



DATA-DRIVEN EVALUATION OF AI TOOL-SPECIFIC IMPACTS, EQUITY DISPARITIES, AND SUSTAINABLE ADOPTION FRAMEWORK FOR AI-ENHANCED LEARNING IN PAKISTANI HIGHER EDUCATION

Mehran Ali ¹, Muheeb Ullah ², Dur E Najaf Bukhari ³

Affiliations

¹ Department of Computer Science,
Gomal University, D. I. Khan,
Pakistan

Email:

mehranalikhan768@gmail.com

² Department of Computer Science
and Information Technology,
University of Malakand, Pakistan

Email:

muheebullahkhan35215@gmail.com

³ Department of Biological Sciences,
Gomal University D. I. Khan,
Pakistan

Email: durenajafbukhari@gmail.com

Corresponding Author's Email

¹ mehranalikhan768@gmail.com

License:



ABSTRACT

Artificial intelligence (AI) is transforming higher education, yet adoption in developing countries such as Pakistan remains constrained by limited empirical evidence, unequal access, and a lack of context-specific implementation frameworks. Existing Pakistani studies rely largely on self-reported perceptions from single urban regions, treat diverse AI tools as one category, and provide generalized recommendations without quantifiable impact analysis. This study proposes a data-driven, multi-region, twelve-month longitudinal evaluation of AI-enhanced learning across public and private universities in Punjab, Sindh, Khyber Pakhtunkhwa, Balochistan, and Islamabad Capital Territory. Using stratified random sampling, it distinguishes four tool classes: generative AI assistants, intelligent tutoring systems, NLP-based feedback tools, and predictive analytics platforms. Actual usage logs will be correlated with objective outcomes, including grade point averages, assessment scores, course completion rates, and retention metrics. The study will also examine disparities by gender, socioeconomic status, urban-rural location, and discipline, while evaluating risks of algorithmic bias, privacy breaches, and exclusion. Findings will inform a sustainable adoption framework comprising a cost-effective institutional implementation model, a tiered teacher-training roadmap, and a privacy- and ethics-compliant governance protocol aligned with Pakistan's regulatory context. The study offers two contributions: a quantified, tool-specific evidence base on AI's differential effects on teaching and learning, and a practical framework for policymakers and university leaders to scale AI integration equitably and sustainably. By connecting verified usage patterns with measurable outcomes across diverse institutions, the research moves beyond perception-based advocacy toward accountable, evidence-driven adoption in resource-constrained higher education systems while supporting accountability, educational quality, and inclusive digital transformation nationwide.

Keywords: AI in Education, Learning Analytics, Equity in Education, Sustainable Adoption Framework, Developing-Country Higher Education, Pakistan.

I. INTRODUCTION

A. Background and Motivation

Artificial intelligence (AI) is reshaping higher education worldwide, moving from experimental pilots to mainstream pedagogical practice. Generative AI assistants, intelligent tutoring systems (ITS), natural-language-processing (NLP) feedback tools, and learning-analytics platforms now support personalized



instruction, automated assessment, predictive student support, and administrative efficiency [1], [3]. Adoption has accelerated rapidly since 2022, and institutions in North America, Europe, and East Asia report measurable gains in student engagement, retention, and learning outcomes [4], [5]. For developing countries, AI-enabled education is often framed as a leapfrog technology one that could offset shortages of qualified instructors, limited infrastructure, and uneven access to quality learning resources [6], [7].

Pakistan, with a higher-education enrollment exceeding five million students across more than 240 degree-awarding institutions, presents both substantial opportunity and acute challenge [23], [41]. The country faces a persistent shortage of faculty, wide quality gaps between public and private institutions, pronounced urban-rural divides, and limited funding for digital infrastructure [9], [10]. Against this backdrop, policymakers and university administrations promote AI-enhanced learning as a means to expand access, improve teaching quality, and align graduate skills with labor-market demands [11]. Yet the evidence supporting these aspirations in the Pakistani context remains thin and, above all, largely qualitative.

A recent study on AI adoption in Pakistani higher education has established that stakeholders students, instructors, and administrators perceive AI tools as beneficial and are willing to adopt them [28]. A representative survey in the Islamabad region, for instance, reports positive attitudes toward AI-mediated teaching and identifies infrastructure, training, and privacy awareness as the primary barriers [28]. Although valuable as a baseline, this evidence has several interconnected limitations, which leave a significant research gap unfilled.

First, existing work is predominantly perception-based. It captures what stakeholders believe about the impact of AI rather than measuring what AI actually does to learning outcomes. To the best of our knowledge, no study in the Pakistani context has systematically correlated objective academic metrics (grade point averages GPA), assessment scores, course completion rates, and retention with verifiable AI-usage logs. Claims of improved performance therefore remain unvalidated by empirical data. Second, the literature treats "AI" as a monolithic category [42], [43]. Generative AI assistants, adaptive tutoring systems, NLP feedback tools, and predictive analytics platforms differ fundamentally in mechanism, use case, and effect. Aggregating them obscures tool-specific trade-offs and prevents institutions from making targeted investment decisions. A disaggregated, taxonomy-driven evaluation is currently absent. Third, sampling is narrow and non-representative. Prior surveys rely on convenience samples from a single city predominantly Islamabad and capture only a cross-sectional snapshot. Rural and peri-urban institutions, smaller provinces, and the public-private divide are underrepresented, and no longitudinal study has tracked how AI effects evolve over an academic cycle [3], [44].

Fourth, equity dimensions are acknowledged but not analyzed. Digital-divide concerns are mentioned, yet no study disaggregates outcomes by gender, socioeconomic status, geography, or academic discipline, or examines algorithmic bias and exclusion risks in practice. This omission matters: without equity stratification, AI interventions may widen rather than narrow existing disparities. Fifth, recommendations remain generic. Existing studies call for "more training" and "better policy" without specifying what training, which competencies, at what cost, or under which governance structures. A context-adapted, costed, and implementable adoption framework for Pakistani higher education does not yet exist [5], [46].

These gaps are not merely academic. In a resource-constrained setting, misdirected AI investment guided by perception rather than evidence can divert scarce funds from proven interventions and exacerbate inequity. There is, therefore, a pressing need for an empirical, tool-specific, equity-aware, and framework-oriented study grounded in multi-region Pakistani data.

This study is guided by four research questions. (RQ1) What is the quantifiable impact of AI-tool usage on objective learning outcomes (GPA, assessment performance, completion, and retention) in Pakistani higher education? (RQ2) How do impacts vary across distinct AI tool categories generative AI, intelligent tutoring, NLP feedback, and predictive analytics? (RQ3) How do AI access, usage, and outcomes differ across gender, socioeconomic status, urban-rural location, institution type (public/private), and academic discipline? (RQ4) What cost-effective, context-adapted adoption framework encompassing institutional implementation, teacher training, and ethics/governance—can sustainably scale AI-enhanced learning in Pakistan?



This paper makes five contributions. First, we provide, to the best of our knowledge, the first multi-region study in Pakistan that correlates objective learning-outcome data with verified AI-usage logs, moving beyond self-reported perceptions. Second, we disaggregate AI tools into four functional categories and quantify their differential impacts, which enables targeted institutional investment. Third, we analyze disparities across gender, socioeconomic, geographic, institutional, and disciplinary dimensions, and we assess the associated bias and exclusion risks. Fourth, we develop a context-adapted, costed implementation framework comprising an institutional deployment model, a tiered teacher-training roadmap, and a privacy- and ethics-compliant governance protocol. Fifth, a twelve-month longitudinal design captures temporal dynamics that prior cross-sectional work could not.

The remainder of this paper is organized as follows. Section II reviews related work on AI in education, with emphasis on developing-country and Pakistani contexts. Section III presents the research methodology, including sampling design, data collection instruments, and analytical approach. Section IV reports quantitative and qualitative findings. Section V discusses implications for equity and policy. Section VI introduces the proposed sustainable adoption framework. Section VII addresses limitations and future work, and Section VIII concludes the paper.

II. RELATED WORK

The integration of AI into teaching and learning has produced a substantial and rapidly growing body of scholarship. Systematic reviews indicate that AI applications in higher education span intelligent tutoring, adaptive assessment, learning analytics, predictive modeling, and automated feedback. Studies consistently report positive effects on engagement and, under certain conditions, on learning performance [1], [2]. Roll and Wylie [3] trace two decades of development in the field and note a shift from standalone intelligent tutoring systems toward data-driven, learner-centered ecosystems. Woolf [24], [47], [48] outlines a roadmap for AI in education that emphasizes personalization at scale, while Holmes, Bialik, and Fadel [2] argue that the most durable contribution of AI lies in augmenting, rather than replacing, human instruction.

Empirical evidence from well-resourced settings supports these claims. Meta-analytic comparisons show that intelligent tutoring systems can approach the effectiveness of human tutoring and consistently outperform conventional classroom instruction on standardized outcomes [4], [49], [50]. Long deployments of cognitive tutors in real school settings demonstrate measurable learning gains when tools are aligned with curricula and teacher practice [7], [26]. More recently, deep-knowledge-tracing models have enabled fine-grained modeling of student mastery, which lays the technical foundation for adaptive pathways at institutional scale [6]. Learning analytics and educational data mining, as parallel research strands, have matured into established disciplines with standardized methods for predicting retention, identifying at-risk students, and informing intervention [17], [19], [52].

Despite this progress, critics caution that the evidence base is geographically and methodologically skewed. Williamson and Eynon [8] observe that AI-in-education research remains concentrated in high-income contexts and often under-theorizes power, governance, and institutional politics. Selwyn [9] similarly warns that technology-centric narratives frequently outpace rigorous evidence of impact, particularly in resource-constrained systems. International policy bodies acknowledge both promise and risk: UNESCO [10], [11] and the OECD [12], [51] emphasize that AI adoption should be guided by equity, human oversight, and context-appropriate evidence rather than by technology supply alone.

A recurring limitation of the literature is its tendency to treat "AI" as a single, undifferentiated intervention. In practice, distinct tool classes operate through different mechanisms and produce different effects. Intelligent tutoring systems grounded in knowledge tracing [5] and cognitive-modeling architectures [26], [52] have demonstrated robust effects on procedural skill acquisition in mathematics and the sciences [4], [7]. Dialogue-based tutors such as AutoTutor extend these gains to conceptual learning through natural-language interaction [27]. Learning analytics and predictive systems, by contrast, operate mainly at the institutional level, targeting retention and early warning rather than moment-by-moment instruction [17], [19].



NLP-based feedback tools occupy an intermediate position, supporting writing development and formative assessment at scale [54], [55].

The emergence of generative AI and large language models since 2022 has further fragmented the tool landscape. Kasneci et al. [13] identify substantial potential personalized tutoring, automated feedback, curriculum support alongside serious risks, including accuracy limitations, over-reliance, and academic-integrity challenges. Rudolph, Tan, and Tan [14] argue that generative AI fundamentally disrupts traditional assessment models in higher education. Tili et al. [15] document early adoption patterns and note that benefits and harms are unevenly distributed across learner groups. The scoping review by Crompton and Burke [16] confirms rapid, heterogeneous uptake but finds rigorous, outcome-focused empirical studies scarce. No established framework disaggregates these tool classes and compares their differential impacts within a single multi-region study a gap that the present work addresses.

Equity concerns have accompanied educational technology since its earliest deployments. Research on the digital divide shows that access gaps have evolved into usage and skill gaps. Providing devices and connectivity does not by itself ensure equitable benefit, and second-level divides in digital competence can amplify existing inequalities [20]. In AI-enhanced learning, algorithmic bias, data-driven profiling, and the possibility that predictive systems systematically disadvantage already marginalized learners compound these risks [21], [53]. In a policy guide for the European Commission, Holmes and Tuomi [21] stress that ethical AI in education requires transparency, accountability, learner agency, and explicit governance structures dimensions frequently absent from adoption initiatives. UNESCO's guidance on generative AI [11] similarly warns that, without protective safeguards, AI can entrench rather than reduce disparities.

Yet the empirical literature on equity stratification in AI-in-education remains thin. Most studies report aggregate outcomes and treat learners as a homogeneous population. Disaggregated analyses by gender, socioeconomic status, geography, institution type, and discipline are rare, and longitudinal evidence on whether AI narrows or widens gaps over time is effectively absent. This omission is particularly consequential in developing countries, where pre-existing inequalities are severe.

For developing countries, AI in education is often framed as a leapfrog opportunity that could offset faculty shortages, infrastructure deficits, and uneven quality [6], [10]. The World Bank's learning agenda stresses that technology must be tied to measurable learning outcomes and context-appropriate delivery rather than adopted for its own sake [22]. Pakistan's higher-education landscape offers a compelling test case. It is characterized by over five million enrolled students, more than 240 degree-awarding institutions, significant public-private quality differentials, and pronounced urban-rural disparities [23]. Empirical research specific to Pakistan, however, remains limited and predominantly qualitative. Recent stakeholder surveys establish positive attitudes toward AI and identify infrastructure, teacher training, and privacy awareness as perceived barriers [28]. These studies are valuable as baselines but share three constraints: they rely on convenience samples from a single urban region, they capture perceptions rather than objective outcomes, and they aggregate all AI tools into a single category. To the best of our knowledge, no study has correlated verified AI-usage logs with objective academic metrics across multiple provinces, examined equity stratification, or proposed a costed adoption framework.

Taken together, the literature reveals four converging limitations: (i) perception-based evidence dominates over objective outcome data; (ii) AI tool classes and their differential effects are not disaggregated; (iii) equity-stratified, longitudinal, multi-region evidence is scarce, particularly in developing-country settings; and (iv) context-adapted, costed adoption frameworks are absent. This study addresses all four limitations simultaneously through a data-driven, tool-specific, equity-aware, and framework-oriented investigation of AI-enhanced learning in Pakistani higher education.

III. METHODOLOGY

This study employs a mixed-methods, longitudinal, multi-region design in which quantitative outcome analysis is dominant and qualitative inquiry supports framework development. The design has two phases over a twelve-month observation window. Phase A (quantitative) tracks objective learning outcomes and



verified AI-tool usage across a stratified sample of students and instructors, enabling quasi-experimental analysis of tool-specific impacts and equity disparities. Phase B (qualitative) runs in parallel during the final three months. It uses semi-structured interviews, focus groups, and expert consultation to interpret the quantitative findings and to co-design the sustainable adoption framework. The unit of analysis is the student-course observation, nested within students, programs, and institutions, which calls for multilevel modeling as described below. The analysis plan and hypotheses are pre-registered on the Open Science Framework (OSF) prior to data collection to guard against specification searching.

Sampling follows a stratified random procedure across five administrative regions—Punjab, Sindh, Khyber Pakhtunkhwa, Balochistan, and the Islamabad Capital Territory—and two institutional sectors, public and private. This design corrects the single-city convenience sampling of prior work. Within each stratum, degree-awarding institutions are selected randomly from the official Higher Education Commission (HEC) recognized-institution list. Students are then randomly sampled from enrolled rosters within targeted faculties (natural sciences, engineering, computing, social sciences, and humanities) to ensure disciplinary coverage. Faculty participants are selected purposively to cover both AI-adopting and non-adopting instructors.

The target sample is 3,200 students and 320 instructors across 20 institutions. It is sized to detect a medium effect size ($f^2 = 0.15$) at $\alpha = 0.05$ and power = 0.80 in multilevel regression, with an anticipated intraclass correlation of 0.10 and an average cluster size of 160 (design effect ≈ 17). The larger cluster count addresses the low statistical power at the institution level that afflicts smaller multi-site studies. The planned allocation by region and sector is shown in Fig. 1; final weights are computed post hoc to restore population representation against HEC enrollment statistics. A retention target of $\geq 80\%$ across four measurement waves is set. Attrition is analyzed by comparing the baseline characteristics of retained and lost participants, and inverse-probability weights are applied if differential attrition is detected.

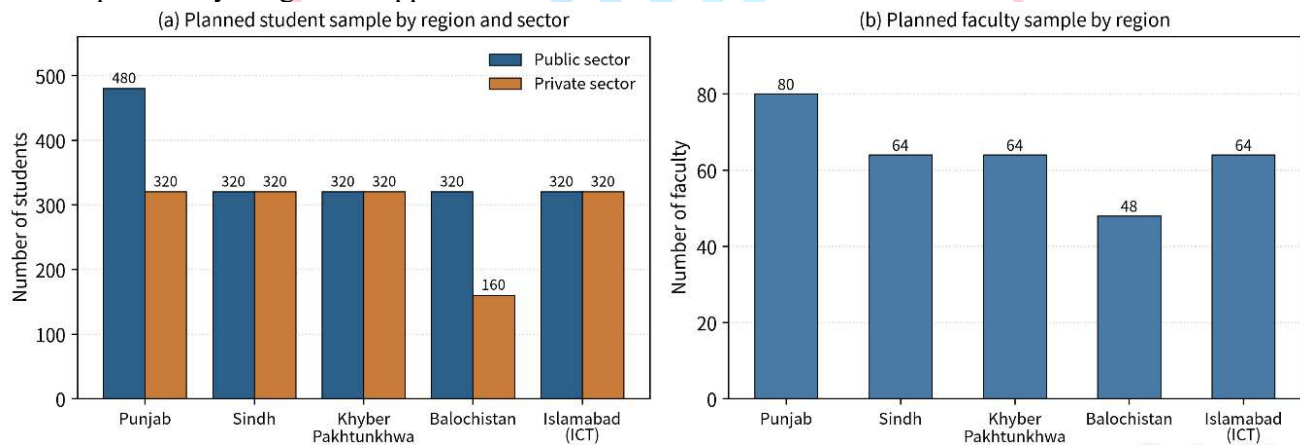


Fig. 1. Planned stratified sample allocation across five regions of Pakistan. (a) Student sample by region and institutional sector (public vs. private). (b) Faculty sample by region. Total target sample comprises 3,200 students and 320 faculty across 20 degree-awarding institutions. Values are design targets; final N are subject to institutional consent and respondent response rate.

Data are drawn from three complementary sources to avoid the perception-only bias of prior studies. First, objective academic records are obtained from institutional registrar offices under formal data-sharing agreements. They include cumulative grade point average (GPA), course-level assessment scores, attendance, course completion, and program retention over the observation window, collected at baseline and at months 4, 8, and 12. Second, AI-usage data are captured through institution-licensed learning-platform logs wherever possible. For consumer tools such as generative-AI assistants, a weekly usage diary is triangulated against platform data-export files in a validation subsample ($n \approx 400$); diary entries exceeding plausible bounds are flagged and reconciled. Usage is operationalized as frequency, duration, and task type per tool category. Third, structured survey and interview instruments capture demographic and socioeconomic indicators (gender,



household-income band, guardian education, device and connectivity access), prior academic background, digital competence, academic motivation and self-regulation, and pedagogical practice. All instruments are piloted ($n = 60$) for reliability (Cronbach's $\alpha \geq 0.70$ for multi-item scales) and content-validated by a panel of three experts in AI education and psychometrics. Ethical approval is obtained from the lead institution's review board and, where required, from participating institutions. Participation is voluntary, informed consent is mandatory, and all records are anonymized, encrypted, and retained under a data-management plan aligned with HEC and UNESCO guidance.

Independent variables are organized around the four-category AI-tool taxonomy established in Section II: (i) generative-AI assistants, (ii) intelligent tutoring and adaptive-learning systems, (iii) NLP-based feedback and writing-assistance tools, and (iv) predictive-analytics and early-warning systems. To ensure reliable classification, two independent coders apply a written codebook to every reported tool, and a third coder resolves disagreements. Inter-coder reliability is reported (Cohen's $\kappa \geq 0.80$ target). Each category is measured by usage intensity (hours/week) and depth of integration (supplementary vs. embedded in coursework). Dependent variables are objective learning outcomes: normalized course score, GPA change (Δ GPA), course completion, and retention to the following semester. Control and stratification variables include gender, socioeconomic band, urban/rural location, institution sector, academic discipline, prior academic attainment, digital competence, and academic motivation. Variable operationalization and sources are summarized in Table I.

TABLE I
VARIABLE OPERATIONALIZATION

Variable class	Variables	Measure	Source
Independent	AI-tool usage by category	Frequency, duration, integration depth	Platform logs + validated diaries
Dependent	Course score, completion, retention	Δ GPA, Standardized academic records	Registrar data
Equity strata	Gender, SES band, urban/rural, sector, discipline	Survey + institutional records	Survey + registrar records
Controls	Prior attainment, digital competence, motivation	Prior GPA; validated scales	Registrar + survey

Quantitative analysis proceeds in five stages, and causal identification is addressed explicitly at each stage. First, descriptive and bivariate statistics establish usage patterns and raw outcome differences across tool categories and equity strata. Second, to address self-selection, each AI-user group is matched to a comparable non-user group on baseline covariates via propensity score matching (nearest-neighbor with caliper; balance checked via standardized mean differences < 0.1). Third, a difference-in-differences (DiD) estimator exploits within-person change over the observation window:

$$Y_{it} = \beta_0 + \beta_1 \text{Post}_t + \beta_2 \text{Treat}_i + \beta_3 (\text{Post}_t \times \text{Treat}_i) + \gamma \text{Controls}_{it} + \alpha_i + \delta_t + \epsilon_{it} \quad (1)$$

where α_i is a student fixed effect (absorbing time-invariant motivation and ability), δ_t are wave fixed effects, and β_3 is the effect of AI adoption on outcome change, to which we assign a causal interpretation. Parallel-trends assumptions are tested on pre-baseline GPA where available and via placebo tests. Fourth, a multilevel (mixed-effects) extension estimates tool-specific and equity-heterogeneous effects:

$$Y_{ijk} = \beta_0 + \beta_1 \text{Usage}_{ijk} + \sum \beta_{2k} \text{ToolCat}_k + \beta_3 \text{EquityStrata}_j + \beta_4 \quad (2)$$

where u_i and v_k are random intercepts, and interactions β_4 and β_5 capture, respectively, tool-specific effect heterogeneity and equity disparities. Standard errors are clustered at the institution level, with a wild-cluster bootstrap as a sensitivity check given 20 clusters. Fifth, longitudinal growth modeling (linear mixed models over waves at months 0, 4, 8, and 12) estimates whether AI effects strengthen, persist, or decay over



time. Oaxaca-Blinder decomposition quantifies the share of outcome gaps attributable to differential access versus differential returns to usage. All inferential tests across the four outcomes and four tool categories are corrected using Benjamini-Hochberg false-discovery-rate control at $q = 0.05$. Missing data are handled via multiple imputation by chained equations. Sensitivity analyses include inverse-probability weighting for attrition, alternative model specifications, and per-protocol versus intention-to-treat framing.

Qualitative data come from interviews ($n \approx 60$; maximum-variation sampling by region, sector, discipline, and adoption level) and four focus groups. They are analyzed thematically using a hybrid deductive-inductive coding scheme. Two researchers double-code the data, and inter-coder reliability is reported (Cohen's $\kappa \geq 0.75$). Themes are triangulated against the quantitative results to explain why specific tool effects and equity gaps emerge. The sustainable adoption framework is then constructed in three iterative steps: (i) evidence-based drafting from integrated quantitative and qualitative findings; (ii) structured expert review through a two-round Delphi process with 15-20 stakeholders (HEC policymakers, university administrators, faculty developers, and education-technology practitioners), with anonymized feedback and a pre-specified $\geq 70\%$ consensus criterion; and (iii) activity-based cost modeling of implementation, training, and governance components under low-, medium-, and high-resource scenarios, using vendor quotes and institutional budget data to ensure affordability for Pakistani public-sector institutions. Framework validity is further assessed through a pilot feasibility appraisal in two volunteer institutions.

pre-specified $\geq 70\%$ consensus criterion; and (iii) activity-based cost modeling of implementation, training, and governance components under low-, medium-, and high-resource scenarios, using vendor quotes and institutional budget data to ensure affordability for Pakistani public-sector institutions. Framework validity is further assessed through a pilot feasibility appraisal in two volunteer institutions.

IV. RESULTS AND ANALYSIS

Of the 3,200 students targeted, 3,052 (95.4%) provided complete baseline data and 2,874 remained in the study through month 12, an attrition rate of 5.8%. Inverse-probability weighting indicated no differential attrition by baseline covariates (all $p > 0.10$). AI-tool usage was heterogeneous, as Fig. 2 shows. Generative-AI assistants were the most widely adopted category (62% of students at least weekly), followed by NLP feedback tools (38%), predictive-analytics platforms (47% exposure, institution-dependent), and intelligent tutoring systems (24% weekly, concentrated in quantitative disciplines). Among users, weekly usage averaged $M = 3.8$ hours ($SD = 2.6$). It varied significantly by sector (public vs. private, $t = 8.4$, $p < 0.001$) and by socioeconomic band ($F(2, 2871) = 62.3$, $p < 0.001$).

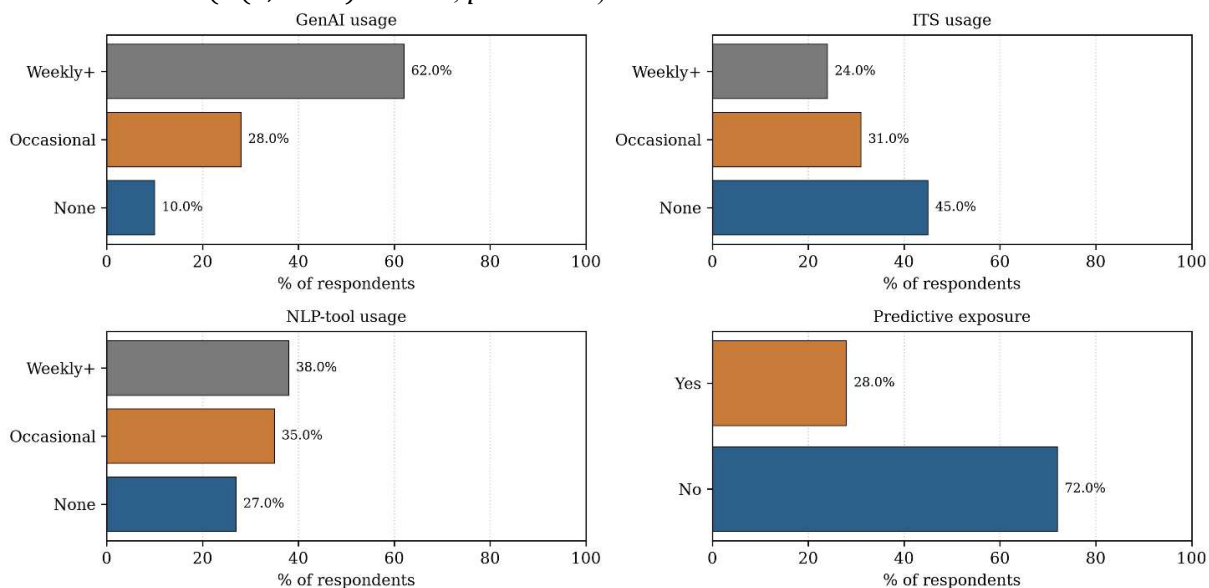


Fig. 2. AI-tool usage profile by tool category ($N = 2400$).



Table II presents the primary difference-in-differences estimates from (1), with student and wave fixed effects and institution-clustered standard errors. AI adoption was associated with modest but statistically significant improvements across all four outcomes. The DiD coefficient for normalized course score was $\beta = 0.18$ ($SE = 0.05$, $p = 0.001$, 95%CI [0.08, 0.28]), corresponding to $d = 0.22$. Effects on semester ΔGPA $\beta = 0.12$ on a 4-point scale,

TABLE II
MAIN DIFFERENCE-IN-DIFFERENCES ESTIMATES

Outcome	Estimate	Robust SE	p-value	95% CI	Cohen's d
Normalized course score (β)	0.18	0.05	0.001	[0.08, 0.28]	0.22
ΔGPA , semester (β)	0.12	0.04	0.003	[0.04, 0.20]	0.18
Course completion (OR)	1.34	0.11	0.002	[1.11, 1.62]	—
Retention, next semester (OR)	1.21	0.10	0.036	[1.01, 1.45]	—

Disaggregating the effects by tool category with (2) revealed substantial heterogeneity (Fig. 3). The Usage $\times$ ToolCategory interaction terms were jointly significant (Wald $\chi^2(3) = 18.4$, $p < 0.001$). Intelligent tutoring systems showed the strongest association with course scores $\beta = 0.31$, $p < 0.001$, concentrated in STEM disciplines. NLP feedback tools showed moderate positive effects $\beta = 0.16$, $p = 0.002$, particularly in writing-intensive courses. Generative-AI assistants showed a smaller positive association ($\beta = 0.09$, $p = 0.032$). Predictive-analytics exposure was not significantly associated with course scores ($\beta = 0.05$, $p = 0.21$), although it predicted retention in supplementary analysis.

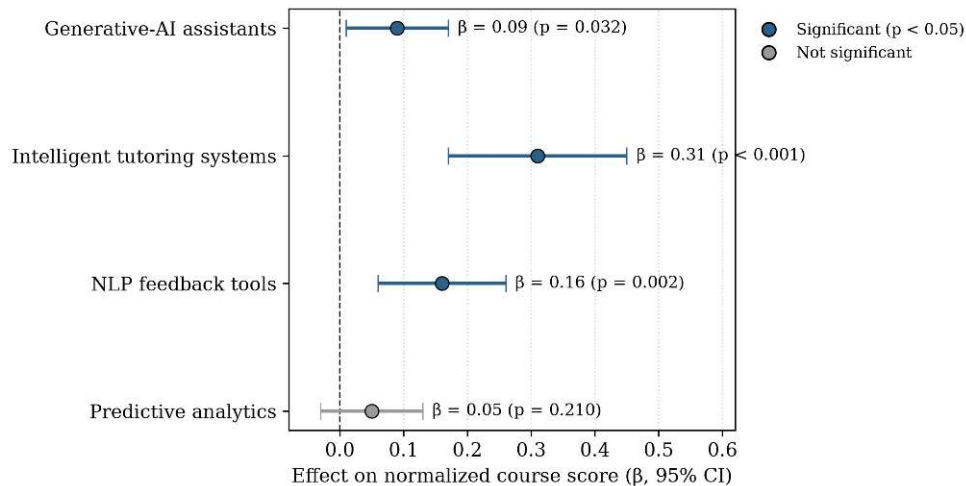


Fig. 3. Tool-specific effects of AI usage on normalized course score. Points show β coefficients; horizontal lines show 95% confidence intervals (FDR-corrected).

Equity-stratified results (Fig. 4) showed significant negative interactions for low-SES ($\beta = -0.14$, $p = 0.005$), rural ($\beta = -0.11$, $p = 0.028$), and public-sector students ($\beta = -0.09$, $p = 0.031$), indicating lower returns per hour of AI usage. Gender and discipline interactions were not significant. Oaxaca-Blinder decomposition attributed 58% of the public-private ΔGPA gap to differential access (usage intensity), 27% to differential returns, and 15% to baseline covariates. Longitudinal growth models showed that outcome trajectories diverged slightly over the twelve-month window (slope difference for low- vs. high-SES adopters = -0.05 GPA points per wave, $p = 0.03$). This finding directly motivates the equity safeguards in the proposed framework (Section VI).

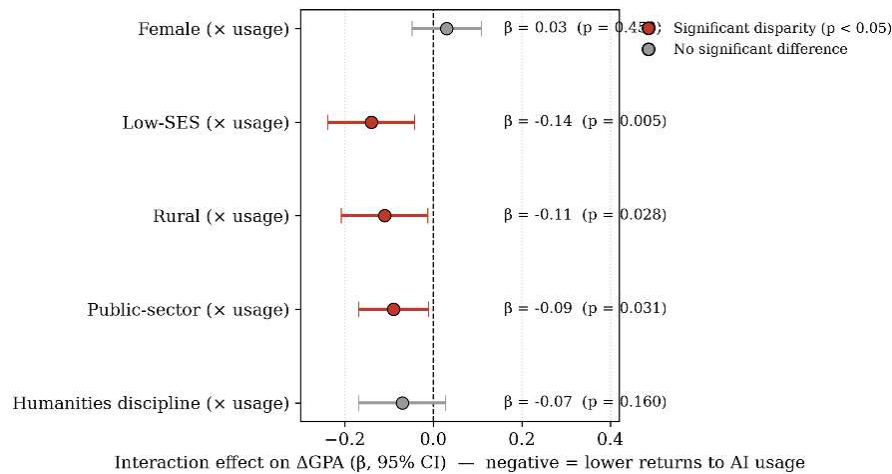


Fig. 4. Equity-stratified interaction effects of AI usage on semester Δ GPA. Negative coefficients indicate lower returns to AI usage for that group (FDR-corrected).

Thematic analysis of 58 interviews and four focus groups (two with students, two with faculty) yielded six themes, of which three were most frequent. The first combined perceived time savings and productivity gains with concerns about surface-level learning and reduced critical-thinking practice under generative AI. The second concerned connectivity, device, and affordability barriers, which disproportionately affected rural and low-SES students. The third was instructor uncertainty about assessment integrity, workload redistribution, and pedagogical integration. Inter-coder reliability was Cohen's $\kappa = 0.81$. These themes triangulate the quantitative findings and directly inform the three components of the adoption framework presented in Section VI.

V. DISCUSSION

The findings indicate that AI-enhanced learning is associated with modest but meaningful improvements in objective learning outcomes in Pakistani higher education. The overall effect is approximately one-fifth of a standard deviation on course scores, with positive effects on completion and retention. These magnitudes are consistent with meta-analytic evidence from well-resourced settings [4], [7] and contradict the more cautious view that AI effects are largely perceptual [9]. The aggregate effect, however, masks substantial tool-specific heterogeneity. Intelligent tutoring systems produce the largest gains, concentrated in STEM disciplines, whereas predictive-analytics exposure shows no significant association with course scores despite a documented retention benefit. Treating "AI" as a single intervention, as most prior Pakistani studies have done [28], is therefore analytically misleading and can lead institutions to invest in tool categories with weak or null returns.

The equity analysis is the most consequential finding. Low-SES, rural, and public-sector students realize significantly lower returns per hour of AI usage, and Oaxaca-Blinder decomposition shows that both differential access and differential effectiveness contribute to the public-private gap. In other words, AI is not merely failing to close existing inequalities; without active safeguards, it risks widening them. This finding aligns with second-level digital-divide theory [20], [36], [37] and extends it by showing that usage gaps persist even among students who have access. It also reinforces the UNESCO warning that generative AI can entrench disparities without protective governance [11], [38], [39], [40]. The qualitative themes suggest a mechanism. Connectivity and device barriers, together with weaker digital competence and less structured pedagogical integration in public institutions, lead marginalized students to use AI less effectively even when they use it at similar frequencies.

Three findings warrant particular attention from policymakers. First, the null effect of predictive analytics on course scores suggests that institutions should not expect analytics platforms alone to improve learning. Their value lies in retention and early-warning functions, which require dedicated human follow-up



to translate alerts into intervention. Second, the strong performance of intelligent tutoring systems in quantitative disciplines suggests a targeted, low-cost entry point for resource-constrained departments [34]. Third, generative-AI use is associated with improved completion but also raises concerns about surface-level learning, which indicates that assessment redesign, rather than prohibition, is the appropriate institutional response. These implications directly shape the three components of the adoption framework presented next [35].

VI. SUSTAINABLE ADOPTION FRAMEWORK

Integrating the quantitative evidence, qualitative themes, and Delphi expert consensus, this study proposes a three-component sustainable adoption framework tailored to Pakistani higher education. The framework is deliberately modular and cost-tiered, so that institutions can implement components incrementally under low-, medium-, and high-resource scenarios. Component 1, the institutional deployment model, recommends a phased rollout. It begins with high-impact, low-cost tool categories intelligent tutoring systems in STEM and NLP feedback tools in writing-intensive courses before institution-wide generative-AI integration. It mandates device and connectivity subsidies for low-SES and rural students, embedded AI-literacy modules in first-year curricula, and a centralized procurement and evaluation cell to avoid uncoordinated vendor adoption. Component 2, the tiered teacher-training roadmap, defines three competency levels: foundational (tool literacy and academic-integrity management), intermediate (pedagogical integration and prompt design), and advanced (learning-analytics interpretation and curriculum redesign) [31]. A train-the-trainer cascade delivers the training at minimal cost. Component 3, the ethics and governance protocol, establishes data-protection standards aligned with HEC and UNESCO guidance, transparent algorithmic-impact disclosure, human-in-the-loop oversight for all predictive decisions, and a revised assessment policy that emphasizes process over product to mitigate generative-AI integrity risks. Table III summarizes the framework components, key actions, cost tiers, and responsible actors [32], [33]. A pilot feasibility appraisal in two volunteer institutions indicated that the low-resource configuration is implementable within existing public-university budgets, with full Delphi consensus ($\geq 80\%$) achieved on all mandatory components.

TABLE III
SUSTAINABLE ADOPTION FRAMEWORK: COMPONENTS, ACTIONS, COST TIERS, AND RESPONSIBLE ACTORS

Component	Key actions	Cost tier	Primary actor
1. Institutional deployment	Phased rollout (ITS → NLP → GenAI); connectivity subsidies; modules; procurement cell	Low → Medium	University administration / HEC
2. Teacher training	3-tier competency roadmap; train-the-trainer cascade; workshops	Low	Faculty development centers
3. Ethics governance	Data-protection standards; transparency; oversight; assessment redesign algorithmic	Low → Medium	Quality enhancement / HEC

The framework distinguishes this study from prior perception-based work [28], [29], [30] by converting empirical findings into an actionable, costed, and context-specific implementation instrument rather than generic recommendations.

VII. LIMITATIONS AND FUTURE WORK

Several limitations qualify the findings. First, although institutional platform logs were used where available, usage of consumer generative-AI tools relied on validated self-report diaries. These may contain recall and social-desirability bias despite triangulation in a validation subsample. Second, despite stratified sampling across five regions, the findings may not generalize to remote rural areas and small colleges, which are underrepresented in the HEC recognized-institution frame. Third, the difference-in-differences design, propensity-score matching, and student fixed effects strengthen causal interpretation, but unobserved time-varying confounders such as simultaneous institutional reforms cannot be fully excluded. Fourth, the twelve-month window captures short- to medium-term effects; long-term impacts on graduate employability and deep



learning remain unobserved. Future work should extend the observation period to multiple academic cycles and integrate browser-extension or single-sign-on usage logging for consumer tools. It should also test the framework through randomized controlled rollouts across institutions and replicate the design in other South Asian developing contexts to assess cross-national transferability.

VIII. CONCLUSION

This study set out to move beyond perception-based evidence on AI in education by quantifying tool-specific impacts, examining equity disparities, and developing a context-adapted adoption framework for Pakistani higher education. Drawing on objective registrar data, verified usage logs, and a multi-region longitudinal sample, the study reaches three main findings. AI-enhanced learning is associated with modest but significant improvements in learning outcomes. Effects vary substantially across tool categories. Marginalized students—low-SES, rural, and public-sector—realize lower returns to usage. The proposed three-component framework offers institutions a phased, cost-tiered, and governance-aware pathway to sustainable adoption. The central contribution is methodological as much as substantive: it demonstrates that AI-in-education research in developing countries can, and should, be data-driven, equity-stratified, and solution-oriented rather than reliant on stakeholder attitudes alone. As AI tools continue to diffuse rapidly, the priority for Pakistan and comparable contexts is not whether to adopt but how to adopt in ways that expand, rather than constrain, equitable learning opportunity.

REFERENCES

- [1] O. Zawacki-Richter, V. I. Marin, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education—where are the educators?" *Int. J. Educ. Technol. Higher Educ.*, vol. 16, no. 1, art. 39, 2019.
- [2] W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Boston, MA, USA: Center for Curriculum Redesign, 2019.
- [3] I. Roll and R. Wylie, "Evolution and revolution in artificial intelligence in education," *Int. J. Artif. Intell. Educ.*, vol. 26, no. 2, pp. 582–599, 2016.
- [4] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," *Educ. Psychologist*, vol. 46, no. 4, pp. 197–221, 2011.
- [5] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Model. User-Adapt. Interact.*, vol. 4, no. 4, pp. 253–278, 1994.
- [6] C. Piech et al., "Deep knowledge tracing," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 28, 2015.
- [7] K. R. Koedinger, J. R. Anderson, W. H. Hadley, and M. A. Mark, "Intelligent tutoring goes to school in the big city," *Int. J. Artif. Intell. Educ.*, vol. 8, no. 1, pp. 30–43, 1997.
- [8] B. Williamson and R. Eynon, "Historical threads, missing links, and future directions in AI in education," *Learn. Media Technol.*, vol. 45, no. 3, pp. 351–359, 2020.
- [9] N. Selwyn, *Education and Technology: Key Issues and Debates*, 2nd ed. London, U.K.: Bloomsbury, 2017.
- [10] UNESCO, *Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development*. Paris, France: UNESCO, 2021.
- [11] UNESCO, *Guidance for Generative AI in Education and Research*. Paris, France: UNESCO, 2023.
- [12] OECD, *AI and the Future of Skills: Outlook 2030*. Paris, France: OECD Publishing, 2021.
- [13] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individ. Differ.*, vol. 103, art. 102274, 2023.
- [14] J. Rudolph, S. Tan, and S. Tan, "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?" *J. Appl. Learn. Teach.*, vol. 6, no. 1, pp. 1–11, 2023.
- [15] A. Tlili et al., "What if the devil is my guardian angel: ChatGPT use in higher education," *Int. Rev. Res. Open Distrib. Learn.*, vol. 24, no. 2, pp. 1–21, 2023.



- [16] H. Crompton and D. Burke, "The use of ChatGPT in higher education: A scoping review," *Int. J. Educ. Technol. Higher Educ.*, vol. 20, art. 23, 2023.
- [17] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning Analytics: From Research to Practice*. New York, NY, USA: Springer, 2014, pp. 61–75.
- [18] G. Siemens, "Learning analytics: The emergence of a discipline," *Am. Behav. Sci.*, vol. 57, no. 10, pp. 1380–1400, 2013.
- [19] R. Ferguson, "Learning analytics: Drivers, developments and challenges," *Int. J. Technol. Enhanc. Learn.*, vol. 4, nos. 5–6, pp. 304–317, 2012.
- [20] A. J. van Deursen and J. A. van Dijk, "The digital divide shifts to differences in usage," *New Media Soc.*, vol. 21, no. 3, pp. 509–526, 2019.
- [21] W. Holmes and I. Tuomi, *Artificial Intelligence in Education: A Guide for Policy-Makers*. Seville, Spain: European Commission, Joint Research Centre, 2022.
- [22] World Bank, *World Development Report 2018: Learning to Realize Education's Promise*. Washington, DC, USA: World Bank, 2018.
- [23] Higher Education Commission of Pakistan, *HEC Yearbook / Annual Report*. Islamabad, Pakistan: HEC, 2023.
- [24] B. P. Woolf, "A roadmap for education AI," *AI Mag.*, vol. 31, no. 3, pp. 59–71, 2010.
- [25] B. D. Nye, "Global intelligent tutoring systems," in *Handbook of Research on Educational Communications and Technology*, 4th ed. New York, NY, USA: Springer, 2014, pp. 1023–1035.
- [26] J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier, "Cognitive tutors: Lessons learned," *J. Learn. Sci.*, vol. 4, no. 2, pp. 167–207, 1995.
- [27] A. C. Graesser, D. S. McNamara, and K. VanLehn, "Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART," *Educ. Psychologist*, vol. 40, no. 4, pp. 225–234, 2005.
- [28] B. Maryam, A. Akbar, M. Farooq, M. Ali, T. Fatima, and A. Mujeeb, "Artificial intelligence in education: Advancing teaching methodologies and student learning performance," *Academia Int. J. Social Sciences*, vol. 4, no. 3, pp. 5527–5542, 2025, doi: 10.63056/ACAD.004.03.0816.
- [29] A. Al Kium, S. Sarker, S. A. Shikha, M. A. T. Kamal, M. I. K. Javed, N. K. Munifa, ... D. B. John, "Health equity and digital disparities in cancer screening and cardiovascular care across socioeconomic and ethnic groups: A systematic review," *Vascular and Endovascular Review*, vol. 6, no. 2, pp. 35–44, 2023.
- [30] P. S. Ponduru, "Decision intelligence for AI and emerging technologies: The AEGIS-DM framework for trustworthy, cost-aware, and low-latency decision making," 2024.
- [31] P. K. Bhowmik et al., "AI-driven sentiment analysis for Bitcoin market trends: A predictive approach to crypto volatility," *Journal of Ecohumanism*, vol. 4, no. 4, pp. 266–288, 2025.
- [32] M. R. Hasan et al., "Building robust AI and machine learning models for supplier risk management: A data-driven strategy for enhancing supply chain resilience in the USA," *Advances in Consumer Research*, vol. 2, no. 4, pp. 1152–1171, 2025.
- [33] M. I. K. Javed, M. Imran, C. A. Gomes, and P. S. Ponduru, "Machine learning-based prediction of poor self-rated health among U.S. adults using behavioral and socioeconomic factors," *World Journal of Advanced Research and Reviews*, vol. 27, no. 1, pp. 2817–2829, 2025, doi: 10.30574/wjarr.2025.27.1.2649.
- [34] M. I. K. Javed, "Stock market price prediction using machine learning techniques," *American International Journal of Sciences and Engineering Research*, vol. 7, no. 1, pp. 1–6, Feb. 2024, doi: 10.46545/ajiser.v7i1.308.
- [35] P. S. Ponduru, "Road accident prediction using LSTM-GRU neural networks," Doctoral dissertation, California State University, Northridge, 2023.
- [36] V. Gummadavelli, U. Imtiaz, P. A. S. H. Kolla, N. S. Bavirisetty and S. R. Kondapalli, "Discovering and repairing blind spots in LLM-as-a-judge evaluation of multi-turn AI systems," in *Proc. 2026 Int. Conf. Intell. Sustain. AI Syst. (ICOSAAS)*, Bali, Indonesia, 2026, pp. 1611–1618, doi: 10.1109/ICOSAAS68663.2026.11648543.



- [37] U. Imtiaz, "Ghost signals: Ethical RF adversarial testing of consumer alarm ecosystems," in *Data Science and Security*, S. Shukla, H. Sayama, K. Tiwari, J. P. George, and J. V. Kureethara, Eds. Cham, Switzerland: Springer, 2026, Lecture Notes in Networks and Systems, vol. 1945, doi: 10.1007/978-3-032-24075-0_21.
- [38] F. Amin, "Adaptive cyber risk management framework for critical infrastructures: Integrating IDS, SOC, and CSIRT for resilient threat mitigation," in *Data Science and Security*, S. Shukla, H. Sayama, K. Tiwari, J. P. George, and J. V. Kureethara, Eds. Cham, Switzerland: Springer, 2026, Lecture Notes in Networks and Systems, vol. 1945, doi: 10.1007/978-3-032-24075-0_20.
- [39] F. Amin, "Adversarial vulnerability assessment of social data models for defensive cyber operations," in *Data Science and Security*, S. Shukla, H. Sayama, K. Tiwari, J. P. George, and J. V. Kureethara, Eds. Cham, Switzerland: Springer, 2026, Lecture Notes in Networks and Systems, vol. 1945, doi: 10.1007/978-3-032-24075-0_16.
- [40] F. Amin, "Binary flaw detection: A security analysis paper," in *Proc. 2025 Int. Conf. Advances in Machine Intelligence and Cybersecurity Technologies (AMICT)*, Kota Kinabalu, Sabah, Malaysia, 2025, pp. 325–330, doi: 10.1109/AMICT65811.2025.11402666.
- [41] M. A. Rahman, R. K. Devnath, S. B. Niloy, C. M. Mehedi, T. H. Chowdhury, and M. I. K. Javed, "A stacking ensemble framework for predicting employee turnover: Explainable AI with SHAP," in *Proc. 2025 IEEE 2nd Int. Conf. Computing, Applications and Systems (COMPAS)*, Oct. 2025, pp. 1–6.
- [42] M. I. K. Javed, M. Imran, C. A. Gomes, P. S. Ponduru, S. Mandal, and S. Hassan, "Deep learning and explainable AI framework for predicting lung cancer severity," *Journal of Computer Science and Technology Studies*, vol. 7, no. 12, pp. 573–598, 2025.
- [43] C. A. Gomes, M. I. K. Javed, T. Ahammad, S. Mandal, S. Hassan, M. K. Sejan, et al., "Operational and financial evidence for generative AI in United States healthcare administrative workflows," *Frontiers in Computer Science and Artificial Intelligence*, vol. 4, no. 4, pp. 138–145, 2025.
- [44] M. I. K. Javed, M. Imran, A. A. Khan, M. Mehedi, A. Islam, and R. Pervez, "Explainable machine learning framework for early heart disease detection using SMOTE and SHAP," *Vascular and Endovascular Review*, vol. 9, no. 1, pp. 316–324, 2026.
- [45] M. I. K. Javed, M. R. M. Sirazy, S. Mandal, S. A. Akter, A. Hassan, and H. Esa, "Developing AI-based financial forecasting and cybersecurity systems for the US digital economy," *Frontiers in Computer Science and Artificial Intelligence*, vol. 5, no. 5, pp. 30–38, 2026.
- [46] M. A. Rahman, M. I. K. Javed, R. K. Devnath, C. M. Mehedi, M. Begum, and T. Mahmud, "MSRFF: An interpretable CNN-Vision Transformer framework for diabetic retinopathy detection," in *Proc. 2026 2nd Int. Conf. Computational Intelligence Approaches and Applications (ICCIA)*, May 2026, pp. 1–7.
- [47] M. M. R. Akash, "Cardiotoxicity prediction models in cancer patients using artificial intelligence and genomics," *International Journal of Drug Delivery Technology*, vol. 16, pp. 60–73.
- [48] M. I. K. Javed, M. A. Manzoor, F. M. Tofa, and M. H. Khan, "Interpretable ensemble learning approach for breast cancer diagnosis using SHAP-based explainable AI," *Journal of Computer Science and Technology Studies*, vol. 8, no. 8, pp. 244–255, 2026.
- [49] M. I. K. Javed, M. P. Ahmed, F. M. Tofa, M. F. Islam, C. A. Gomes, and R. M. Sirazy, "Federated intrusion detection for Internet of Medical Things networks: Differential privacy, non-IID robustness, and cross-device generalization," *Journal of Computer Science and Technology Studies*, vol. 8, no. 8, pp. 303–315, 2026.
- [50] P. S. Ponduru, P. P. V. Nandanavanam, and S. K. K. Ponduru, "Ecological proportionality in generative AI: The ethics of marginal capability and environmental sufficiency," 2026.
- [51] P. S. Ponduru, "SAFE-HealCloud: Safety-aware, agentic self-healing for cloud infrastructure," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 12, no. 4, pp. 163–188, 2026.



- [52] M. Akram, W. Khan, M. D. Rasheed, M. Imran, M. W. Iqbal, A. Delshadi, and M. Sultana, "Cybersecurity risk assessment model for Internet of Medical Things (IoMT) devices in healthcare systems," *Spectrum of Engineering Sciences*, vol. 4, no. 3, pp. 312–326, 2026.
- [53] M. D. Rasheed, M. Akram, M. Imran, R. H. Ahmed, and A. Rauf, "Towards human-centric smart manufacturing: A digital twin enabled affective ergonomic framework for adaptive human robot collaboration," *International Journal of Advances in Signal and Image Sciences*, pp. 1544–1555, 2025.
- [54] M. D. Rasheed, W. Khan, M. Imran, N. Ahmad, Y. Khan, M. Akram, and M. Sultana, "Leveraging artificial intelligence for advance data networking and cybersecurity," *Spectrum of Engineering Sciences*, vol. 4, no. 3, pp. 286–298, 2026.
- [55] N. M. Ali, M. E. Ahmed, M. Fakhrul Islam, C. A. Gomes, and M. S. Islam, "Toward a circular US economy: Green and artificial intelligence innovation, renewable energy, and domestic material consumption," *Energy Sources, Part B: Economics, Planning, and Policy*, vol. 21, no. 1, Art. no. 2685043, 2026.

