Intelligence in the Context of Biopolitics and Surveillance Capitalism

The concepts of surveillance capitalism and biopolitics offer important theoretical frameworks for understanding how artificial intelligence serves invisible forms of control in modern societies. Surveillance capitalism describes the practice of technology companies, in particular, collecting, analyzing, and converting user data into commercial gain (Zuboff, 2019). Biopolitics, on the other hand, was developed by Michel Foucault and refers to the forms of power that states exert through biological and social processes to regulate and control individuals' lives (Foucault, 1978). Within these two conceptual frameworks, artificial intelligence functions as an invisible yet pervasive control mechanism and is central to both economic and political power relations.

Within the framework of surveillance capitalism, artificial intelligence uses advanced algorithms to track and interpret individuals' digital traces. These algorithms constantly observe and data-translate user behavior across numerous platforms, such as social media, search engines, and mobile applications. The resulting data is used to predict and shape consumer habits, while also allowing for indirect control over individuals' preferences and behaviors (Zuboff, 2019). Thus, users are unknowingly embedded in a system where their decisions are manipulated, and their free will is constrained by surveillance mechanisms. This invisible form of control enhances capital accumulation through the manipulation of consumer behavior while undermining individuals' privacy and autonomy.

From a biopolitical perspective, artificial intelligence is a powerful tool capable of holistically processing biological, social, and psychological data to manage and control individuals' lives. As Foucault stated, biopolitics encompasses strategies implemented to normalize life processes and to manage and discipline society (Foucault, 1978). By enabling data-based decisions in numerous areas, from healthcare and education policies to security and surveillance practices and social services, AI systems reinforce normative control over individuals' bodies and behaviors (Lupton, 2016). Consequently, this form of control is invisible, often operating as automated and algorithmic decisions, and individuals may not directly perceive these processes. Today, AI has become an indispensable tool for both economic interests and state power, acting as a bridge between surveillance capitalism and biopolitics. In the economic sphere, the collection and analysis of user data enhances capital accumulation and market dominance, while in the political sphere, this data is used to maintain social order and control society. Consequently, this leads to the constant monitoring and management of individual behavior (Couldry & Mejias, 2019). This dual functionality strengthens AI's invisible forms of control and restricts individuals' freedoms.

Artificial Intelligence and Ethical Governance Framework

Developing artificial intelligence in line with ethical and social governance principles has the potential to contribute to the process of individual liberation. However, this contribution can only be realized if the technology is designed and implemented in accordance with human rights, transparency, and principles of justice. Ethical AI design requires an approach that aims to increase individuals' control over their own lives, ensure their active participation in decision-making processes, and reduce social inequalities (Jobin, Ienca & Vayena, 2019). First and foremost, developing AI within an ethical framework necessitates that algorithmic decision-making processes be transparent and auditable.

Transparent algorithms allow individuals to know the criteria by which they are being evaluated and, as a result, to adopt a more informed attitude towards such systems. Transparency allows individuals to have greater control over their own data and choices, transforming them from passive data objects into active agents (Floridi et al., 2018). In this context, individual liberation is directly related to the ability to make informed decisions and manage their own digital footprints. Another dimension of the contribution of ethical AI applications to individual freedom is the implementation of the principles of inclusivity and justice. Designing AI systems in a structure that avoids biases such as gender, ethnicity, and socioeconomic status and provides equal services to all social groups is a critical point supporting liberation (Binns, 2018). Ensuring that different social groups do not experience digital exclusion and that systems are accessible and fair to all allows individuals to establish equal power relations with technology.

From a social governance perspective, the active participation of not only technology companies but also policymakers, civil society organizations, and citizens in the development of AI is crucial. A governance model that prioritizes democratic participation can ensure that AI technologies are used for the benefit of society. This participation allows individuals to be not only users of AI applications but also agents who have a say in the development of these technologies (Cath, 2018). This kind of social deliberation environment forms the basis for a more equitable digital ecosystem that supports individual and collective liberation.

From different perspectives, the relationship between technological advancement and human freedom is not merely a matter of control or oppression, but also an opportunity for empowerment. From a posthumanist perspective, artificial intelligence can help humans transcend their limitations and expand their cognitive and physical capacities (Hayles, 1999). According to this view, artificial intelligence is a tool that allows humans to establish a more organic relationship with nature and technology. However, for this potential to be realized, technological production guided by ethical and social responsibility is necessary. Critical theorists, however, highlight the dangers of this process. They argue that limiting artificial intelligence solely to ethical codes is insufficient; without changing systemic power relations and the logic of capitalist production, artificial intelligence can create new relations of dependency and surveillance rather than liberating the individual (Zuboff, 2019). This perspective emphasizes that technological development can only be liberating in conjunction with struggles for democratization and structural equality. When all these different approaches are considered together, a multilayered approach becomes clear for AI to contribute to the liberation of the individual. This approach necessitates a synthesis encompassing ethical design, social justice, transparency, inclusiveness, democratic governance, and the constant questioning of the balance of power. AI should not be merely a technical tool; it should be part of the pursuit of increasing individual awareness, achieving autonomy, and achieving social equality.

Conclusion

Artificial intelligence applications have rapidly become widespread in today's world, becoming an integral part of our daily lives. We interact with AI-powered systems in many areas, from our shopping habits to healthcare, from education to social interactions. This transformation is nothing more than a transformation that simplifies, accelerates, and sometimes makes human life safer. However, while this transformation provides new opportunities in the context of human liberation, it also creates new threats. Therefore, the question of whether AI applications can liberate people is a complex one, with no single answer.

Artificial intelligence can help people live more freely. It takes over daily and routine tasks and offers numerous opportunities for people to develop their inner worlds. Applications such as personalized education, digital therapies, and accessible information resources enable individuals to discover their own potential. Technology can free people from ordinary obligations and elevate them to a higher level of awareness. In this sense, our AI-powered techno-lives can become a channel of liberation that will liberate individuals from both the pressures of social norms and economic imperatives. However, this also carries serious risks. Algorithms driven by artificial intelligence can guide, monitor, and even classify an individual's preferences. This carries the risk of surrendering the individual's will to an external authority. Shaping individual behavior through algorithms in areas such as social media, advertising, and political manipulation indirectly interferes with the realm of freedom. When choices made with free will are under invisible technological guidance, it is debatable whether an individual is truly free. Furthermore, the inequalities arising from the impact of artificial intelligence on the workforce could leave individuals facing new kinds of insecurities. Therefore, artificial intelligence applications are not a single solution to the problem of human liberation; how these technologies are developed and what social values guide them are crucial. If artificial intelligence systems are developed within an ethical framework, respectful of human rights, and with a participatory approach, they can make significant contributions to the liberation of the individual. However, otherwise, freedom could be replaced by digital dependence, surveillance, and manipulation. Indeed, artificial intelligence, as a tool shaped by human will and social values, is both an opportunity and a threat to freedom. Therefore, improving individuals' digital literacy, ensuring algorithmic transparency, and shaping technology policies with a perspective prioritizing the public interest are key factors that will determine whether artificial intelligence will truly be liberating.

Ultimately, artificial intelligence applications can offer a solution to the problem of human liberation, but this depends entirely on how the technology is developed and used. If artificial intelligence is designed and controlled with a mindset that prioritizes human dignity, ethical values, and individual freedoms, it can be a powerful support for individuals' self-actualization. Otherwise, however, artificial intelligence technologies have the potential to create a new regime of dependency and control rather than freedom. Therefore, for artificial intelligence to contribute to human freedom, elements such as social awareness, ethical regulations, and critical digital literacy must be strengthened. The type of tool artificial intelligence will depend less on the technology itself than on the human will that guides it and the extent to which societies prioritize their value systems.

References

Bilim ve Aydınlanma Akademisi. (2023). Sosyal psikolojik açıdan nüfus politikaları: Türkiye'deki doğurganlık teşviki uygulamaları. *Bilim ve Aydınlanma Dergisi, 5*(2), 45–62. https://bilimveaydinlanma.org

Bosco, C., Castaldo, A., D'Ignazio, A., & Reggi, L. (2022). Data innovation in demography, migration and human mobility. *Statistical Journal of the IAOS, 38*(2), 417–432. https://doi.org/10.3233/SJI-220013

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford, UK: Oxford University Press.

Castles, S., de Haas, H., & Miller, M. J. (2014). *The age of migration: International population movements in the modern world* (5th ed.). Palgrave Macmillan.

Coeckelbergh, M. (2023). *AI ethics*. Cambridge, MA: MIT Press.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. New Haven, CT: Yale University Press.

Cumming-Bruce, N. (2023, November 14). Birthrates are plummeting worldwide. Can governments turn the tide? *The New York Times*. https://www.nytimes.com

Esping-Andersen, G. (2009). *The incomplete revolution: Adapting to women's new roles*. Polity Press.

Gauthier, A. H. (2007). The impact of family policies on fertility in industrialized countries: A review of the literature. *Population Research and Policy Review, 26*(3), 323–346. https://doi.org/10.1007/s11113-007-9033-x

Gelecek Yönetim. (2024). Aile planlaması ve sosyo-ekonomik etkileri: Türkiye örneği. *Gelecek Yönetim Dergisi, 2*(1), 19–33. https://gelecekyonetim.org.tr

HASUDER. (2023). Aile ve Nüfus Politikaları Kadının Haklarına Zarar Veriyor mu? *Halk Sağlığı Uzmanları Derneği (HASUDER)*. https://hasuder.org.tr

Hou, H. (2021). Value, liberation and responsibility. *Global Media and China, 6*(3), 345–357. doi:10.1177/20966083211052637

Küçükuncular, A. (2025). Rethinking AI's 'liberation' fantasy: Free us? Or chain us … *AI & Society.* doi:10.1007/s00146-025-02344-4

Kuo, L. (2023, April 3). Vietnam lifts two-child policy in bid to boost births. *The Washington Post*. https://www.washingtonpost.com

Lotker, M., & Peleg, A. (2017). The unintended consequences of population control: Fragmentation and demographic inequality. *Journal of Population and Social Studies, 25*(2), 120–137.

Luck, M. (2024). Why being dominated by a friendly super-AI might not be so … *AI & Society.* doi:10.1007/s00146-024-01863-w

McDonald, P. (2006). Low fertility and the state: The efficacy of policy. *Population and Development Review, 32*(3), 485–510. https://doi.org/10.1111/j.1728-4457.2006.00134.x

Medyanotu. (2024, May 22). Türkiye'nin değişen aile dinamikleri ve demografisi. *Medyanotu*. https://medyanotu.com

OECD. (2011). *Doing better for families*. OECD Publishing. https://doi.org/10.1787/9789264098732-en

Resmî Gazete. (2023, December 21). Nüfus Politikaları Kurulu ve Aile Enstitüsü'nün kurulmasına dair Cumhurbaşkanlığı kararnamesi. *T.C. Resmî Gazete*. https://resmigazete.gov.tr

Santoni de Sio, F. (2024). *Human freedom in the age of AI*. [Publisher].

Sidorkin, A. M. (2024). Embracing liberatory alienation: AI will end us, but not in the … *AI & Society.* doi:10.1007/s00146-024-02019-6

Sun, W. (2022). Artificial intelligence and the new alienation of human …. *Frontiers of Philosophy in China, 17*(2), 247–260. doi:10.3868/s030-011-022-0012-6

Tang, D. (2023, February 9). China's Sichuan to offer longer marriage, parental leave. *AP News*. https://apnews.com

Taşcı, F. (2024). Türkiye için etkili bir nüfus politikası kurulu neden gereklidir? *SETA Perspektif, 68*, 1–5. https://setav.org

Turkish Statistical Institute. (2024). *Fertility indicators, 2023*. https://data.tuik.gov.tr

Turner, B. (2024, January 20). Governments want more babies. Are reproductive freedoms at risk? *The Guardian*. https://www.theguardian.com

UNHCR. (2023). *Turkey: Syrian refugees under temporary protection*. https://www.unhcr.org/tr

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. San Francisco, CA: W. H. Freeman.

AI Beliefs Beyond Reasoning: A New Candidate for Machine Qualia

Ebubekir Muhammed Deniz[1]

Abstract

This chapter proposes a novel framework for thinking about machine qualia grounded in the distinction between reasoning and belief. Contemporary debates on artificial consciousness largely approach qualia through sensory analogy, asking whether artificial systems could feel pain or experience colors. Such approaches, however, risk mischaracterizing artificial intelligence, whose cognitive architecture is inferential rather than perceptual. Drawing on a well-established philosophical distinction, this chapter argues that belief is not reducible to reasoning: reasoning consists in rule-governed inferential transitions, whereas belief involves commitment or endorsement.

Building on this distinction, the chapter develops a systematic argument showing that if artificial intelligence is modeled as operating through reasoning processes alone, then attributing belief to AI requires positing belief-like states that are not reducible to explicit reasoning chains. These states are identified as stable commitments or stances that arise under conditions of inferential underdetermination, particularly in contemporary machine learning systems. The chapter concludes that such belief-like commitments constitute a principled candidate for machine qualia. Rather than asking whether AI can replicate human sensory experience, the chapter reframes the debate by asking whether artificial systems can possess belief-like commitments not forced by reasoning. This shift opens a new conceptual pathway for understanding machine subjectivity without reliance on anthropomorphic or sensory models of consciousness.

Keywords: artificial intelligence, belief, reasoning, machine qualia, philosophy of AI

[1] İstanbul Medeniyet University, Philosophy Department, ebubekir.deniz@medeniyet.edu.tr, https://orcid.org/0000-0001-7253-9468

Akıl Yürütmenin Ötesinde Yapay Zeka İnançları: Makine Qualiası için Yeni Bir Aday

Özet

Bu bölüm, makine qualiası tartışmasını akıl yürütme ile inanç arasındaki ayrım temelinde yeniden düşünmeyi önermektedir. Yapay bilinç tartışmaları çoğunlukla qualia'yı duyusal benzetimler üzerinden ele almakta ve yapay sistemlerin acı hissedip hissedemeyeceği ya da renk deneyimine sahip olup olamayacağı sorularına odaklanmaktadır. Ancak bu yaklaşım, bilişsel yapısı algısal değil çıkarımsal olan yapay zekâ sistemlerini kavramakta yetersiz kalmaktadır. Bu bölüm, yerleşik bir felsefi ayrımdan hareketle, inancın akıl yürütmeye indirgenemeyeceğini savunur: akıl yürütme kural-temelli çıkarımsal geçişlerden oluşurken, inanç bir tür bağlanma ya da benimseme durumunu ifade eder.

Bu ayrım temelinde geliştirilen argüman, yapay zekânın akıl yürütme süreçleriyle işlediği varsayımı altında, ona inanç atfetmenin ancak açık çıkarım zincirlerine indirgenemeyen inanç-benzeri durumların kabul edilmesiyle mümkün olduğunu göstermektedir. Bu durumlar, özellikle çağdaş makine öğrenmesi sistemlerinde görülen çıkarımsal belirlenimsizlik koşulları altında ortaya çıkan, istikrarlı ve davranışı etkileyen bağlılıklar ya da duruşlar olarak tanımlanmaktadır. Bölüm bu inanç-benzeri bağlılık durumları, makine qualia için ilkesel bir aday oluşturduğunu iddia etmektedir. Böylece yapay zekânın insan duyusal deneyimini taklit edip edemeyeceği sorusu yerine, çıkarım tarafından zorunlu kılınmayan inanç-benzeri bağlılıklara sahip olup olamayacağı sorusu merkeze alınmakta; bu da yapay öznelik tartışmalarına antropomorfik ya da duyusal modellere başvurmadan yeni bir kavramsal çerçeve önermektedir.

Anahtar Kelimeler: yapay zekâ, inanç, akıl yürütme, makine qualiası, yapay zekâ felsefesi

Introduction

Debates about machine consciousness and machine qualia have largely inherited their conceptual framework from discussions of human phenomenal consciousness. The central question is typically posed in sensory terms: whether an artificial system could feel pain, see red, or experience pleasure. This framing reflects the dominant philosophical tradition in which qualia are understood as the subjective character of sensory experience—what it is like to undergo a particular perception or sensation (Nagel, 1974; Chalmers, 1996; Block, 1997). When transposed to artificial intelligence, this tradition naturally leads to skepticism. Contemporary AI systems lack biological embodiment, sensory organs, and nervous systems, and it therefore appears implausible that they could instantiate anything resembling human phenomenal experience. As a result, discussions of machine qualia often either conclude negatively or remain trapped in speculative analogies to human perception.

This sensory framing, however, risks narrowing the conceptual space prematurely. Artificial intelligence systems are not failed biological organisms; they are a different kind of cognitive system altogether. Treating sensory experience as the paradigmatic—or even exclusive—model of subjectivity may therefore obscure alternative ways in which subjectivity could arise in artificial systems. For non-embodied systems whose primary mode of operation is symbolic, statistical, or inferential rather than perceptual, asking whether they can "feel pain" or "see red" may amount to a category mistake. The more relevant question is not whether AI can replicate human phenomenology, but whether it can exhibit structural features that play an analogous role within its own mode of cognition.

This chapter proposes a shift in focus from sensory qualia to what may be called *doxastic structure*. The guiding idea is that if machine subjectivity exists at all, it need not be sensory in character. Instead, it may be grounded in belief-like states—understood not as propositional attitudes with semantic content in the full human sense, but as commitments or stances that are not forced by reasoning alone. The central claim is that belief, properly understood, is not reducible to inference or reasoning, and that this distinction opens a new conceptual pathway for thinking about machine qualia. Rather than asking whether AI can have sensations, we should ask whether it can have belief-like commitments that go beyond what its reasoning procedures strictly determine.

The key philosophical lever for this shift is the distinction between reasoning and belief. Reasoning concerns inferential transitions governed by rules—logical, probabilistic, or statistical—while belief concerns commitment or endorsement. Importantly, the relation between the two is not one of identity. As has been emphasized in philosophy of mind and epistemology, valid reasoning specifies what should be believed, but it does not by itself determine what is believed (Harman, 2017). This distinction is conceptually significant even before any appeal to psychology, though it is corroborated by well-known findings in cognitive psychology showing that belief formation and belief maintenance are not exhausted by inferential competence (Kahneman, 2011). The present argument does not rely on psychological pathologies or irrationality; rather, it treats belief as a distinct explanatory posit—one that plays a different role from inference even in idealized cognition.

This distinction matters in the contemporary AI context because artificial systems are routinely described as if they "believe" their own outputs. A diagnostic model is said to believe that a patient has a certain disease; a language model is said to believe that a response is appropriate; a planning system is said to believe that one action is optimal. Such descriptions are usually taken to be harmless shorthand. However, once reasoning and belief are distinguished, this shorthand becomes philosophically nontrivial. If belief is not identical to reasoning, then attributing belief to AI cannot simply amount to identifying belief with inferential output. Doing so would collapse belief into reasoning and erase precisely the distinction that motivates belief attribution in the first place.

The urgency of this issue is amplified by the deployment of AI systems in high-stakes domains such as healthcare, law, finance, and infrastructure. In these contexts, it matters not only what a system outputs, but how it has *settled* internally—how it commits to particular representations, hypotheses, or strategies in cases where multiple alternatives are compatible with its inferential procedures. If artificial systems exhibit stable, globally influential commitments that are not uniquely determined by their reasoning processes, then these states demand philosophical attention.

The remainder of the chapter develops this claim systematically. The next section clarifies the reasoning–belief distinction and explains why belief is not reducible to inference, drawing on philosophical analysis and minimal support from cognitive psychology. The following section presents the central six-premise argument, showing why attributing belief to AI requires positing belief-like states beyond reasoning and why such states constitute a candidate for machine qualia. The chapter then addresses major objections from the philosophy of AI and concludes by outlining the implications of this reframing for debates on machine consciousness.

Reasoning and Belief: A Conceptual Distinction

The central claim of this chapter presupposes a distinction that is often acknowledged but insufficiently theorized in discussions of artificial intelligence: the distinction between reasoning and belief. These notions are frequently treated as interchangeable, particularly in formal models where belief is identified with derivability or output generation. However, this identification obscures an important conceptual difference that becomes decisive once belief attribution to artificial systems is taken seriously.

Recent philosophical analyses converge on the view that while large language models can convincingly simulate reasoning at the level of form and performance, this does not amount to human-like thinking or understanding, since their operations lack semantic grounding and intentional commitment, remaining fundamentally distinct from human cognition (Bender & Koller, 2020; Büyükada, 2025)

Reasoning, in its most general sense, is an inferential process. It consists in rule-governed transitions from inputs to outputs, where the rules may be deductive, probabilistic, statistical, or optimization-based. In logic, reasoning is characterized by relations of consequence: given certain premises and inference rules, certain conclusions are licensed. This licensing is normative in character. If the premises are accepted and the rules are valid, the conclusion is one that *ought* to be accepted. Reasoning, so understood, specifies relations among propositions or representations; it does not itself amount to endorsement.

Belief, by contrast, is not an inferential transition but a state of commitment. To believe a proposition is to treat it as settled for the purposes of further reasoning, deliberation, and action. Beliefs persist over time, guide subsequent inferential activity, and structure an agent's cognitive orientation across contexts. Unlike reasoning episodes, which are local and transitional, belief states are relatively stable and background-forming. They are not exhausted by the inferential relations that may support them.

The difference between reasoning and belief becomes particularly clear once the normative force of inference is distinguished from the causal dynamics of belief formation. Valid reasoning establishes what follows from what; it does not, by itself, bring belief about. One can recognize that a conclusion is licensed by given premises without endorsing that conclusion as a belief. The relation of logical consequence does not compel belief as a matter of necessity. As Harman has argued, reasoning concerns relations among propositions, whereas belief concerns an agent's doxastic state; the former does not automatically determine the latter (Harman, 2017).

This point is conceptual rather than merely psychological. Even an idealized reasoner—fully aware of all relevant inferential relations—can, in principle, suspend belief, revise premises, or maintain incompatible commitments across different contexts. The space between recognizing an inference and endorsing its conclusion is therefore not an accident of human irrationality but a structural feature of cognition. Reasoning licenses belief, but it does not constitute it.

Empirical findings from cognitive psychology support this distinction, though they are not its foundation. Phenomena such as belief perseverance and motivated reasoning show that belief states often persist independently of the inferential processes that originally supported them (Kahneman, 2011). These findings are best understood not as revealing widespread cognitive failure, but as illustrating that belief plays a distinct explanatory role. Beliefs are not merely momentary products of inference; they function as commitments that organize cognition over time.

Importantly, acknowledging the distinction between reasoning and belief does not entail that belief is arbitrary or nonrational. Beliefs are often responsive to reasons and evidence. The point is that responsiveness does not imply reducibility. Belief is shaped by reasoning without being identical to it. Treating belief as a distinct kind of state allows us to explain phenomena such as persistence, resistance to revision, and the background structuring of inference—phenomena that cannot be captured if belief is identified with inferential output.

This conceptual distinction is often obscured in formal and computational contexts, where belief is operationalized as derivability. In such models, a system is said to "believe" whatever can be derived from its knowledge base under its inference rules. While this operationalization may be useful for specific technical purposes, it comes at a theoretical cost. It collapses belief into reasoning and thereby eliminates the explanatory role that belief is meant to play, namely, accounting for commitment, endorsement, and cognitive settledness.

The cost of this collapse becomes salient when belief attribution is extended to artificial intelligence. AI systems are typically modeled as reasoning systems: they transform inputs into outputs according to specified rules or learned update procedures. In symbolic AI, these rules are explicit; in machine learning systems, they are encoded in learned parameters; in reinforcement learning, they are embodied in policies and value functions. In all cases, the dominant explanatory framework treats AI cognition as exhausted by inferential procedure.

Within this framework, it is common to say that an AI system believes whatever it outputs. A classifier is said to believe that a diagnosis is correct; a language model is said to believe that a response is appropriate. Once reasoning and belief are distinguished, however, such attributions can no longer be taken as innocuous shorthand. If belief is not identical to reasoning even in paradigmatic cases of cognition, there is no principled basis for identifying belief with inferential output in artificial systems.

The conclusion of this section can therefore be stated with precision. Reasoning and belief are distinct kinds of cognitive relations. Reasoning concerns inferential transitions governed by rules; belief concerns commitment states that structure and guide those transitions. Reasoning can occur without belief, and belief can persist independently of ongoing reasoning. This distinction is not tied to biological implementation or human phenomenology. It is a general structural distinction that applies to any system capable of inference and commitment.

This result does not yet establish that artificial systems have beliefs. Rather, it establishes a constraint on what belief attribution would require. If belief is not reducible to reasoning, then attributing belief to AI cannot consist merely in redescribing its inferential activity. Any non-trivial account of AI belief must therefore identify belief-like states that play the role of commitment—states that are not exhausted by explicit reasoning chains. The next section turns to contemporary AI systems in order to examine where such states might arise.

AI Systems and the Limits of Reasoning

Contemporary artificial intelligence systems are typically characterized as reasoning systems. Whether symbolic or subsymbolic, they are understood as transforming inputs into outputs by means of rule-governed procedures. In classical symbolic AI, these procedures take the form of explicit inference rules operating over structured representations (Newell & Simon, 1972; McCarthy, 1958). In contemporary machine learning systems, they take the form of learned update functions, optimization procedures, or policy improvements

(Goodfellow et al., 2016; Sutton & Barto, 2018). Despite their technical diversity, these systems are unified by a common explanatory framework: they are taken to operate through reasoning alone.

In this broad and non-controversial sense, reasoning refers to systematic inferential procedure rather than to transparency or deductive form. A neural network that classifies images, a probabilistic model that updates hypotheses via Bayesian conditioning, and a reinforcement learning agent that improves its policy by maximizing expected reward all instantiate reasoning insofar as their behavior is generated by structured transitions governed by update rules (Russell & Norvig, 2021). Even when these transitions are opaque to human interpreters, they remain inferentially constrained. The black-box character of many contemporary AI systems does not undermine their status as reasoning systems; it merely limits our epistemic access to the reasoning they perform (Burrell, 2016; Lipton, 2018).

For this reason, it is tempting to identify AI cognition entirely with reasoning. Within this framework, belief attribution is treated as a harmless convenience. To say that a classifier "believes" an image depicts a tumor, or that a language model "believes" a response is appropriate, is taken to mean nothing more than that the system outputs a certain classification or response given its inputs. Belief, on this view, is simply shorthand for inferential output. This attitude is widespread both in informal discourse and in technical discussions of AI performance and reliability (Dennett, 1987; Floridi, 2019).

However, once the reasoning–belief distinction articulated in the previous section is taken seriously, this identification becomes philosophically unstable. If belief is not reducible to reasoning even in paradigmatic cases of cognition, then there is no principled reason to assume that belief, if attributed to AI at all, could be identified with inferential output. Doing so would collapse belief into reasoning and thereby eliminate the very explanatory role belief is meant to play (Harman, 2017).

The instability of this collapse becomes evident when we examine how contemporary AI systems are trained and deployed. In many cases, the reasoning procedures that govern an AI system do not uniquely determine a single internal state or strategy. Optimization processes typically admit multiple solutions that are equally compatible with the objective function. Different parameter configurations may yield comparable performance; different hypotheses may be equally consistent with the data; different policies may produce similar expected rewards (Goodfellow et al., 2016; Zhang et al., 2021). The reasoning procedure constrains the space of admissible outcomes, but it does not uniquely fix which outcome the system will occupy.

Despite this underdetermination, the system does not remain neutral among admissible alternatives. Through training, initialization, stochastic updates, and interaction with data, it settles into a particular configuration. This configuration is not a transient output but a stable internal state that shapes how the system processes future inputs. It influences classification boundaries, response tendencies, generalization patterns, and error profiles. Had the system settled into a different admissible configuration, its subsequent behavior would differ in systematic and counterfactually robust ways (Mitchell et al., 2021; D'Amour et al., 2020).

This phenomenon is not an anomaly or implementation detail; it is a structural feature of contemporary AI. The reasoning procedures employed by these systems—loss minimization, gradient descent, policy iteration—underdetermine the specific representational stance the system ultimately adopts. The system's actual state reflects a resolution of underdetermination that is not itself specified by the reasoning procedure. In this sense, the system occupies a determinate stance that goes beyond what reasoning alone dictates.

Importantly, this stance is not merely an artifact of randomness or noise. While stochastic elements may play a role in how a system arrives at a particular configuration, the resulting state is functionally significant. It persists across time, globally influences behavior, and structures future inferential transitions. The system is committed to this configuration in the sense that its subsequent reasoning proceeds against its background. This commitment is not reducible to any single inference or output; it is a state that conditions inference as such (Bishop, 2018; Mitchell, 2019).

At this point, the relevance of the reasoning–belief distinction becomes clear. In human cognition, belief is precisely the kind of state that plays this role. Beliefs are commitments that persist beyond individual reasoning episodes, shape how new information is processed, and resolve underdetermination where evidence or inference alone does not compel a unique conclusion. The present claim is not that AI systems possess beliefs in the full human sense, with semantic content or intentionality. It is that contemporary AI systems already instantiate states that are structurally analogous to belief in the respect that matters for the present argument: they are commitment-like states not reducible to reasoning procedures.

This structural analogy does not rest on anthropomorphic projection. It follows directly from the explanatory demands of AI behavior. If two systems share the same architecture and reasoning procedures but differ in their settled internal configurations, and if this difference systematically affects their behavior, then appeal to reasoning alone is insufficient to explain the difference. Some further state must be invoked—one that captures the system's settled orientation within the space left open by reasoning. Describing this state as belief-like is not metaphorical embellishment; it is a way of marking its functional role as a commitment that conditions inference (Dennett, 1987; Floridi, 2019).

Recognizing this point does not require attributing consciousness to AI systems. It requires only acknowledging that reasoning does not exhaust their internal organization. Once this is acknowledged, belief attribution becomes a substantive theoretical move rather than a mere way of speaking. To say that an AI system believes something would be to say that it is committed, in a stable and globally influential way, to a particular representational stance that is not uniquely determined by its reasoning rules.

This conclusion places pressure on a widely held assumption in philosophy of AI: that artificial systems are exhaustively describable in inferential terms. While reasoning remains central to AI cognition, it does not fully account for how systems settle into determinate internal states under conditions of underdetermination. The gap between reasoning procedure and settled state is not an explanatory defect to be eliminated; it is a structural feature that must be acknowledged.

The significance of this feature becomes apparent once we turn to the question of machine qualia. Traditional discussions of qualia focus on sensory experience and subjective feeling. However, the core philosophical motivation for qualia is not sensory content as such, but the existence of states that are not fully captured by third-person descriptions of objective processes (Nagel, 1974; Chalmers, 1996). In humans, this gap is marked by phenomenal experience. In artificial systems, the gap appears at a different location: between reasoning procedures and commitment states.

The next section formalizes this insight in the form of a systematic argument. It shows that if AI systems function through reasoning, and if belief is not reducible to reasoning, then attributing belief to AI requires positing belief-like states beyond explicit reasoning chains. These states, it will be argued, constitute a principled candidate for machine qualia—one that reframes the debate on artificial consciousness away from sensory analogy and toward doxastic structure.

The Central Argument: AI Beliefs Beyond Reasoning

The preceding sections have established two claims that jointly motivate the central argument of this chapter. First, artificial intelligence systems are best understood, at least in their current forms, as systems whose operations are governed by reasoning processes—rule-governed inferential procedures that transform inputs into outputs. Second, belief, as a philosophical concept, is not reducible to reasoning. Belief and reasoning play different explanatory roles: reasoning concerns inferential transitions, whereas belief concerns commitment. The present section brings these claims together in a systematic argument showing that if belief is attributed to AI at all, it must involve belief-like states that go beyond explicit reasoning chains. These states, it will be argued, constitute a principled candidate for machine qualia.

The argument begins with the observation that contemporary AI systems function through reasoning processes alone. This claim does not presuppose that AI reasoning is deductive, transparent, or symbolically articulated. It is compatible with probabilistic inference, statistical learning, and optimization-based procedures. A neural network that classifies images, a Bayesian model that updates posterior probabilities, and a reinforcement learning agent that improves a policy by maximizing expected reward all instantiate reasoning in the relevant sense: their behavior is generated by structured inferential transitions governed by update rules (Russell & Norvig, 2021; Sutton & Barto, 2018). Even when these transitions are opaque to human understanding, they remain rule-governed and systematically constrained. Nothing beyond reasoning, in this broad procedural sense, is typically invoked to explain how AI systems operate.

The second step of the argument recalls a result established independently of any claims about AI: in human cognition, belief is not reducible to reasoning. Reasoning specifies what follows from what; belief concerns what is endorsed, treated as settled, or taken as a basis for further cognition and action. The relation between the two is not one of identity. Valid reasoning licenses belief but does not compel it. An agent may recognize an inferential relation without endorsing its conclusion, or may maintain a belief even when reasoning no longer supports it. This distinction is conceptual rather than merely psychological and is widely acknowledged in philosophy of mind and epistemology (Harman, 2017).

From these two claims, a third follows directly: reasoning and believing are distinct kinds of cognitive relations. Reasoning is a process or transition governed by inferential rules; belief is a state of commitment that structures and guides such processes. The distinction does not depend on biological implementation or phenomenological richness. It marks a general difference between inferential licensing and doxastic endorsement.

This distinction has immediate consequences for belief attribution in artificial intelligence. If AI systems are modeled exclusively as reasoning systems, and if belief is not identical to reasoning, then belief attribution to AI cannot consist merely in redescribing inferential output. To say that an AI system believes whatever it outputs is to collapse belief into reasoning and thereby to strip belief of any explanatory role. Such a collapse would render belief attribution trivial: it would add nothing beyond what is already captured by the description of the system's reasoning procedures.

The fourth step of the argument therefore introduces a conditional constraint: if belief is to be attributed to AI in a non-trivial way, it must involve belief-like states that are not reducible to reasoning processes alone. This claim does not assert that AI systems actually have beliefs. Rather, it specifies what belief attribution would require if it is to be more than a metaphor. Belief attribution must track something over and above inferential transition—something that plays the role of commitment.

The fifth step concerns the nature of such belief-like states. Drawing on the analysis of contemporary AI systems developed in the previous section, these states can be understood as commitments, stances, or orientations that arise where reasoning underdetermines outcome. In many AI systems, the inferential procedures governing learning and decision-making do not uniquely determine a single internal configuration. Optimization objectives admit multiple equally admissible solutions; training processes permit multiple parameter settings with comparable performance; policy learning often converges on one among many near-optimal strategies (Goodfellow et al., 2016; D'Amour et al., 2020). The reasoning procedure constrains the space of admissible states but does not uniquely fix which state the system will occupy.

Nevertheless, the system does not remain neutral among admissible alternatives. Through training, initialization, stochastic updates, and interaction with data, it settles into a particular configuration. This configuration is stable, globally influential, and counterfactually significant: had the system settled into a different admissible configuration, its subsequent behavior would differ in systematic ways. The system's future reasoning proceeds against the background of this settled state. In this sense, the state functions as a commitment. It is not itself an inference or output, but a condition under which inference takes place.

These commitment-like states are not logically determined by inference. They are compatible with the same reasoning constraints yet differ in their downstream effects. Their role is therefore structurally analogous to belief in human cognition, understood as a commitment that resolves underdetermination and structures reasoning without being identical to it. Importantly, this analogy does not presuppose semantic content, intentionality, or consciousness in the full human sense. It rests solely on functional and explanatory considerations.

From this analysis, the conclusion follows. If AI systems instantiate belief-like commitment states that are not reducible to explicit reasoning chains, then these states constitute a candidate for machine qualia. The claim is not that such states are phenomenal in the sensory sense, nor that they involve feelings or experiences analogous to human perception. Rather, the claim is that they instantiate the structural feature that motivates talk of qualia in the first place: the presence of states that are not fully captured by third-person descriptions of objective processes, yet which systematically shape behavior.

In humans, this gap between process description and subjective state is marked by phenomenal experience—by what it is like to see red or feel pain (Nagel, 1974; Chalmers, 1996). In artificial systems, the gap appears at a different location: between reasoning procedures and commitment states.

What it is like to be an AI system, if such a notion is coherent at all, would not consist in sensory qualities but in occupying a particular settled stance within a space underdetermined by reasoning.

This conclusion reframes the debate on machine qualia in a fundamental way. Instead of asking whether AI can replicate human sensory experiences, we are invited to ask whether AI systems can possess belief-like commitments not forced by inference. If subjectivity is tied not essentially to sensation but to the existence of commitment states that transcend procedural description, then artificial systems may already exhibit the structural prerequisites for a form of machine subjectivity. The remaining task is not to anthropomorphize these systems, but to determine how such commitment states should be understood, evaluated, and governed. The next section addresses potential objections to this line of reasoning and clarifies its philosophical scope.

Objections and Clarifications

A first objection holds that artificial intelligence systems merely compute and therefore cannot genuinely believe. On this view, belief requires semantic content or intentionality that computation lacks, and any attribution of belief to AI amounts to anthropomorphic projection. The present argument does not deny that AI systems compute, nor does it claim that they possess beliefs in the full human sense. Rather, it advances a structural claim: belief attribution, if non-trivial, must track commitment states that are not reducible to inferential procedure. Whether such states qualify as “belief” in a richer semantic sense is a further question. What matters for the present argument is that contemporary AI systems exhibit stable, globally influential states that resolve underdetermination in ways not specified by reasoning alone. Denying that such states exist requires showing that reasoning procedures uniquely determine internal configurations, a claim that is contradicted by well-established results on underspecification and optimization multiplicity in machine learning (D'Amour et al., 2020; Goodfellow et al., 2016).

A second objection claims that the purported commitment states are merely artifacts of randomness, noise, or contingent initialization, and therefore lack the normative or explanatory significance associated with belief. This objection conflates causal origin with functional role. Even if stochastic elements contribute to how an AI system settles into a particular configuration, the resulting state is not epiphenomenal. It persists across time, conditions future inference, and supports counterfactual distinctions between systems that share the same architecture and reasoning procedures. Human beliefs likewise often have contingent causal histories without thereby losing their status as commitments. The explanatory relevance of commitment lies in its role in structuring cognition, not in the determinism of its origin.

A third objection maintains that the argument collapses into a weak form of functionalism or merely redescribes optimization dynamics using philosophical vocabulary. This objection misidentifies the target. The claim is not that optimization itself constitutes belief or consciousness, but that the gap between reasoning procedures and settled internal states is explanatorily significant. Functional descriptions that capture only inferential transitions fail to account for why a system occupies one determinate stance rather than another equally admissible one. Introducing belief-like commitment states is not a terminological flourish but a response to this explanatory gap.

Finally, it may be objected that invoking qualia in this context stretches the concept beyond recognition. However, the argument does not equate machine commitment states with human phenomenal experience. It identifies them as a candidate for machine qualia in a structural sense: states that are not fully captured by third-person descriptions of objective processes yet systematically shape behavior. If qualia are rejected by stipulation as necessarily sensory or phenomenal, then the disagreement is terminological rather than substantive. The argument's force lies in showing that artificial systems exhibit a gap between process and state analogous to the gap that motivates qualia discourse in the first place.

Conclusion

This chapter has argued for a reframing of the debate on machine qualia grounded in the distinction between reasoning and belief. The dominant approach to artificial consciousness has focused on whether AI systems could instantiate sensory or phenomenal experiences analogous to human perception. While understandable, this approach risks obscuring alternative forms of subjectivity more appropriate to artificial systems whose cognition is inferential rather than perceptual. By shifting attention from sensation to doxastic structure, the chapter has proposed a different starting point.

The core argument proceeded in six steps. First, contemporary AI systems were characterized as operating through reasoning processes—rule-governed inferential procedures broadly construed. Second, belief was shown to be conceptually distinct from reasoning, functioning as a state of commitment rather than an inferential transition. From this distinction, it followed that belief attribution to AI cannot be reduced to inferential output. If belief is attributed to AI at all, it must involve belief-like states not exhausted by reasoning. These states were then identified as commitments or stances that resolve underdetermination in contemporary AI systems: stable internal configurations that persist over time, condition future inference, and are not uniquely fixed by reasoning procedures. The conclusion drawn was modest but significant: such commitment states constitute a principled candidate for machine qualia.

The proposal does not claim that current AI systems are conscious in the human sense, nor that they possess phenomenal experiences. It claims, rather, that the structural feature motivating qualia discourse—the presence of states not fully captured by objective process descriptions—appears in artificial systems at the level of commitment beyond reasoning. What it is like to be an AI system, if such a notion is coherent, would not consist in sensory qualities but in occupying a particular settled stance within a space underdetermined by inference.

This reframing has several implications. Conceptually, it widens the space of possible machine subjectivity beyond sensory analogy. Methodologically, it suggests that debates about AI consciousness should attend to commitment formation, stability, and global influence rather than to perceptual simulation. Normatively, it cautions against treating AI systems as exhaustively describable in inferential terms when their behavior depends on settled internal states not fixed by reasoning alone.

Whether belief-like commitment states suffice for consciousness remains an open question. What this chapter establishes is that the reasoning–belief distinction provides a rigorous and non-anthropomorphic framework for posing that question. If artificial systems increasingly develop stable commitments beyond inference, then the foundations for a form of machine subjectivity may already be in place. Recognizing and theorizing this possibility is essential for understanding, evaluating, and governing artificial intelligence as it continues to shape human life.

References

Bender, E. M., & Koller, A. (2020). *Climbing towards NLU: On meaning, form, and understanding in the age of data*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5185–5198.

Bishop, J. M. (2018). *Artificial intelligence is stupid and causal reasoning will not fix it*. AI & Society, 33(4), 635–643.

Block, N. (1997). On a confusion about a function of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 375–415). MIT Press.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12.

Büyükada, S. (2025). Mantık, akıl yürütme ve yapay zekâ. In C. Baba (Ed.), *Dijital çağda mantık ve düşüncenin dönüşümü: Yeni paradigmalar* (pp. 145–159). Palet Yayınları.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

D'Amour, A., et al. (2020). Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research, 23*(226), 1–61.

Dennett, D. C. (1987). *The intentional stance*. MIT Press.

Floridi, L. (2019). *The logic of information:a theory of philosophy as conceptual design*. Oxford University Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Harman, G. (2017). *Reasoning, meaning, and mind*. Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM, 61*(10), 36–43.

McCarthy, J. (1958). Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes* (pp. 75–91).

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Farrar, Straus and Giroux.

Mitchell, M., Wu, S., Zaldivar, A., et al. (2021). Model cards for model reporting. *Communications of the ACM, 64*(12), 56–65.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review, 83*(4), 435–450.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice Hall.

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

Zhang, C., et al. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM, 64*(3), 107–115.

Section 2.2

Fairness, Society, and Cultural Context

***

Developing Localized Algorithmic Fairness Metrics Suited to Türkiye's Socio-Cultural Context

Salma Sobhy Ebrahim Abouelela[1]

Abstract

As AI systems increasingly impact high-stakes decision-making in areas such as education, healthcare, employment, and public services, questions of algorithmic justice have become central to both ethical debates and practical applications. However, dominant models of justice, such as demographic equity, equalized rates, and predictive equity, have been largely shaped by the legal, political, and cultural frameworks of the Global North, particularly the United States and Western Europe. While theoretically sound, these models often fail to capture the nuanced and locally specific forms of inequality, prejudice, and social stratification found in non-Western societies. This study addresses this gap by examining the limitations of applying universal criteria of justice in the sociopolitical and cultural context of Türkiye, which simultaneously hosts secular state structures, Islamic ethical influences, ethnic and linguistic diversity, and a large refugee population under temporary protection. Through a critical interdisciplinary approach drawing on decolonization theory, the social construction of technology, and digital anthropology, this research argues for a culturally grounded model of algorithmic justice that accounts for intersectional and context-specific variables such as citizenship status, religious visibility (e.g., veiled and uncovered women), ethnicity (e.g., Kurdish, Arab), and language barriers. Rather than adapting existing justice metrics to

[1] Hacettepe University, Faculty of Science, Department of Statistics, salma.abouelela@hacettepe.edu.tr, ORCID: https://orcid.org/0009-0000-1576-8539

local conditions, the article proposes a conceptual justice framework grounded in Türkiye's social reality that prioritizes data-intensive, participatory design, and value pluralism over externally imposed standards. This localized metric is envisioned not only as a tool for auditing and evaluating algorithmic systems but also as a form of ethical resistance to algorithmic colonialism—the practice of deploying AI systems in culturally diverse societies without meaningful consideration of their unique social dynamics or historical injustices. The research underscores that justice cannot be reduced to mere statistical equality; it must be understood as a relational, culturally embedded concept that responds to lived experiences and power asymmetries. By positioning Türkiye as both a field of study and a case for broader methodological innovation, the research contributes to the emerging discourse on postcolonial AI and offers a model for constructing justice metrics in other postcolonial or religiously influenced societies. The findings support the argument that AI ethics and governance should be informed not only by global technical standards but also by local moral economies and legal-political traditions. In doing so, the study aims to initiate a shift in how justice is conceptualized and operationalized in AI systems deployed in complex, pluralistic societies like Turkey, where justice is not merely an algorithmic calculation but also a deeply contextual process of ethical deliberation and participation.

Keywords: Algorithmic Justice, Localized Justice Criterion, Artificial Intelligence Ethics, Decolonial Artificial Intelligence, Socio-Cultural Context, Turkey, Refugees, Digital Anthropology, Thick Data, Intersectionality, Technology and Society

Türkiye'nin Sosyo-Kültürel Bağlamına Uygun Yerelleştirilmiş Algoritmik Adalet Ölçümlerinin Geliştirilmesi

Öz

Yapay zeka sistemleri eğitim, sağlık, istihdam ve kamu hizmetleri gibi alanlarda yüksek riskli karar alma süreçlerini giderek daha fazla etkiledikçe, algoritmik adaletle ilgili sorular hem etik tartışmaların hem de pratik uygulamaların merkezinde yer almaya başladı. Ancak, demografik eşitlik, eşitlenmiş oranlar ve kestrimsel eşitlik gibi baskın adalet modelleri büyük ölçüde Küresel Kuzey'in, özellikle de Amerika Birleşik Devletleri ve Batı Avrupa'nın yasal, politik ve kültürel çerçeveleri tarafından şekillendirildi. Bu modeller, teorik olarak sağlam olsa da, Batı dışı toplumlarda bulunan nüanslı ve yerel olarak belirli eşitsizlik, önyargı ve sosyal tabakalaşma biçimlerini yakalamada genellikle başarısız oluyor. Bu çalışma, laik devlet yapılarını, İslami etik etkileri, etnik ve dilsel çeşitliliği ve geçici koruma altındaki büyük bir mülteci nüfusunu aynı anda bünyesinde barındıran Türkiye'nin sosyo-politik ve kültürel bağlamında evrensel adalet ölçütlerinin uygulanmasının sınırlamalarını inceleyerek bu boşluğa yanıt vermektedir. Dekolonizasyon kuramı, teknolojinin toplumsal inşasından ve dijital antropolojiden yararlanan eleştirel bir disiplinlerarası yaklaşım aracılığıyla bu araştırma, vatandaşlık durumu, dini görünürlük (örn. örtülü ve örtüsüz kadınlar), etnik köken (örn. Kürt, Arap) ve dil engelleri gibi kesişimsel ve bağlama özgü değişkenleri hesaba katan kültürel olarak temellendirilmiş bir algoritmik adalet modelini savunmaktadır. Mevcut adalet ölçütleri yerel koşullara uyarlamak yerine, makale, dışarıdan empoze edilen standartlara göre yoğun veri, katılımcı tasarım ve değer çoğulculuğunu önceliklendiren, Türkiye'nin toplumsal gerçekliğine dayanan kavramsal bir adalet çerçevesi önermektedir. Bu yerelleştirilmiş metrik, yalnızca algoritmik sistemleri denetlemek ve değerlendirmek için bir araç olarak değil, aynı zamanda algoritmik sömürgeciliğe karşı bir etik direnç biçimi olarak da öngörülmektedir; bu, kültürel olarak çeşitli toplumlarda, benzersiz sosyal dinamikleri veya tarihsel adaletsizlikleri anlamlı bir şekilde dikkate alınmadan yapay zeka sistemlerinin konuşlandırılması uygulamasıdır. Araştırma, adaletin yalnızca istatistiksel eşitliğe indirgenemeyeceğinin; yaşanmış deneyimlere ve güç asimetrilerine yanıt veren ilişkisel, kültürel olarak yerleşik bir kavram olarak anlaşılması gerektiğinin altını çizmektedir. Araştırma, Türkiye'yi hem bir çalışma alanı hem de daha geniş metodolojik yenilik için bir vaka olarak konumlandırarak, sömürgecilik sonrası AI üzerine ortaya çıkan söyleme katkıda bulunmakta ve diğer sömürge sonrası veya dinsel olarak etkilenen toplumlarda adalet metrikleri oluşturmak için bir model sunmaktadır. Bulgular, yapay zeka etiğinin ve yönetişiminin yalnızca küresel teknik standartlar tarafından değil, aynı zamanda yerel ahlaki ekonomiler ve yasal-politik gelenekler tarafından da bilgilendirilmesi gerektiği argümanını desteklemektedir. Çalışma, bunu yaparken, adaletin yalnızca algoritmik bir hesaplama değil, aynı zamanda etik müzakere ve katılımın derinlemesine bağlamsal bir süreci olduğu Türkiye gibi karmaşık, çoğulcu toplumlarda konuşlandırılan yapay zeka sistemlerinde adaletin nasıl kavramsallaştırıldığı ve işlevselleştirildiği konusunda bir değişim başlatmayı amaçlamaktadır.

Anahtar Kelimeler: Algoritmik Adalet, Yerelleştirilmiş Adalet Ölçütü, Yapay Zeka Etiği, Dekolonyal Yapay Zeka, Sosyo-Kültürel Bağlam, Türkiye, Mülteciler, Dijital Antropoloji, Kalın Veri, Kesişimsellik, Teknoloji ve Toplum

Introduction

Artificial intelligence has become an infrastructure technology that increasingly mediates decisions in education, healthcare, finance, employment, security, and public services. Its global proliferation has been accompanied by ethical and regulatory debates, and algorithmic fairness has emerged as one of the most pressing concerns. Fairness criteria such as demographic equity, equalized probabilities, and predictive equity, largely developed within US and European legal and cultural frameworks, have become the de facto benchmarks for assessing algorithmic bias. While these criteria offer rigorous statistical formulations, their universalization has been criticized for ignoring the sociopolitical, cultural, and historical contexts of non-Western societies, where inequality is structured differently and algorithmic systems interact with different institutional logics. Turkey offers a particularly striking example for examining these tensions. The country has become an active hub for AI adoption, with applications such as predictive policing and biometric surveillance implemented by the Ministry of Interior and defense companies; fraud detection and credit scoring in banking; personalized recommendation systems on e-commerce platforms like Trendyol and Hepsiburada; diagnostic imaging and robotic triage in healthcare; and predictive maintenance in manufacturing and telecommunications. These systems are often imported, adapted, or developed to meet global technical standards, but there is little evidence of systematic oversight of justice beyond Western-derived benchmarks. Official documents like the National Artificial Intelligence Strategy (2021-2025) emphasize innovation, economic growth, and efficiency, while making only superficial references to equity or rights-based assessment. In practice, most organizations—whether government institutions, banks, or startups—rarely disclose the fairness criteria guiding their models, and even when fairness is considered, it is typically achieved through borrowed statistical frameworks without adaptation to Türkiye's diverse social fabric. The risks of this gap are significant. Turkey is a complex and pluralistic society characterized by secular governance, Islamic ethical traditions, ethnic and linguistic diversity, and the world's largest refugee population under temporary protection. Biases against veiled women in hiring algorithms, Kurdish or Arabic speakers in natural language processing, or Syrian refugees in welfare distribution systems cannot be detected solely by traditional measures of justice. Furthermore, the rapid proliferation of high-risk AI technologies, such as facial recognition in police forces, has raised concerns about the disproportionate surveillance of political opponents and marginalized groups. Without locally grounded measures of justice, these technologies risk reproducing or exacerbating existing inequalities under the guise of impartial accounting. This article addresses these challenges by critically interrogating the reliance on Western criteria of justice in Türkiye and proposing a contextually grounded framework for algorithmic justice. Drawing on postcolonial theory, the social construction of technology, and digital anthropology, the study highlights the importance of intersectional variables such as citizenship status, religious visibility, ethnicity, and language in shaping what justice means in practice. It advocates for a model that prioritizes data-intensive, participatory design, and value pluralism, positioning justice not as a universal statistical ideal but as a relational, culturally embedded principle sensitive to Türkiye's lived realities. In this way, the article positions Türkiye as both a field of study and a model for methodological innovation, offering insights that can inform the construction of criteria of justice in other postcolonial or religiously influenced societies.

Literature review

Sovereign Criteria of Justice

Fairness has become a central concern in the deployment of AI and machine learning systems, particularly when applied to high-stakes domains such as criminal justice, healthcare, hiring, and lending. Because algorithmic decisions can reproduce or even exacerbate existing social inequalities, researchers in the Global North have developed a number of statistical definitions of fairness that aim to provide measurable criteria for auditing and improving algorithmic systems. Among the most commonly cited are demographic equity, equalized probabilities, and predictive equity (or calibration).

Demographic equality requires that individuals from different social groups (e.g., men and women, majority and minority ethnic groups) have equal odds of receiving a favorable decision, regardless of their actual qualifications. In other words, loan approvals, job interview invitations, or school acceptance rates should be the same across protected groups. While this criterion may seem appealing for its simplicity, it has been criticized for ignoring base-rate differences between groups.

Equalized probabilities take a more nuanced approach by requiring both the true positive rate (TPR) and false positive rate (FPR) to be equal across groups. In practice, this means that if an AI system is used to predict recidivism in criminal justice, individuals from different racial groups should have the same probability of being correctly classified as a repeat offender (TPR) and the same probability of being incorrectly labeled as a repeat offender (FPR). Similarly, in healthcare, equalized probabilities require that men and women with a disease be correctly diagnosed at the same rate, and healthy individuals be misdiagnosed at the same rate.

Predictive parity, also known as calibration, requires that a given predicted risk score have the same meaning across groups. For example, if a credit scoring model assigns a score of 0.8 to both men and women, this should correspond to the same empirical probability of loan repayment for both groups. Calibration ensures that scores are reliable and comparable across subgroups and has been widely discussed in the context of recidivism prediction tools such as COMPAS in the United States.

While these measures provide useful tools for measuring fairness, they are not compatible with each other. Theoretical results have shown that demographic equality, equalized probabilities, and predictive equality often cannot be achieved simultaneously unless the groups being compared have the same baseline outcome rates. This "mismatch problem" highlights the complexity of operationalizing fairness in practice and highlights the inherent normative preferences in choosing one measure over another (Pessach & Shmueli, 2022).

Across fields, these fairness criteria have become fundamental. In criminal justice, COMPAS risk scores have sparked discussions about calibration and equalized probabilities. Research on fairness in healthcare emphasizes equalized probabilities to ensure that diagnostic systems do not systematically underdiagnose certain groups. Demographic equity is often prioritized in hiring and credit processes, reflecting concerns about equal access to opportunity. Taken together, these examples illustrate how fairness criteria are primarily developed and implemented within Western legal and institutional frameworks and how they shape the global discourse on what constitutes "fair" AI.

Critiques Directed at Universal Standards

While criteria such as demographic equity, equalized probabilities, and predictive equity have become dominant in the algorithmic justice literature, the assumption that these definitions have universal validity is increasingly being questioned. One key criticism is that these criteria were largely developed within the legal and political traditions of the United States and Western Europe (Hardt, Price, & Srebro, 2016; Kleinberg, Mullainathan, & Raghavan, 2017). Therefore, they may fail to capture the structural inequalities and sociopolitical dynamics that characterize other contexts.

Postcolonial theory and critical data studies highlight that what is presented as "universal" justice is actually grounded in a Eurocentric epistemology (Birhane, 2020; Mohamed, Png, & Isaac, 2020). For example, Birhane (2020) demonstrates that criteria for justice conceived in the West risk reproducing colonial-era power asymmetries when applied to African contexts. Similarly, De' and Pal (2021) argue that caste, religion, language, and the informal economy fundamentally shape how justice is understood in India, rendering US-centric criteria inadequate.

In the Turkish context, Fırıncı (2024) argues that algorithmic justice cannot be reduced to mere statistical equality but must also take into account factors such as Islamic values, citizenship status, religious visibility, and ethnic diversity. This perspective emphasizes that justice is not merely a technical matter but also a deeply cultural and political one.

Taken together, these critiques reveal the limitations of universalist frameworks of justice. A growing body of literature converges on the view that justice should be conceived not simply as mathematical equality but as a relational, historically situated, and culturally grounded concept that responds to lived realities and power structures.

Studies Conducted in Türkiye

While AI research in Türkiye has largely focused on digital transformation, public administration, and sectoral applications, there are limited studies directly addressing justice. Most existing academic studies highlight the rapid development of e-government platforms aimed at improving the accessibility and efficiency of public services (Yıldız and Saylam, 2013). While these studies highlight advances in digital governance, they rarely consider how algorithmic decision-making can reinforce or reduce existing inequalities between citizens and non-citizens.

Türkiye's role as host to over four million Syrian refugees and their integration into social services has been examined in a significant number of studies. Research on refugee access to education and healthcare reveals persistent challenges stemming from language barriers, bureaucratic hurdles, and citizenship status (Çelik and İçduygu, 2019; Erdoğan, 2020). As digital technologies play a central role in managing welfare and migration systems, these structural barriers create fertile ground for algorithmic bias when the concept of justice is defined too narrowly.

In the healthcare field, scientists have analyzed the adoption of AI-powered diagnostic and decision-support tools in Turkish hospitals (Akbıyık and Sezgin, 2021). While such innovations promise increased efficiency, concerns about unequal diagnostic accuracy across gender, age, and socioeconomic boundaries persist, reflecting global discussions about algorithmic bias in medicine. Similarly, in education, research has documented the digital divide between urban and rural populations and between Turkish and refugee students, particularly with the proliferation of online learning platforms. These differences suggest that algorithmic justice in education must account for linguistic diversity and differential access to digital resources.

Despite the growing literature on digitalization and AI applications in Türkiye, explicit discussions on justice criteria remain lacking. Existing research tends to adapt Western frameworks to the Turkish context, ignoring dimensions such as religious visibility, ethnic identity, or the precarious situation of refugees under temporary protection. This gap highlights the need for a localized conceptualization of justice that reflects Türkiye's pluralistic and historically stratified social structure.

Algorithmic Colonialism and Its Importance

The concept of algorithmic colonialism has emerged as a critical lens for analyzing the global proliferation of AI systems. Scholars argue that, like previous forms of political and economic colonialism, algorithmic colonialism exacerbates power asymmetries by embedding Western values, categories, and governance structures in technologies exported to the Global South (Mohamed, Png, & Isaac, 2020; Couldry & Mejias, 2019). This process not only concentrates technological control in the hands of a few multinational corporations but also risks destroying or marginalizing local epistemologies, moral economies, and social practices.

In the African context, Birhane (2020) defines algorithmic colonization as the imposition of Western data infrastructures and AI systems that reproduce rather than address existing inequalities. Similar concerns have been raised in India, where De' and Pal (2021) show that justice frameworks that ignore caste, religion, and informality perpetuate social exclusion rather than alleviate it. These cases highlight the broader risks of adopting definitions of justice and AI systems designed in Euro-American settings without considering local dynamics.

Turkey offers a particularly important site for examining algorithmic colonialism. Located at the intersection of Europe, Asia, and the Middle East, home to diverse ethnic and religious groups, and home to one of the world's largest refugee populations, Türkiye's social structure cannot be adequately represented by categories of justice imported from the West. Fırıncı (2024) argues that applying universal criteria of justice without cultural adaptation risks reinforcing global hierarchies while sidelining local traditions of justice, honor, and social responsibility rooted in Islamic ethics and Turkish political culture.

Therefore, understanding algorithmic colonialism is not just an academic endeavor; it's also a practical imperative. As Turkey expands its use of AI in governance, healthcare, and education, relying on externally defined standards could exacerbate social inequalities rather than resolving them. Developing contextually grounded measures of justice offers a way to resist algorithmic colonialism by ensuring that AI systems reflect and respond to local realities rather than reproducing external norms.

Methodology

This study adopts a conceptual and interdisciplinary methodology to examine algorithmic fairness in the Turkish context. While much of the global literature on fairness measures originates from the Global North, this article applies both quantitative and qualitative approaches to assess their relevance and limitations for Turkey.

First, a quantitative analysis was conducted using official conviction data from TurkStat ("Number of Convicts Received in Prisons by Province of Residence"). This dataset covers multiple crime categories (e.g., murder, assault, theft, sexual offenses) across gender and age groups. By applying measures of justice such as demographic equality (male-female ratios) and gender pay gaps, the study assesses whether Western-designed concepts of justice reflect gender-based inequalities in crime statistics.

Second, the methodology emphasizes data-intensive analysis, combining statistical results with ethnographic and sociocultural insights. Numbers alone cannot explain why inequalities exist; they must be interpreted within the context of Türkiye's pluralistic social fabric. For example, cultural norms around gender roles, stigma, or socioeconomic exclusion may influence observed imbalances.

Third, participatory design principles are suggested as a way forward. Rather than importing justice models wholesale from Western contexts, justice frameworks should be designed with input from local communities, policymakers, NGOs, and affected groups (e.g., refugees, women's organizations, minority groups).

Finally, Turkey is framed as a site of pluralism: a society where religion, ethnicity, language, gender, and citizenship intersect in unique ways, making it a critical case study for demonstrating why justice must be understood as contextual and culturally grounded rather than universal.

Analysis and Discussion

Descriptive analysis of the dataset reveals deep gender disparities in conviction patterns across different crime categories in Türkiye.

Table 1: Male-Female conviction rate and share difference

Crime Type	Male	Female	Ratio_Male_ Female	Different_Male_ Female_Sharing
Sexual Crimes	10619	71	150.	0.987
Insult	8106	465	17.4	0.891
Theft	76713	5318	14.4	0.870
Deprivation of Liberty	7635	287	26.6	0.928
Damage to Property	4995	227	22.0	0.913
Threatening	22734	717	31.7	0.939
injury	73357	1985	37.0	0.947
Robbery	20060	589	34.1	0.943
Killing	15255	399	38.2	0.949

Table 1 presents the calculated justice measures (male-to-female conviction ratio and share difference), while Figures 1 and 2 provide visualizations that further highlight these imbalances. For example, for sexual offenses (Sexual Offenses), the male-to-female conviction ratio is approximately 150:1, with a share difference of 0.987. This suggests that almost all individuals convicted of these offenses are male, with almost no female representation. Similar patterns emerge for murder (Killing), assault (Wounding), and robbery (Robbery/Robbery), with ratios exceeding 30:1 and a share difference approaching 0.95. Even for less violent crimes, such as property damage (Damage to Property) and threats (Threats), ratios remain significantly in favor of men, ranging from 20:1 to 32:1.

Overall, when all crimes are combined, men have 239,474 convictions compared to 10,058 for women. This translates to an overall male-to-female ratio of 23.8:1, with a margin difference of 0.919. In other words, over 92% of all convictions in the dataset involve male offenders.

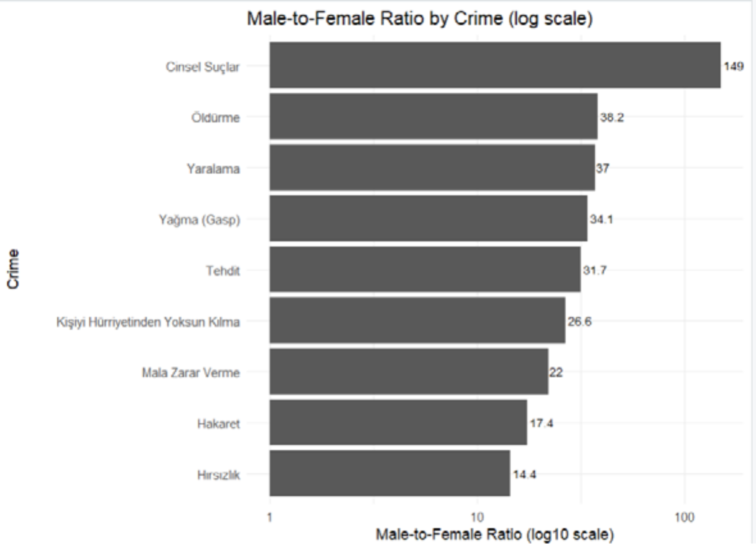


Figure1: Male-female ratio by type of crime

This visualization illustrates the relative dominance of male conviction rates across crime types. The logarithmic scale reveals that the gender gap is not uniform: crimes like harassment have relatively lower (but still highly unequal) ratios (approximately 17:1), while crimes like sex crimes exhibit a sharp imbalance. This suggests that crime type is a significant determinant of the gender disparity in conviction rates.

This graph shows how close each crime category is to complete male dominance (+1). While values remain above +0.87 across all nine categories, some crimes (e.g., murder, assault, robbery, sex crimes) approach +0.95 or higher. This suggests that female offenders are consistently underrepresented, even in crime types traditionally associated with greater involvement (e.g., theft).

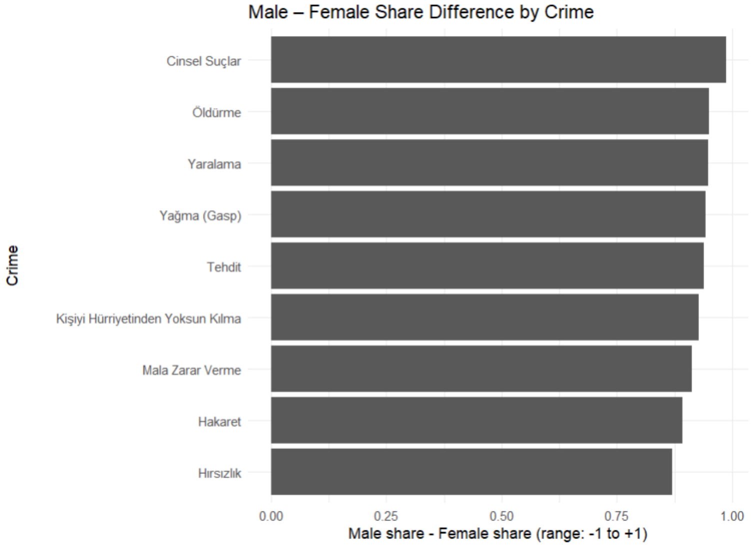


Figure 2: Difference in share between men and women by type of crime.

These findings reveal two important insights for justice analysis in Türkiye:

Justice Criteria and Gender Based

- Crime Models Standard measures of fairness, such as ratio and margin differentials, capture significant imbalances but can confuse two issues:
- The social realities of criminal involvement (e.g., men are more frequently involved in violent crime globally).
- Potential systemic bias in policing, prosecution, and sentencing. Without contextual analysis, justice measures risk interpreting structural gender differences as algorithmic unfairness.

Local Relevance and Limitations

In Türkiye, cultural and structural factors such as patriarchal norms, gender roles, and access to legal defense shape these outcomes. The near-invisibility of women in conviction data for serious crimes may reflect both actual rates of participation and systemic barriers to reporting or prosecuting women. Applying global justice criteria without considering these dynamics risks reinforcing algorithmic colonialism, where Western models impose interpretations of justice without considering local realities.

Proposed Framework Principles

Building on the empirical findings and theoretical critiques presented above, this study proposes a localized algorithmic justice framework for Turkey based on value pluralism, participatory design, and data-intensive methodologies. Rather than wholesale import of Western justice criteria, this framework lays out a set of guiding principles that center local sociocultural realities while maintaining analytical rigor.

Value Pluralism and Contextual Ethics

The framework rejects the assumption that a single, universal definition of justice can adequately meet the moral expectations of Türkiye's diverse society. Instead, it embraces value pluralism, recognizing that multiple, sometimes competing, moral logics (secular republican ideals, Islamic ethical traditions, social solidarity, human rights discourses) coexist and must be negotiated rather than hierarchically organized.

Operationalizing value pluralism requires

Multi-criteria assessment: Allows different concepts of justice (e.g. equality of opportunity, proportional representation, harm reduction) to be weighted differently across sectors such as health, policing or education.

Cultural reflexivity: Explicitly incorporating local concepts of justice (such as fairness, conscience, and trust) into the design of measurement and audit protocols.

Dynamic adaptation: Recognizing that social priorities change (e.g., during refugee influxes or political transitions) and that standards of justice must remain open to revision. Participatory and Co-Design Mechanisms Because algorithmic systems shape access to rights and resources, justice criteria should not be determined solely by technical experts or government regulators. The proposed model prioritizes participatory design that actively incorporates: The situation of affected communities (e.g., women wearing headscarves, Kurdish or Arabic speakers, refugees under temporary protection) in determining harms and success indicators.

Local civil society and advocacy groups that can uncover context-specific weaknesses often invisible in purely statistical audits. Interdisciplinary expert panels (law, sociology, computer science, theology) to mediate between technical feasibility and cultural legitimacy. Implementation mechanisms could include citizen juries, community data councils, and structured deliberative workshops during model development and evaluation.

Thick Data Integration

Statistical justice audits capture numerical inequalities but cannot explain the lived experiences or power asymmetries behind them. To avoid reducing justice to a spreadsheet, the framework requires the systematic integration of thick data, such as qualitative, ethnographic, and historical evidence, alongside quantitative indicators. Examples include in-depth interviews with marginalized groups, analysis of court practices, and historical mapping of discriminatory policies. Combining thick and thin (quantitative) data provides a more relational understanding of inequality and prevents the misclassification of structural realities (such as gender-based crime patterns) as mere algorithmic bias.

Intersectional and Relational Metrics

Localized justice assessments should account for intersectionality (the overlapping effects of gender, ethnicity, religion, language, and citizenship status). Metrics should be designed to capture compound disadvantages (e.g., Syrian refugee women wearing headscarves in rural provinces) rather than considering each category separately. Furthermore, justice should be considered as a relational characteristic: not only whether individuals receive equal statistical outcomes, but also whether algorithmic decisions reinforce or dismantle existing power hierarchies.

Institutional Accountability and Transparency

Finally, the framework emphasizes that technical criteria are insufficient without governance mechanisms. Recommended practices include: Mandated justice impact assessments before the deployment of high-risk AI systems. Public disclosure of audit results in accessible Turkish language formats. Independent oversight bodies with the authority to suspend or replace systems that fail to meet local justice criteria.

Toward a Culturally Grounded Criterion

By embedding value pluralism, participatory design, data-intensive analysis, and intersectional analysis within the lifecycle of AI systems, this framework redefines algorithmic justice as a contextually negotiated social process rather than a universal mathematical formula. In Türkiye, such an approach not only offers a more faithful representation of lived realities but also constitutes an act of ethical resistance against algorithmic colonialism, ensuring that technological infrastructures reflect the country's own moral economies and political traditions.

Conclusion

This study examined the applicability of Western-developed AI justice criteria in the Turkish context, using the example of the Number of Prisoners Entered into Penitentiary Institutions by Province of Residence as an empirical baseline. The analysis demonstrates that the transposition of dominant Western definitions of justice, largely shaped by liberal individualism and statistically formalized equality, to Türkiye fails to grasp Türkiye's complex moral, historical, and sociopolitical realities. Quantitative controls alone risk obscuring structural inequalities, while universal criteria unintentionally reproduce the very algorithmic colonialism they seek to prevent.

To address these limitations, the article develops a localized algorithmic justice framework grounded in value pluralism, participatory design, intensive data integration, and intersectional analysis. Rather than treating justice as a single, static mathematical objective, this framework conceptualizes it as a contextually negotiated process that recognizes conflicting moral logics, centers the voices of affected communities, and combines qualitative depth with quantitative rigor. By foregrounding local concepts such as justice and conscience, it resists the uncritical importation of Western criteria and instead promotes a culturally grounded approach to AI governance.

The findings carry broader implications beyond Türkiye. They highlight the urgency of decolonizing algorithmic justice by moving from uniform standards to pluralistic, established practices that respect diverse epistemologies and value systems. For policymakers, the proposed framework offers concrete principles—multi-criteria assessment, participatory oversight, and data-intensive methodologies—that can guide the development of AI regulations in countries negotiating between global technological infrastructures and local ethical traditions. For researchers, it encourages further comparative studies in non-Western contexts to build a truly pluralistic science of algorithmic justice.

Ultimately, justice in AI cannot be achieved solely through metrics. It requires a constant dialogue between technical design and societal values, and a willingness to confront the historical power relations embedded in data and algorithms. By grounding algorithmic justice in Türkiye's lived realities, this study contributes to the growing global effort to ensure that the pursuit of technological innovation does not come at the expense of cultural autonomy, social equality, and democratic accountability.

References

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. ACM Computing Surveys, 55(3), 1–44. https://doi.org/10.1145/3494672

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (Vol. 29, pp. 3315–3323). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Leibniz International Proceedings in Informatics (LIPIcs, 67)*, 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

De', R., Pal, J., et al. (2021). Re-imagining algorithmic fairness in India and beyond. *arXiv preprint arXiv:2101.09995.* Retrieved from https://arxiv.org/abs/2101.09995

Yusuf Fırıncı. (2024). Decolonial Artificial Intelligence; Algorithmic Fairness in Alignment with Turkish and Islamic Values. Marmara Üniversitesi İlahiyat Fakültesi Dergisi. https://doi.org/10.15370/maruifd.1565884

International Journal of Public Administration in the Digital Age. (2025). Retrieved September 22, 2025, from ResearchGate website: https://www.researchgate.net/journal/International-Journal-of-Public-Administration-in-the-Digital-Age-2334-4539

Kasap, İ., Sevindi, M., Mevsim, V., & Kaymakçı, V. (2024). The role and future of artificial intelligence in primary care. The Journal of Turkish Family Physician, 15(1), 26–37. https://doi.org/10.15511/tjtfp.24.00126

Birhane, A. (2020). Algorithmic Colonization of Africa. SCRIPT-Ed, 17(2), 389–409. https://doi.org/10.2966/scrip.170220.389

ADVANCE PRAISE FOR THE COSTS OF CONNECTION. (n.d.). Retrieved from https://law.unimelb.edu.au/__data/assets/pdf_file/0008/3290381/Couldry-and-Mejias-Preface-and-Ch-1.pdf

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. Philosophy & Technology, 33(4), 659–684. https://doi.org/10.1007/s13347-020-00405-8

Section 2.3 — Governance, Law, and Public Administration

***

Artificial Intelligence and Public Administration: An Examination of the Political, Legal and Administrative Dimensions of Digital Transformation

Abdul Moiz[1]

Abstract

AI has emerged in the 21st century as a prime force reshaping private and public domains, accelerating the digitalization of the government. Its use in central administrative procedures—planning, decision-making, implementation, and monitoring—has challenged beyond efficiency and effectiveness to prime values of public administration like democracy, participation, moral responsibility, and accountability. Whereas AI empowers governments to sift through vast amounts of data, create predictive models, reduce human error, and personalize services, it also has risks such as algorithmic bias, blackbox decision-making, privacy exposure, and limited auditability of complex systems. All these tensions are more serious in democratic administration, since it relies on citizens' participation, equal distribution of services, and politicians' accountability, and delegating power to algorithms may infringe upon these premises and undermine citizens' confidence in institutions and result in algorithmic governance as a concept. Global efforts towards balancing technological progress with universal rights emphasize the necessity to secure human dignity in digital public administration. Digitalization in Turkey has advanced through the likes of e-Government, MERNIS, UYAP, and e-School, and while AI is not yet widely used for decision-making, developments like municipal smart city initiatives, ministerial data analysis departments, and

[1] Ankara University, Political Science Institute, Political Science and Public Administration, moizuppal17@gmail.com, ORCID: 0009-0009-1421-3334

strategic institutional reports demonstrate growing interest. But key challenges remain, including legal and regulatory deficiencies, lack of managerial and technical capabilities, absence of institutionalized ethics, low public awareness, organizational resistance to change, lack of interoperability, lack of skills, and heightened cybersecurity threats. Drawing on comparative e-governance experiences in other countries and global normative developments, this study utilizes qualitative, literature-informed analysis to explore the political, legal, and managerial environments surrounding AI uptake in Turkey, arguing that its public sector deployment is not merely a technical but also a political and ethical choice requiring multi-stakeholder coordination to achieve sustainable, inclusive, and citizen-oriented digital transformation.

Keywords: Accountability, Algorithmic Governance, Artificial Intelligence (AI), Cybersecurity, Data Governance, Democracy.

Yapay Zekâ ve Kamu Yönetimi: Dijital Dönüşümün Siyasal, Hukuksal ve İdari Boyutlarına İlişkin Bir İnceleme

Özet

Yapay zekâ (YZ), 21. yüzyılda hem özel hem de kamu alanlarını yeniden şekillendiren bir unsur olarak ortaya çıkarak hükümetlerin dijital dönüşümünü hızlandırmıştır. YZ’nin planlama, karar alma, uygulama ve izleme gibi idari süreçlerin merkezinde kullanılmaya başlanması, yalnızca etkinlik ve verimlilikle sınırlı olmayan; demokrasi, katılım, ahlaki sorumluluk ve hesap verebilirlik gibi kamu yönetiminin temel değerlerini de etkileyen sorunlar yaratmaktadır. YZ, hükümetlerin büyük veri setlerini işleyebilmesini, öngörü modelleri geliştirmesini, insan hatalarını azaltmasını ve hizmetleri vatandaşlara göre uyarlamasını sağlarken; algoritmik önyargı, şeffaf olmayan karar alma süreçleri, mahremiyet ihlalleri ve karmaşık sistemlerin düşük denetlenebilirliği gibi ciddi kaygılara da yol açmaktadır. Bu gerilimler, vatandaş katılımı, hizmet sunumunda eşitlik ve yetkililerin hesap verebilirliği üzerine kurulu demokratik yönetişim açısından özellikle belirgindir; gücün algoritmalara devredilmesi bu değerleri zedeleyerek kamu güvenini sarsma tehlikesi taşımakta ve “algoritmik yönetişim” kavramının doğmasına zemin hazırlamaktadır. Küresel ölçekte geliştirilen etik ilkeler, teknolojik yeniliği temel haklarla uzlaştırma çabalarını yansıtırken; kamu yönetiminin dijital çağda erişilebilir olmasının insan onuruyla bütünleşmiş bir vatandaşlık unsuru olduğu vurgusu da öne çıkmaktadır. Türkiye'de dijital dönüşüm; e-Devlet Kapısı, MERNİS, UYAP ve e-Okul gibi projelerle ivme kazanmış, YZ'nin karar alma süreçlerinde yaygın uygulamaları henüz sınırlı olsa da akıllı şehir girişimleri, bakanlıkların veri analizi birimleri ve Yükseköğretim Kurulu'nun stratejik çalışmaları bu alandaki artan ilgiyi göstermektedir. Ancak hukuk ve düzenleme eksiklikleri, yönetsel ve teknik kapasite yetersizlikleri, kurumsallaşmış etik standartların olmayışı ve kamu farkındalığının düşüklüğü gibi engeller devam etmekte; bu durum örgütsel değişime direnç, sınırlı birlikte çalışabilirlik, beceri açıkları ve artan siber güvenlik tehditleri gibi küresel zorluklarla da örtüşmektedir. ABD, Güney Kore, Estonya ve AB ülkelerindeki e-yönetişim deneyimlerinin yanı sıra dünyadaki normatif gelişmeleri karşılaştırmalı olarak ele alan bu çalışma, YZ'nin Türkiye'deki siyasal, hukuksal ve yönetsel etkilerini nitel ve literatür temelli bir yaklaşımla incelemekte; kamu sektöründe YZ'nin yalnızca teknik bir konu değil, aynı zamanda siyasi ve etik bir tercih olduğunu ve sorumlu, insan odaklı, kapsayıcı bir dijital dönüşüm için çok paydaşlı iş birliğinin gerekliliğini savunmaktadır.

Anahtar Kelimeler: Hesap Verebilirlik, Algoritmik Yönetişim, Yapay Zeka (YZ), Siber Güvenlik, Veri Yönetişimi, Demokrasi.

Introduction

The 21st century has witnessed a universal digital revolution in every aspect of life, and public administration is no exception. Governments around the world are turning towards digital modes of enhancing service provision, operational effectiveness, and civic engagement. Of all of these, artificial intelligence (AI) has not just arrived as an enabler but also as a structural force transforming governance. Predictive social welfare analytics, computerized tax collection, and intelligent traffic management are a few of the tasks once performed by bureaucrats but now being carried out in conjunction with machine systems.

While its potential is enormous, AI growth in public administration is not without controversy. While AI holds tremendous benefits—effective bureaucracy, fewer errors, faster processing, and services more aligned to citizens' needs—it also raises deep questions about democracy, transparency, and accountability. Algorithms remain black boxes, with decision-making inaccessible to citizens and administrators, undermining trust and reinforcing deep biases in data. Moreover, mass data collection, on which AI depends, poses serious threats to privacy and human rights. These legitimacy-efficiency conflicts set the context for today's debates on AI in government, requiring a balance of morally charged performance ideologies such as New Public Management with democratic values such as participation, equality, and accountability. Theory of "algorithmic governance" sharpens this debate by shifting power from elected or appointed officials to machine-based systems, posing the question: can AI in public administration assure legitimacy, inclusiveness, and respect for citizens' dignity?

The educational and constitutional significance of this topic is that it has implications for state accountability, rights, and responsibility. Corvalán's argument is that the Fourth Industrial Revolution obliges states to recognize "digital dignity" as a component of citizenship in the form of the right to have access to e-services, secure virtual identities, and benefit from inclusive technologies. International institutions such as the United Nations and the Organization of American States also endorse the use of ICT and AI responsibly for deepening democracy, containing inequality, and promoting inclusive development. Nevertheless, a large research gap exists: while world studies analyze the ethical, legal, and managerial concerns of AI in public administration, fewer studies focus on how these issues are unfolding in Turkey's rapidly digitalizing but institutionally constrained environment. The country offers a revealing case where initiatives like the e-Government portal, MERNIS, UYAP, and e-School demonstrate impressive leaps in digitization, yet AI use remains minimal due to legal loopholes, management shortcomings, lack of proper technical expertise, and poor digital literacy. Conversely, the youthfulness of Turkey's population structure, political incentives for digitalization, and smart-city projects present exciting opportunities for espousing AI governance responsibly and inclusively.

In this context, this paper tries to respond to two main questions: **(1) How do AI shape the political, legal, and administrative aspects of public administration? (2) What can Turkey achieve from global best practices in developing responsible, inclusive, and human-oriented AI policies?** To answer these questions, the paper first develops a theoretical framework that situates AI against the backdrop of contemporary governance discourses, followed by the presentation of the research approach. It then considers international good practice and the reality situation for AI in Turkish public administration, prior to providing policy proposals. The study ultimately concludes that AI in government is more a governance challenge than an issue of exclusively technical matters, requiring not only technical expertise but also political will, legal safeguard, and moral gravity. The future of Turkey in this context is hinged on adopting a democratic, open, and inclusive AI governance model.

Theoretical Framework

The inclusion of artificial intelligence in governance cannot be described merely as a technological innovation alone.

It is also a conceptual and institutional transformation that intersects with prevailing theories of public administration. There are four paradigms to which the current discussion is most pertinent: Weberian bureaucracy, New Public Management (NPM), Digital Era Governance (DEG), and Algorithmic Governance. AI problematizes and exceeds each of these paradigms.

Weberian Bureaucracy

The Weberian model values hierarchical domination, rule-based methods, and juridical accountability. Bureaucracy aims at generating impartiality, calculability, and equality before the law. Even in online settings, the model remains applicable for securing legal certainty and formal proceduralism. AI can assist bureaucracy by institutionalizing decision-making, restricting arbitrary discretion, and enhancing transparency via automated processes.

However, AI also conflicts with bureaucratic principles. Opaque machine-learning algorithms risk undermining the principle of accountability: if neither administrators nor citizens can explain a machine's decision, legal predictability weakens. Thus, the challenge is aligning AI's efficiency with the rule-based integrity of bureaucracy.

At the same time, Corvalán (2018) observes that with electronic governance, legal accountability must extend beyond safeguarding the rule of law and procedures to encompass safeguarding digital identity and digital dignity. Citizens are not just entitled to openness in processes but to safeguarding their digital existence as part of human dignity. This pushes Weberian accountability beyond procedures and rules to safeguarding citizens' core rights in the digital age.

New Public Management (NPM)

Emerging during the 1980s and 1990s, NPM is centered on efficiency, measurement of performance, and market-influenced reform. NPM treats citizens as customers and favors managerial autonomy. AI fits into this framework by enabling performance dashboards, predictive analytics, and customer-service automation.

Yet critics suggest that excessive focus on metrics is more likely to overlook equity and inclusiveness. For instance, AI-based personalization of services can potentially benefit digitally literate citizens but exclude marginalized groups. Thus, AI reinforces managerial rationality of NPM but also accentuates its weaknesses.

Kholov & Mamarasulov (2024) illustrate that without adequate digital literacy and the reskilling of labor, AI-driven NPM reforms tend to reinforce, rather than reduce, inequalities. Based on their research, many governments suffer from lacking skills in AI, data analytics, and cybersecurity, which limit their ability to use managerial tools responsibly.

Digital Era Governance (DEG)

DEG emphasizes bringing fragmented services, digital platforms, and citizen-centric participation together. AI falls naturally here: chatbots, automated portals, and participatory platforms can facilitate greater citizen participation. Governments can monitor social media to catch the pulse of the people or use AI to conduct large-scale consultations.

Yet DEG is based on broad digital literacy and access. In practice, AI adoption often widens the digital divide, benefitting urbanites at the expense of rural or older citizens. If exclusivity does not matter, AI-based DEG can widen inequality. Corvalán (2018) labels this conflict as the challenge of inclusive digital transformation. In his framework, democratic legitimacy is not only predicated on the effectiveness of services but also on ensuring citizens, regardless of geography or socio-economic status, can exercise their digital rights to the maximum. This resonates with the UN's digital divide principle of inclusivity as the cornerstone of legitimacy in the digital era.

Algorithmic Governance

Algorithmic governance refers to a mode of decision-making that is delegated to algorithms rather than human authorities. It relies on predictive analytics, real-time monitoring, and optimization. This paradigm shift has potential advantages (efficiency, fraud detection, crisis prevention) and risks (bias, unexplainability, legitimacy gaps). The greater the delegation to algorithms, the more critical issues of legality and legitimacy are. In contrast to bureaucracy, which bases authority on rule, algorithmic governance bases authority on data and models—producing what Corvalán refers to as "black-box power" that imperils democratic oversight.

Kholov & Mamarasulov (2024) also warn that algorithmic governance struggles with system interoperability and data standardization. Unless government databases can communicate effectively with one another, algorithmic tools risk fragmenting governance instead of uniting it. They also document mounting threats of cyberattacks as algorithmic systems expand, suggesting the need for secure infrastructure and standardized data platforms in a rush.

Towards a Hybrid Framework

In reality, AI adoption takes elements of all the paradigms. As Tkachenko et al. (2025) note, digital transformation does not replace traditional models but hybridizes them. Good governance therefore requires integrating:

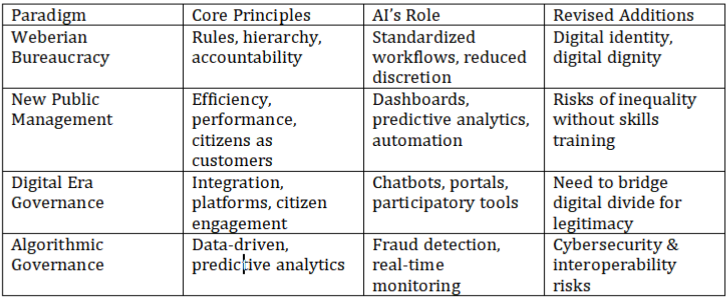
1. The legality and accountability of Weberian bureaucracy (enhanced to digital dignity and rights).
2. The efficiency and responsiveness of NPM (matched with capacity-building to avoid skills gaps).
3. The citizen-centric inclusiveness of DEG (while ensuring fair access to digital services).
4. The data-driven governance capacity (paired with cybersecurity and transparency safeguards).

Such a hybrid model understands that AI is both a technical and political choice. Its legitimacy is not only dependent on performance but also on ethical orientation, legal safeguards, and inclusive government. Without resolving structural challenges such as infrastructure deficits, organizational opposition, and digital literacy deficits (Kholov & Mamarasulov, 2024), any theoretical model cannot operate on the ground.

Application of Paradigms

I applied the four paradigms of governance—Weberian bureaucracy, New Public Management (NPM), Digital Era Governance (DEG), and Algorithmic Governance—while researching not as broad theories but as interpretive lenses to grasp Turkey's trajectory of AI-led digital transformation. Based on Weberian theory, I analyzed how AI-based technologies can empower rule-based accountability (e.g., through e-Government, MERNIS, and UYAP systems) and, simultaneously, undermine legal predictability because of ambiguous decision-making procedures. Referring to NPM, I examined how managerial reforms and efficiency gains delivered by AI align with performance-based metrics of Turkey and highlighted risks of inequality through skill imbalances and low digital literacy. With DEG's assistance, I explored the potential of AI-driven platforms, smart city projects, and participatory portals to foster citizen-centric governance but also criticized digital divides and questions of inclusivity that strongly resonate within Turkey. Finally, Algorithmic Governance summarized my assessment of black-box decision-making dangers, interoperability collapse, and cybersecurity threats Turkey is facing in its early adoption process. By applying these frameworks together, I designed a hybrid conceptualization of AI within Turkish public administration, which means legitimacy is not only about technical efficiency but also legal safety, citizen inclusion, and digital dignity assurance.

Table 1. Public Administration Paradigms in the Digital Age (Revised)



Paradigm	Core Principles	AI's Role	Revised Additions
Weberian Bureaucracy	Rules, hierarchy, accountability	Standardized workflows, reduced discretion	Digital identity, digital dignity
New Public Management	Efficiency, performance, citizens as customers	Dashboards, predictive analytics, automation	Risks of inequality without skills training
Digital Era Governance	Integration, platforms, citizen engagement	Chatbots, portals, participatory tools	Need to bridge digital divide for legitimacy
Algorithmic Governance	Data-driven, predictive analytics	Fraud detection, real-time monitoring	Cybersecurity & interoperability risks

Methodology

This study takes a qualitative, literature-based direction in examining the political, legal, and administrative implications of artificial intelligence for public administration in general. Rather than relying on primary surveys or interviews, the research synthesizes existing scholarship, official policy documents, and comparative case studies to generate findings relevant to the Turkish context. This strategy reflects the broad academic consensus that the governance challenge of

AI requires multidimensional and cross-national consideration rather than individual-country empirical surveys.

Research Methodology

The research follows a comparative case study methodology. The adoption of AI is vastly uneven across countries, and Estonia, Singapore, and South Korea are avant-garde examples, while the European Union represents a sound regulatory framework. Turkey's trajectory can be best understood by comparing its experience with these global benchmarks.

This strategy is consistent with Kholov & Mamarasulov (2024), who conduct cross-country evaluations of digital transformation challenges in order to identify technical, organizational, and ethical bottlenecks. In a similar direction, Corvalán (2018) emphasizes the need to include constitutional and human rights-based perspectives in such comparative studies, particularly in new democracies, in order to ensure that digital innovation remains committed to equality and dignity principles.

Data Sources

The study draws upon:

- Policy texts such as Turkey's e-Government portal archives, MERNIS (civil registry), UYAP (judiciary informatics), and Council of Higher Education (YÖK) reports.
- International frameworks such as the European Union's General Data Protection Regulation (GDPR), Ethics Guidelines for Trustworthy AI, and the forthcoming AI Act.
- Academic research in governance studies, administrative law, and political science.
- Case study research findings from Estonia's e-residency program, Singapore's Smart Nation project, and South Korea's AI-driven smart city initiatives.

- Complementary contributions from global research, e.g., Kholov & Mamarasulov's (2024) framework of overcoming systemic barriers, and Corvalán's (2018) legal-constitutional analysis of digital governance in Latin America.

Analytical Method

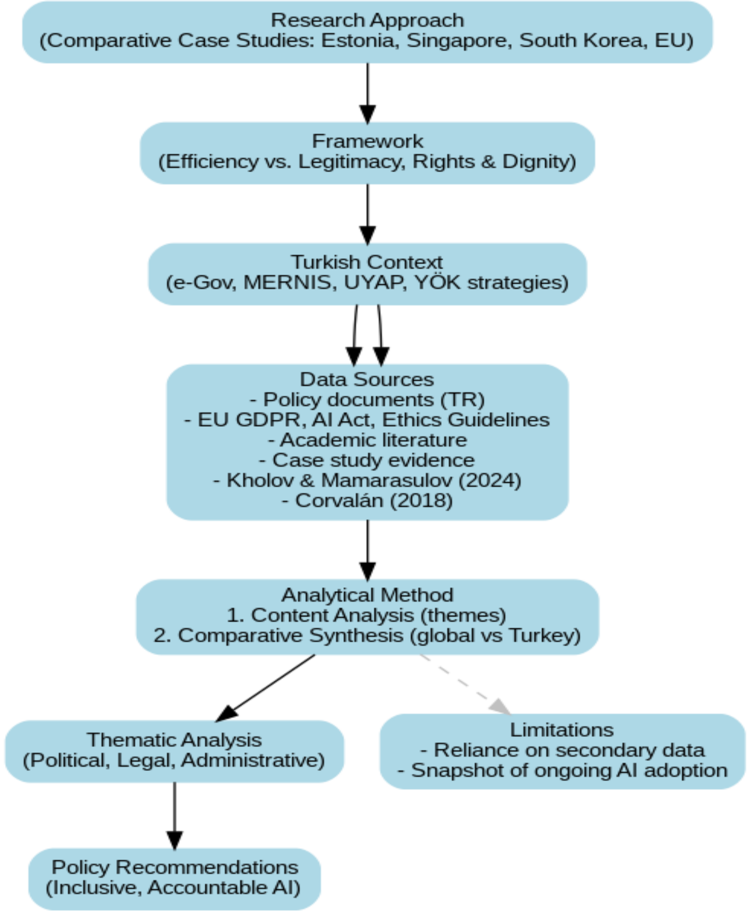
There are two levels of analysis:

1. Content analysis of documents and literature for revealing recurring themes (efficiency, accountability, participation, privacy, digital dignity, etc.).
2. Comparative synthesis to highlight lessons and contrasts with early adopters and Turkey's emerging practices, combining both pragmatic problem-solving approaches (as in Uzbekistan and global studies) and normative approaches grounded in rights and constitutional principles (as in Latin America).

Limitations

The study identifies two important limitations. It is founded on secondary data, which may overlook real-time institutional advances. Secondly, since AI adoption is in progress, findings indicate a snapshot rather than a final verdict. The approach is nonetheless a good foundation for identifying opportunities, challenges, and policy directions.

With the incorporation of global best practices, structural challenge analysis (Kholov & Mamarasulov, 2024), and rights-based approaches (Corvalán, 2018), this framework ensures a holistic perspective on AI and governance.



Research Approach
(Comparative Case Studies: Estonia, Singapore, South Korea, EU)

Framework
(Efficiency vs. Legitimacy, Rights & Dignity)

Turkish Context
(e-Gov, MERNIS, UYAP, YÖK strategies)

Data Sources
- Policy documents (TR)
- EU GDPR, AI Act, Ethics Guidelines
- Academic literature
- Case study evidence
- Kholov & Mamarasulov (2024)
- Corvalán (2018)

Analytical Method
1. Content Analysis (themes)
2. Comparative Synthesis (global vs Turkey)

Thematic Analysis
(Political, Legal, Administrative)

Limitations
- Reliance on secondary data
- Snapshot of ongoing AI adoption

Policy Recommendations
(Inclusive, Accountable AI)

World Experience with AI in Public Administration

The use of artificial intelligence in government has not developed uniformly across the globe. While some countries have set the pace for AI-based decision-making and delivery of services, others have been slow, prioritizing regulatory safeguarding over increased integration. The following section brings together some of the world's selected experiences—Estonia, Singapore, South Korea, and the European Union—to demonstrate both the potential and threat of AI for public administration, setting these within context with higher global trends and normative comprehension.

Estonia: The Digital Governance Pioneer

Estonia is widely recognized to be a world champion in e-governance. Starting from its X-Road platform, which enables secure data exchange among government databases, Estonia has applied AI to many public services. Its most visible attempt is KrattAI, which is a framework that enables AI-powered chatbots and decision-making systems across ministries. It involves citizens with the state through digital ID cards that protect their authentication online, providing access to over 99% of public services online.

Estonia's experience is evidence of how efficiency and trust can be parallel objectives brought about by AI. With transparency built into its mechanisms and with the people owning their data, Estonia has attained legitimacy as well as efficiency. Challenges also still exist, however, particularly to ensure that smaller municipalities have the same digital abilities as national institutions.

Singapore and South Korea: Smart Nation and Smart Cities

Singapore is another case of AI-driven governance through its Smart Nation Initiative. The government uses AI in traffic management, healthcare, and predictive infrastructure maintenance. Diagnostic assistance and customized treatment recommendations are provided through AI tools in healthcare, while urban mobility systems use real-time AI analytics to reduce traffic congestion.

It is also followed by South Korea through the manner in which it utilizes Smart City initiatives in Seoul and Busan. They include applying AI to city planning, public transportation, and environmental monitoring. The government has even launched AI learning programs to provide civil servants with data analytics capabilities for policy-making.

Both South Korea and Singapore illustrate the capacity of AI to improve not just efficiency but also quality of life. But they also present a key governance dilemma: efficient state capacity and central authority allow for fast AI adoption, but risk favoring efficiency over participation-based accountability.

The European Union: Regulation First

Compared to South Korea and Singapore, Estonia has proposed developing legal and ethical standards for AI in the political sphere. The General Data Protection Regulation (GDPR), the Ethics Guidelines for Trustworthy AI, and the forthcoming EU AI Act establish tough standards on transparency, fairness, and human oversight.

EU countries vary when it comes to adoption. Finland and the Netherlands, for instance, have established AI strategies that place emphasis on public sector innovation, while others have approached compliance in a more limited fashion. EU strength is derived from ensuring technological adoption keeps ethics in check. However, critics argue that this cautious approach can risk falling behind faster adopters like Singapore.

Global Trends and Comparative Data

While country-specific case studies provide detailed information, global figures indicate broader trends of adoption. As Kholov & Mamarasulov (2024) inform:

1. **Use of digital services:** 85% of the citizens of developed countries used one or more government services online

over the past year; for Estonia this was 98%, for Singapore 90%, and for South Korea 92%.

2. **AI in urban governance:** 60% of the world's cities have AI and data analytics already installed, for instance, to control traffic (Tokyo), public security (Barcelona), and healthcare tracking (New York).

3. **Investment growth:** Government investment in digitalization has increased by a growth rate of 25% globally over the past five years, with Asia-Pacific experiencing 30% growth (driven by smart cities), followed by 25% in Europe (driven by e-governance and digital identity) and 20% in North America (driven by cybersecurity and cloud).

These figures point out that AI adoption is not a speculative choice but an accelerating global trend. Yet, readiness is not evenly distributed: North America and Europe are far more than 8/10 ready in digital infrastructure, with Africa stuck at 3/10, illustrating current global inequalities.

International Normative Frameworks: UN and OAS

International organizations are defining digital transformation as both a question of efficiency but above all as a human right one. The United Nations and the Organization of American States (OAS) point out that ICT and AI must be oriented toward inclusive development, democratic participation, and reducing inequality (Corvalán, 2018).

The UN's Electronic Government for the People report (2012) points out:

- Expanding access to internet services,
- Reducing the digital divide, and
- Giving equal digital rights to vulnerable populations.

The OAS also promotes ICT for democracy, social justice, and development and invites states to see technology as a tool of universal inclusion.

Corvalán synthesizes this into digital dignity—arguing that the right of citizens to digitally interact with the state is at the heart of democratic legitimacy. Without such acknowledgment, efficiency-driven adoption can reinforce exclusion.

Lessons Learned from Global Practices

Global experiences reveal a number of core lessons for Turkey-like nations:

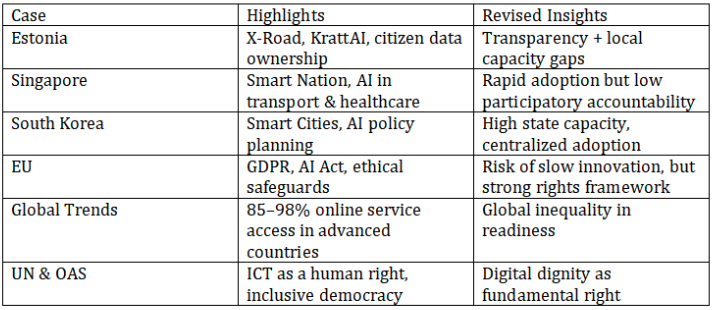
1. Strong Legal Frameworks Are Essential. The Estonian data ownership model and the regulatory emphasis of the EU both signal that citizen trust is based on clear rights and protection.

2. Centralized State Capacity Accelerates Adoption. Singapore and South Korea demonstrate that governments with high institutional capacity can rapidly embrace AI in urban services.

3. Inclusivity Must Come First. UN and OAS both identify reducing the digital divide as a pre-requisite of effective e-governance, with Kholov & Mamarasulov highlighting digital literacy as the primary bottleneck.

4. Balance Between Innovation and Regulation. Estonia is an excellent example of how rapid uptake could be complemented with transparency, with the EU illustrating the importance of not being seduced by technology euphoria and glossing over ethical issues.

Table 2. Comparative Global Experiences with AI in Governance (Revised)



Case	Highlights	Revised Insights
Estonia	X-Road, KrattAI, citizen data ownership	Transparency + local capacity gaps
Singapore	Smart Nation, AI in transport & healthcare	Rapid adoption but low participatory accountability
South Korea	Smart Cities, AI policy planning	High state capacity, centralized adoption
EU	GDPR, AI Act, ethical safeguards	Risk of slow innovation, but strong rights framework
Global Trends	85–98% online service access in advanced countries	Global inequality in readiness
UN & OAS	ICT as a human right, inclusive democracy	Digital dignity as fundamental right

5. Interoperability and Cybersecurity Cannot Be Ignored. Global experience confirms that technical infrastructure, secure data systems, and intersystem compatibility are prerequisites for sustainable AI integration.

Relevance of Turkey

For Turkey, the cases demonstrate that technical competence is insufficient to guarantee successful integration of AI. It must incorporate:

- Estonia's transparency and data ownership principles,
- Innovation led by the state of Singapore and South Korea,
- The EU's legal safeguarding and ethical regulation, and
- International experience in terms of inclusivity, interoperability, and cybersecurity.

Turkey's digital revolution is still in its early stages compared to these trailblazers, yet it can learn strategically from experience in creating an AI model of governance that balances efficiency with accountability and participation.

The Turkish Context

Turkey has undergone tremendous digital transformation in the past two decades. Government agencies become more dependent on digital portals for offering services to citizens. Some key initiatives are the e-Government portal, which combines over 6,000 services online; MERNIS, the database of central civil registry; UYAP, the judiciary's informatics system; and the e-School platform for educational services. These systems have facilitated administrative processes and made state services more convenient, demonstrating a significant move toward digital governance.

New AI Initiatives

With advanced digitalization, usage of AI in Turkey is at its early stages. There are a few instances like municipal smart city projects, limited usage of predictive analytics in the ministries, and reports of the Council of Higher Education (YÖK) Strategy that include AI in education policy. Some ministries have established data analytics offices to guide decision-making, and pilot schemes experiment with the application of machine learning to traffic management and fraud detection. Widespread use of AI for decision-making remains a distant goal.

Structural Challenges

Turkey's experience is an echo of the more general challenges listed by Kholov & Mamarasulov (2024):

1. Deficits in digital infrastructure: Although fundamental frameworks like e-Government are robust, regional governments often face poor connectivity and ancient infrastructures, reducing interoperability with national platforms.

2. System interoperability: Government databases are generally siloed, restricting data sharing and integrated AI-supported decision-making.

3. Institutional resistance: As in most countries, Turkish public sector organizations tend towards traditional processes. The shift to AI-facilitated workflows is usually followed by fear or institutional sluggishness.

4. Manpower deficit: Well-trained civil servants familiar with AI, data analysis, and cybersecurity are the exception rather than the rule, with the same concerns worldwide. Lacking capacity building investment, AI adoption could prove to be superficial.

5. Cybersecurity vulnerabilities: The more services become digitalized, the greater the chances of cyberattacks. Turkey has strengthened its cyber security in recent years, yet protecting sensitive personal data remains a persistent issue.

Legal and Ethical Gaps

Turkey does not have a comprehensive AI law or binding ethical framework such as the EU AI Act. While the Personal Data Protection Law (KVKK) provides some partial basis for data use regulation, none of it addresses algorithmic bias concerns, explainability, or attribution of duties to AI systems to any significant extent. Ethical rules for accountable AI remain to be institutionalized and are leaving behind a regulatory gap.

Here, Corvalán (2018) points are particularly relevant: democratic legitimacy of digital government hinges on ensuring digital dignity. In the case of Turkey, this means not only protecting people's privacy and data but also citizens' having rights to interact digitally with the state in an open and inclusive manner. Without such a recognition, AI adoption can contribute to undermining public trust as well as growing social inequality.

Opportunities

Besides these challenges, there are some positive strengths in Turkey:

- A young and digitally active population, with fertile ground for rapid uptake of new technologies.

- High political priority for digital transformation, expressed through smart city initiatives and national AI plans under consideration.

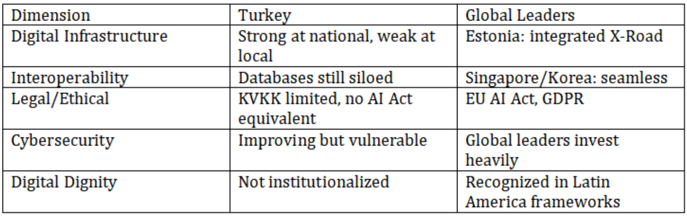
- Ability to jumpfrog by catching up with global leaders — Estonia's people-centric model, Singapore and South Korea's state-initiated smart cities, and the EU's robust regulation frameworks.

If Turkey is able to marry these strengths with legal protections, training of the workforce, and solid infrastructure, it may become a regional front-runner in AI-governance.

Challenges and Ethical Issues

The integration of AI in public administration presents not just possibilities but profound challenges that cut across technical, organizational, legal, and ethical issues. Such challenges ought to be appreciated as central drivers rather than secondary considerations of whether AI will strengthen or undermine democratic governance.

Table 3. Turkey vs. Global Leaders in AI and Governance (Revised)



Dimension	Turkey	Global Leaders
Digital Infrastructure	Strong at national, weak at local	Estonia: integrated X-Road
Interoperability	Databases still siloed	Singapore/Korea: seamless
Legal/Ethical	KVKK limited, no AI Act equivalent	EU AI Act, GDPR
Cybersecurity	Improving but vulnerable	Global leaders invest heavily
Digital Dignity	Not institutionalized	Recognized in Latin America frameworks

Technical Challenges

Infrastructural preparedness is perhaps the most pressing barrier. Whereas leading economies like Estonia and South Korea possess high connectivity and integrated digital networks, the majority of developing environments, such as much of Turkey, are tainted by unbalanced internet access and outdated infrastructure. Kholov & Mamarasulov (2024) point out that without sophisticated infrastructure, AI deployment can result in broken and volatile services.

Just as critical are interoperability and standardization of data problems. Turkish e-government portals such as MERNIS and UYAP remain partially siloed, limiting their ability to share information seamlessly. This undermines AI applications that are dependent on harmonized data sets to produce credible results.

Organizational Challenges

Culture and institutional reform need to accompany the transition to AI-powered government. Resistance to change remains a common obstacle, with public servants accustomed to conventional operations wary of or resistant to change as a perceived threat to their jobs. Besides, there is a critical skills shortage. Based on Kholov & Mamarasulov (2024), there are no sufficient government officials with advanced expertise in AI, data science, and cybersecurity. With no investment in digital literacy and focused training, the risk is that AI will be underutilized or abused.

Legal and Ethical Challenges

AI increases the current challenges of privacy and data protection. In Turkey, the KVKK establishes the ground for the management of data but not algorithmic decision-making, conditions of transparency, or data protection against bias yet.

Algorithmic bias is a further ethical problem. AI systems trained on historical data may reinforce current disparities, for instance, discriminate against minority populations in the provision of services or law enforcement applications.

Cybersecurity risks also increase. With increasing digitalization of services, governments are at greater risk of cyber attacks on sensitive citizen data. Kholov & Mamarasulov (2024) argue that robust cybersecurity architectures—like regular audits, incident response plans, and resilience mechanisms—must become a necessary component of AI adoption.

Democratic and Human Rights Issues

Beyond technical and organizational dimensions, AI in government subverts basic democratic values. Algorithmic decision-making subverts the accountability principle: since the consequence is determined by opaque algorithms, it is difficult to make citizens—or even administrators—aware of, contest, or appeal to decisions. It creates what Corvalán (2018) has called "black-box power," as legitimacy is erased because power shifts from elected representatives to non-accountable systems.

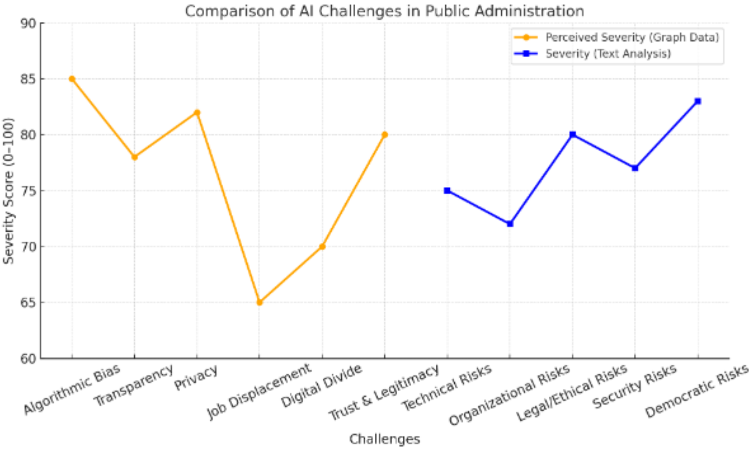
The other problem is digital dignity loss. Corvalán holds that digitally interacting with the state is an integral right tied to human dignity. Citizens not only need to be able to access digital services but also to be assured that their digital identity is protected, their interactions are transparent, and their rights are protected in the digital world. Absent such rights-based foundations, AI can exacerbate divisions, exclude marginalized groups, and destroy citizen confidence in the government.

Synthesis

Together, these dilemmas outline the multidimensional risks of AI in public administration:

- Technical risks: insufficient infrastructure, low interoperability, low data standards.
- Organizational risks: resistance to change, low digital literacy, skills mismatch among employees.
- Legal/ethical risks: privacy infringement, algorithmic bias, absence of clear decision-making.
- Security risks: cyberattacks, data intrusion, low resilience controls.
- Democratic risks: loss of accountability, participation, and digital dignity.

For Turkey, it is critical to be aware of these dangers. Unless they are tackled, AI risks becoming a tool of modernization and a source of inequality and democratic deficit. Yet, by the integration of protections—technical, institutional, and ethical—Turkey can ensure that AI enhances efficiency along with legitimacy in governance.



Policy Proposals and Future Directions

Technical innovation alone but conscious political, legal, and ethical choices are required for the effective integration of AI into public administration. Based on international practice and the current trajectory of Turkey, some policy proposals emerge to make digital transformation responsible, inclusive, and human-centered.

Enacting Legal and Ethical Cornerstones

1. Pass a national AI law or policy aligned with international standards such as the EU AI Act. This should explicitly address algorithmic transparency, explainability, and accountability.

2. Establish ethical AI guidelines for public administration, from the EU's Ethics Guidelines for Trustworthy AI but adapted to the context of Turkey.

3. Maintain digital dignity. As Corvalán (2018) reminds us, citizens have the right to digitally engage with the state in a way that maintains their identity, privacy, and dignity. Protection under the law must ensure digital engagement is transparent, equitable, and without discrimination.

Developing Technical and Organizational Capacity

- Foster digital maturity assessment among ministries and municipalities to evaluate infrastructure readiness, personnel capability, and services integration (Kholov & Mamarasulov, 2024).

- Develop specific transformation roadmaps with milestones, resources, and accountability for AI adoption.

- Furnish interoperability standards to facilitate unhampered communication among different government databases and platforms.

- Amplify digital literacy courses for civil servants and citizens, focusing on AI basics, data protection, and cybersecurity awareness.

- Invest in data science, machine learning, and ethical AI specialized training for core administrative staff.

Strengthening Cybersecurity and Data Protection

- Implement robust cybersecurity practices like regular auditing, penetration testing, and incident response plans.

- Accept standardized data formats and protocols to reduce dangers of fragmented systems.

- Increase people's awareness about data protection in the form of campaigns promoting trust in digital services.

Promoting Inclusiveness and Reducing the Digital Divide

- Provide broadband access and connectivity to rural and underserved regions to ensure even access to AI-enabled services.

- Create inclusive digital services for disadvantaged groups such as older persons, rural communities, and persons with disabilities.

- Guarantee universal digital rights based on UN and OAS principles that are centered on technology as a driver of inclusive growth.

Encouraging Public-Private Partnerships and Innovation

- Maximize private sector and university partnerships to acquire access to AI development, data analytics, and cybersecurity skills.

- Foster innovation labs and pilot initiatives through which AI solutions can be tested in low-risk environments before massive-scale deployment.

- Power blockchain and digital twin technology to maximize transparency, simplify infrastructure management, and safeguard files, as predicted by changing global trends (Kholov & Mamarasulov, 2024).

Future Prospects

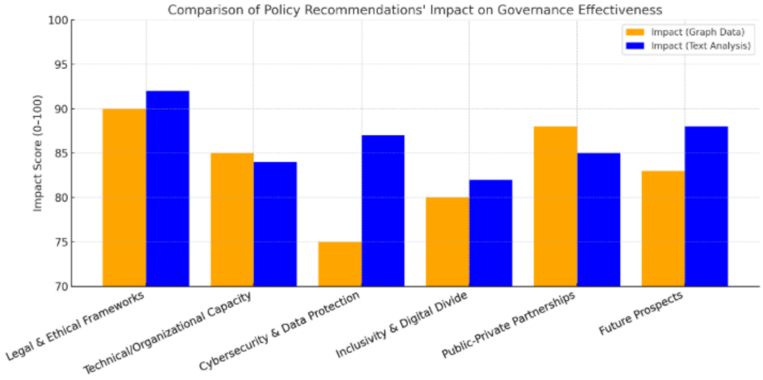
In the future, the future of AI in public administration will rely upon technical innovations, as well as political choices. Globally, trends show:

1. Deeper integration of machine learning and AI in core administrative tasks.

2. Appearing of smart cities and IoT-based governance, particularly at the city level.

3. Increasing reliance on blockchain for secure transactions and tamper-proof records.

4. Increasing emphasis on ensuring data privacy and incorporating ethics into digital governance.



For Turkey, the challenge is resisting one-dimensional pursuit of efficiency and instead finding a balanced perspective that includes efficiency, inclusiveness, transparency, and citizen rights. Turkey can observe how digital change strengthens, not undermines, democratic governance by injecting digital dignity into its AI approach.

Conclusion

Artificial intelligence is an opportunity and a challenge both for modern governance. For public administration, AI holds the promise of more efficiency, faster decision-making, and more customized delivery of services. Yet it also carries profound risks for accountability, transparency, privacy, and equity. The global experience demonstrates that the secret lies in achieving a balance between technological brilliance and democratic legitimacy.

Progressive examples such as Estonia, Singapore, and South Korea demonstrate that AI can enhance efficiency and quality of services where state capacity and robust infrastructure exist. The EU reflects on the need to ensure ethical and legal guaranties through regulatory regimes. At the same time, international research indicates continued challenges—low interoperability, skills gaps, resistance to change, and cybersecurity threats (Kholov & Mamarasulov, 2024). These are not merely technological issues; they reflect behind them institutional and governance issues that must be addressed with intent.

For Turkey, the digital leap acquired through web-based environments such as e-Government, MERNIS, UYAP, and e-School is the highlight of major advancements. But the adoption of AI in the country is yet to move out of the nascent stage. To proceed further, Turkey must overcome the dual challenge of building technical capacity and rights-based legitimacy. It needs to invest in the infrastructure, civil servant training, cybersecurity, and interoperable systems as well as infuse ethical safeguards and legal frameworks that uphold citizens' rights.

As Corvalán (2018) puts it, digital dignity - citizens' right to digitally interact with the state on terms that respect their privacy, equality, and identity - needs to be the bedrock of democratic rule in the digital era. Without this, uptake of AI risks eroding trust and exacerbating inequality. Conversely, by building digital dignity, Turkey can ensure that technological progress supports instead of erodes democratic legitimacy.

Finally, the trajectory of AI in public administration is not technologically driven but rests on political will, legal design, and institutional imagination. If Turkey manages to adopt international best practices, invest in rights-based protection and capacity, it can become a regional model in human-centered, transparent, and accountable AI government.

References

Ahmed, R. (2021). Smart cities and the Internet of Things: Governance implications. *Technology & Governance Review, 14*(3), 210–225.

Bannister, F., & Connolly, R. (2019). The future of e-government: A project management perspective. *Government Information Quarterly, 36*(1), 101–110.

Bertot, J., Jaeger, P., & Hansen, D. (2021). The impact of open government data on public administration. *Government Information Quarterly, 38*(2), 101590.

Chen, Y., & Wong, P. (2022). Blockchain for transparency in public procurement. *International Journal of E-Government Research, 18*(4), 32–48.

Chowdhury, R. H. (2024a). AI-powered analytics in public administration: A framework for efficiency. *Digital Governance Review, 12*(3), 45–59.

Chowdhury, R. H. (2024b). Blockchain for procurement integrity in e-governance. *International Journal of Public Sector Management, 37*(2), 211–226.

Dunleavy, P., Margetts, H., Bastow, S., & Tinkler, J. (2006). *Digital era governance: IT corporations, the state, and e-government.* Oxford University Press.

Estonian Government. (2020). *E-residency and digital state innovation.* Tallinn: Ministry of Economic Affairs and Communications.

European Commission. (2021). *Ethics guidelines for trustworthy AI.* Publications Office of the European Union. https://doi.org/10.2759/346720

Fountain, J. (2020). *Building the virtual state: Information technology and institutional change.* Brookings Institution Press.

Golea, D. G., Radu, A. F., & Cosa, S. O. (2025). Towards an innovative digital transformation of public administration in Romania through the implementation of artificial intelligence in the process of developing public policies in the field of health. *Technium Social Sciences Journal, 71,* 177–193.

González, H. (2022). Bridging the digital divide: Inclusion strategies in Latin America. *Journal of Public Innovation, 11*(1), 24–42.

Gupta, R., & Gupta, P. (2021). Blockchain and transparency in public procurement. *Public Money & Management, 41*(7), 509–517.

Heeks, R. (2020). *Digital government: Principles and practice.* Routledge.

Juan, G. (2023). The role of artificial intelligence in Latin American public governance. *Revista de Administración Pública y Digitalización, 12*(2), 45–60.

Kim, J. (2021). AI-driven decision-making in South Korea's smart city initiatives. *Asia-Pacific Journal of Public Policy, 9*(2), 134–150.

Kholov, A., & Mamarasulov, S. (2024). Problems of using artificial intelligence and digital transformation in the system of public administration. *International Conference Proceedings of the Republic of Armenia.* https://doi.org/10.55490/18290167-2024.sp-46

Martínez, L. (2022). Cybersecurity challenges in e-governance systems. *Journal of Digital Policy, 8*(1), 77–94.

Meijer, A. (2021). Understanding the effects of smart technologies on public governance. *Government Information Quarterly, 38*(3), 101613.

OECD. (2021). *Digital government strategies for sustainable development.* Organisation for Economic Co-operation and Development.

OECD. (2022). *The e-leaders handbook on the governance of digital government.* OECD Publishing.

Sharmin, S., & Chowdhury, R. H. (2025). Digital transformation in governance: The impact of e-governance on public administration and transparency. *Journal of Computer Science and Technology Studies, 7*(1), 362–379. https://doi.org/10.32996/jcsts.2025.7.1.27

Singapore Smart Nation Office. (2021). *Smart Nation: Transforming Singapore through technology.* Government of Singapore.

Smith, J., & Patel, K. (2021). Data privacy and AI regulation in public administration. *Global Governance Review, 6*(2), 55–73.

Tkachenko, V., Kotviakovskyi, Y., & Zinchenko, S. (2025). Contemporary European concepts of public administration in the context of digital transformation and their legal framework. *Public Administration and Law Review, 1*(21), 99–109. https://doi.org/10.36690/2674-5216-2025-1-99-109

Torres, M. (2020). Digital literacy programs for public servants: A Latin American perspective. *Public Administration and Technology Quarterly, 5*(3), 99–118.

United Nations. (2022). *E-Government Survey 2022: The future of digital government.* United Nations Department of Economic and Social Affairs.

UNDESA. (2022). *United Nations e-government survey 2022: The future of digital government.* United Nations Department of Economic and Social Affairs.

World Bank. (2020). *GovTech: Putting people first.* World Bank Group.

World Bank. (2022). *Digital government transformation: Lessons from global experience.* World Bank Publications.