KFUPM

Journal of Undergraduate Research International

Volume 1, Issue 1, September 2025
ISSN 1532-458X

Journal of Undergraduate Research International

Volume 1, Issue 1, September 2025

Contents

Research Articles

Introducing the KFUPM Journal of Undergraduate Research International

"Research is the foundation of problem-solving."

At KFUPM, we believe that meaningful change begins with rigorous investigation. Research is not only the path to discovery—it is the driving force behind innovation, progress, and impact. Even if academia is not your final destination, research remains one of the most powerful tools for understanding and transforming the world as it is the essence of problem solving.

This is why undergraduate research has become a cornerstone of the academic experience at KFUPM. We encourage every student—not just future academics and researchers—to engage deeply with research as a way of critical thinking, developing, and contributing to knowledge. In fact, we do not just encourage; we require. This is why undergraduate research has been identified as one of the "Seven Habits of KFUPM Students".

It is from this philosophy that the KFUPM Journal of Undergraduate Research International was born. This journal provides a platform for undergraduate researchers from around the world to share their work, communicate their ideas, and showcase the impact of student-led investigations across disciplines. It celebrates not only the results of research, but also the process—the questions, the failures, the breakthroughs, and the passion that defines scholarly inquiry.

Though it is rooted at KFUPM, the journal is global in its vision. It welcomes contributions from undergraduate students everywhere and seeks to elevate the role of research at the undergraduate level—where the seeds of great ideas so often take root.

We hope this journal will grow to become a hub of excellence, curiosity, and collaboration, and that it will inspire young researchers to pursue bold questions and meaningful projects with confidence and creativity. This is just the beginning. Together, let us spark a movement that empowers undergraduates to lead with ideas, shape their disciplines, and contribute to humanity—one paper at a time.

Dr. Muhammad Al-Saggaf
President, KFUPM

Message from the Editor-in-Chief

It is with great pride and excitement that I present to you the inaugural issue of "Journal of Undergraduate Research International", our university's first undergraduate research journal. This publication marks a significant milestone—not only for our institution, but also for the many passionate student researchers who are shaping the future of academia.

This journal is to provide a platform where students can engage in the complete academic publishing experience for the very first time — from original research and thoughtful writing to peer review and final publication. We believe that such experience not only strengthens academic and professional skills but also inspires a lifelong appreciation for inquiry, intellectual rigor, and the sharing of knowledge.

I want to extend my deepest gratitude to Dr. Nayef M. Alsaifi, for establishing the journal, the editorial board, our faculty advisors, and the many reviewers who generously gave their time and expertise. Most importantly, I thank the student authors whose work fills these pages. You are the heart of this journal.
As we begin this journey, we welcome your feedback and your insights.

Dr. Jiabao Yi
Editor-in-Chief
Distinguished Professor, Chemical Engineering Department

Meet the International Editorial Board

Meet the International Editorial Board

Research Articles

KFUPM
Journal of Undergraduate Research International

Diamond Open Access 1

Article Accounting and Finance

Effects of Urban Comfort on a Heterogeneous Labor Force

Lifa Wang[1] and Xilin Zhang[2]*

Cite https://doi.org/10.64589/juri/209729

Submitted: May 07, 2025 **Revised:** July 24, 2025 **Accepted:** August 20, 2025

ABSTRACT

In China, the labor market is undergoing profound structural shifts in both supply and demand. As the economy transitions from high-speed to high-quality development, new competencies are required of the workforce. In this study, we used a microlevel perspective to examine the drivers of labor mobility. By integrating principal component analysis and the entropy weight method, we constructed a composite index for urban livability, theorized its role in shaping heterogeneous labor migration, and empirically tested these relationships using a conditional logit model. The analysis is based on 1% of the national population sample surveys conducted from 2000 to 2015, along with corresponding urban characteristics. The findings reveal that urban livability significantly influences destination choice, and there are notable variations between worker groups. Policy recommendations include enhancing urban livability to attract talent, tailoring recruitment strategies to local labor demands, and reforming the *hukou* (residency registration) system to reduce access barriers. This study offers theoretical and practical insights into how urban environments affect the mobility decisions of diverse labor populations.

Keywords: heterogeneous labor mobility, urban livability, conditional logit model

1. INTRODUCTION

As China transitions from high-speed growth to high-quality development, cities are no longer competing solely for capital and resources but also for human talent. Traditional economic incentives, such as wage differentials and job availability, remain important, but they are no longer the sole determinants of labor mobility. Urban livability, encompassing the natural environment, public services, cultural amenities, and overall comfort, has emerged as a critical factor shaping residential and employment decisions. However, migration patterns reveal a paradox: although large cities offer higher economic returns, many skilled workers are opting for smaller, more livable cities or returning to their hometowns to achieve a better balance between career development and life satisfaction. This raises key questions: To what extent does urban livability influence labor mobility, particularly among heterogeneous groups with diverse preferences and backgrounds, and how do these preferences evolve amid structural changes in the economy and society?

This study addresses these questions by constructing a comprehensive urban livability index and empirically testing its impact on labor mobility decisions using microlevel survey data and advanced econometric methods. By focusing on heterogeneous labor groups, the analysis reveals nuanced migration patterns not fully explained by economic variables alone.

1.1. Research Background.

Urbanization and industrialization have significantly expanded labor mobility across all sectors and regions. According to China's *Seventh National Census*, the floating population reached 375.82 million in 2020, approximately 26% of the total population, representing a 69.73% increase from 2010. Since 2000, the number of mobile workers has increased by over 210%, meaning that approximately one in five Chinese citizens is a migrant. In particular, cross-provincial migration continues to increase, especially concentrated in economically advanced urban clusters such as Beijing, the Yangtze River Delta, and the Pearl River Delta. However, migration to coastal megacities has slowed, giving way to trends such as labor "return migration" and the phenomenon of "escaping Beijing, Shanghai, and Guangzhou."

Further, China's economic shift toward quality development has been accompanied by evolving social values. As incomes and living standards rise, workers increasingly prioritize quality of life, and labor demand is diversifying. This transition challenges the long-standing assumption that economic factors alone determine mobility. From the mass rural migration of the 1990s to present-day "labor shortages" and reverse flows of highly educated talent from megacities, patterns have changed. In this study, we investigated whether urban livability has become a decisive factor in workers' destination choices and whether its influence differs between labor groups.

1.2. Contributions of This Study.

(1) Theoretical Contribution: This study extends migration theory by integrating a multidimensional urban livability index into the analysis of labor mobility, with explicit attention to labor heterogeneity. It bridges the gap between economic and amenity-based migration models.

(2) **Practical Contributions**: The findings offer actionable guidance for local governments seeking to attract or retain talent. By quantifying the impact of specific livability components, cities can design targeted interventions to enhance their attractiveness.

(3) **Economic Contribution:** Understanding the drivers of labor mobility in the current era has direct implications for urban planning, regional development, and the efficient allocation of human capital. Enhancing urban livability holds substantial potential to attract high-quality labor, thereby supporting sustainable economic growth.

1.3. Literature Review. Definitions of heterogeneous labor and urban livability vary between developmental stages. Drawing on the existing literature, this section clarifies both concepts and situates them within relevant theoretical contexts. Further, we review classical and contemporary theories of labor mobility and urban livability and analyze prior research across three key dimensions: (1) the heterogeneity of labor at the individual level, (2) role of economic factors in shaping mobility, and (3) influence of noneconomic variables, particularly urban livability, on migration decisions.

Existing studies on labor mobility identify both economic and noneconomic determinants. Economic factors include wages, housing prices, and unemployment rates. Noneconomic dimensions encompass public services, environmental quality, and urban livability. Research perspectives range from macro-level analyses focusing on household registration systems and national development strategies to microlevel approaches that examine individual preferences in response to city-specific attributes when choosing migration destinations.

1.3.1. Heterogeneity in Labor Mobility. Much of the existing literature treats labor as a homogeneous entity, overlooking how different worker types respond differently to the same influencing factors. Although such homogeneity may have applied during the early stages of industrialization, when labor oversupply allowed for standardized capital–labor combinations, subsequent economic development has produced labor shortages and increasingly differentiated demand, making heterogeneity more important. For example, Scully (1969) was among the first to recognize labor heterogeneity, observing that reverse migration in early 20th-century America stemmed from divergent regional income levels and was largely driven by differences among labor groups. Xiaofang (2013) analyzed the spatial mismatch between population and production in China, finding that cities such as Shanghai exhibited industrial overconcentration, whereas regions such as Inner Mongolia and Yunnan faced underutilization. She attributed these patterns to labor heterogeneity. In addition, some researchers have linked labor heterogeneity to regional income disparities. Wang and Fan (2004) argued that large-scale migration from central and western China to the eastern regions helped reduce income gaps by boosting productivity and wages in receiving areas. By contrast, Zhao and Li (2007) contended that skilled labor agglomeration intensified inequality, as higher-skilled workers benefited disproportionately relative to low-skilled counterparts. However, early research often treated labor as a homogeneous group, neglecting variation across skill level, age, and background (Scully, 1969). More recent studies, however, emphasize labor heterogeneity amid accelerated economic transformation (Sun, 2013). For example, Wang and Fan (2004) demonstrated that regional disparities in China are significantly shaped by the selective mobility of different labor groups. This is consistent with global findings showing that skill-based migration patterns have reshaped regional economies in the United States and Europe (Diamond, 2016; Storper et al., 2015).

1.3.2. Economic Drivers of Heterogeneous Labor Mobility. Because of individual differences in background and preferences, workers respond unevenly to economic incentives. Takatoshi and Thisse (2002) were among the first to incorporate labor heterogeneity into the New Economic Geography framework, arguing that disparities in household income, education, and personal values shape how workers perceive wage differentials and influence their locational choices. Moreover, housing costs have received significant attention as a proxy for living expenses. Housing prices play a dual role in migration decisions: labor inflows increase demand and push prices upward, whereas rising costs can deter further migration by raising the cost of living. However, whether housing acts as a "pull" or "push" factor depends on labor group heterogeneity. For example, Zhang, Zhang, and Yao (2019) found that increasing housing costs generally reduce migration intent. However, the deterrent effect is more pronounced among less-educated, younger, and female workers, whereas highly educated individuals tend to be less sensitive to these costs. This finding underscores the importance of recognizing labor heterogeneity in analyzing economic determinants of migration. Although economic factors such as wage differentials and housing prices remain central to migration decisions (Tabuchi & Thisse, 2002; Zhang et al., 2019), a growing body of literature highlights the importance of noneconomic influences, particularly urban amenities and livability (Tiebout, 1956; Glaeser et al., 2001; Florida, 2020). Recent research indicates that as economies mature, nonmonetary determinants, such as environmental quality, public services, and lifestyle preferences, become increasingly influential (Gennaioli et al., 2014). The "consumer city" paradigm suggests that urban growth increasingly hinges on a city's ability to offer a high quality of life (Glaeser et al., 2001). In China, several studies have shown that amenities such as educational resources, healthcare, and ecological quality significantly affect migration intentions (Zhang & Fang, 2019; Wang et al., 2021). Importantly, heterogeneity exists in amenity valuation: younger and more highly educated migrants tend to prioritize livability, whereas older or less-skilled groups are less sensitive to these factors (Liu & Shen, 2020).

1.3.3. Urban Livability and Other NonEconomic Influences on Heterogeneous Labor Mobility. Noneconomic factors, such as public services and urban livability, have long been recognized in labor mobility research. Tiebout (1956) introduced the "voting with their feet" theory, which suggests that individuals select residential locations based on preferred combinations of public goods and local tax structures. Around the same period, the geographer Ullman emphasized that migration is not solely economically driven, climatic comfort and quality of life also serve as powerful attractors. Similarly, Graves (1979) argued that regional utility differences arise not only from income but

also from nonmonetary attributes such as livability. He demonstrated that, in the absence of full income equalization, spatial variation in amenities may prompt migration. In the Chinese context, Zhang (2019) explained that when labor supply and demand experience external shocks and income adjustments fail to compensate for comfort disparities fully, livability must be directly incorporated into migration models. Glaeser et al. (2001) further advanced the "consumer city" concept, where urban growth increasingly depends on a city's capacity to deliver services and lifestyle amenities. Accordingly, livability has emerged as a central factor, beyond wages, in attracting labor. Diamond (2016) analyzed census microdata from the United States between 1980 and 2000, finding that migrants tend to prefer regions with lower housing costs and higher livability. However, preferences diverge by skill level: high-skilled workers prioritize quality of life, whereas low-skilled workers respond more strongly to wage–rent trade-offs.

1.4. Gaps in the Existing Literature. A review of the literature reveals several limitations. Most current studies investigate noneconomic factors influencing labor mobility using individual indicators such as public service quality or environmental conditions. However, relatively few incorporate urban livability as a composite metric that captures broader urban characteristics. In terms of data, many studies rely on aggregated macro-level statistics, which obscure individual-level heterogeneity. Although the China Labor Force Dynamic Survey (CLDS) is a commonly used microlevel dataset, its nationwide coverage began only in 2012, limiting most empirical analyses to 2014 or 2016. Consequently, studies often rely on data from 2005 or 2016, resulting in limited temporal coverage. Furthermore, while some scholars examine drivers of labor mobility at the macro level, few explore these mechanisms from a microanalytical perspective. This study addresses that gap by focusing on individual-level determinants of labor movement.

1.5. Theoretical Background. This study draws on two foundational frameworks: *Random utility theory* and the concept of *amenity migration*. According to random utility theory, individuals make locational choices by maximizing their perceived utility, which depends on both observable city characteristics and individual-specific preferences (McFadden, 1974). Within this framework, urban livability enters the utility function as a critical explanatory variable, interacting with personal traits such as age, education, and *hukou* (residency) status. Amenity migration theory (Graves, 1979; Glaeser et al., 2001) further explained why noneconomic factors, such as climate, culture, and public services, can outweigh purely monetary considerations, particularly for high-skilled or mobile populations. As the urban economy matures, cities increasingly compete based on "soft" factors, and these amenities become central to migration decisions. By synthesizing these theories, this study hypothesizes that (1) urban livability is an independent and significant determinant of labor inflow, and (2) the strength of this effect varies according to labor heterogeneity.

2. METHODOLOGY

We employed microlevel data from China's 1% population sample surveys for 2000, 2005, 2010, and 2015. A conditional logit regression model was used to estimate the impact of urban livability on the probability of labor inflow. The model tested the core hypotheses and incorporated robustness checks to ensure the reliability of the results.

2.1. Model Specification and Data Description.

2.1.1. Model Specification. We adopted a conditional logit model to examine labor migration decisions across multiple potential destination cities. Each worker was assumed to face a set of urban alternatives and to select the one that maximized their expected utility. The random utility associated with choosing city j is given by Eq. (1).

$$U_{ij} = \alpha X_{ij} + \varepsilon_{ij} \ (i = 1, 2, 3...j = 1, 2, 3...) \tag{1}$$

Here, i denotes an individual worker, and j represents a potential destination city. The utility that worker i derived from choosing city j is denoted by U_{ij}, where X is a vector of characteristics specific to city j. If $U_{ij} > U_{ik}$ for any alternative city k, then the worker chose city j. The probability that worker i selected destination j is given by Eq. (2).

$$\text{Prob}(y_i = j \mid x) = \frac{\exp(\alpha X_{ij})}{\sum_{j=1}^{J} \exp(\alpha X_{ij})} \tag{2}$$

When individual i evaluates city j, they face a binary choice: migration to that city (i.e., labor inflow) or not. This decision is captured using a binary indicator. If migration occurred, the observation is coded as 1; otherwise, it is coded as 0, as shown in Eq. (3).

$$\text{Prob}(\text{choice}_{ij} = 1) = \frac{\exp(\alpha X_{ij})}{\sum_{j=1}^{J} \exp(\alpha X_{ij})} \tag{3}$$

Because we analyzed the impact of urban livability on heterogeneous labor mobility, recognizing that individuals respond differently to city characteristics based on their own attributes, the empirical model incorporated interaction terms between urban livability and worker-specific traits. If Z_i denotes the characteristics of worker i, Y_{ij}represents the level of urban livability in city j, and X_{ij} include other city-level attributes. The utility function is given by Eq. (4).

$$\text{Prob}(\text{choice}_{ij} = 1) = \frac{\exp(\alpha X_{ij} + \beta Y_{ij} + sZ_i * Y_{ij})}{\sum \exp(\alpha X_{ij} + \beta Y_{ij} + sZ_i * Y_{ij})} \tag{4}$$

2.1.2. Variable Description and Data Sources.

(1) Dependent Variable: Individual Labor Data

The primary individual-level labor data used in this study were drawn from the 1% *National Population Sample Survey Microdatabase* for 2000, 2005, 2010, and 2015 (Table 1). Migrant workers were identified as individuals whose current place of residence was in a different prefecture-level city from their registered *hukou* location. Those residing in a non-*hukou* prefecture were classified as inflow migrants.

The sample was restricted to individuals aged 16–65 and excluded those whose current activity status was listed as "in school." In the dataset, gender was coded as 1 for males and 2 for females, whereas *hukou* type was coded as 1 for agricultural and 2 for nonagricultural.

Table 1. *Indicators for the urban livability index*

Primary Indicator	Secondary Indicator	Tertiary Indicator
High quality of resources	Medical treatment	Number of hospital beds per 10,000 people Number of doctors per 10,000 people
	Education	Pupil–teacher ratio in primary and secondary schools Student–teacher ratio in general secondary schools Number of college students per 10,000 people
Living convenience	Traffic	Number of buses per 10,000 people Number of taxis per 10,000 people
	Living consumption	Number of hotel and catering industry employees Residential service and repair employment Number of post offices (county level)
Environmental comfort	Natural environment	Annual sunlight hours Average January temperature
	Artificial healthy environment	Green coverage rate of built-up areas (%) Centralized treatment rate of sewage (%) Harmless disposal rate of household garbage (%)
Cultural richness	Library holdings	Library holdings per 100 people
	Theaters	Number of cinemas

(2) Independent Variables: City-Level Characteristics

City-level variables referred to data collected at the prefecture level or above. These characteristics were sourced primarily from the China *City Statistical Yearbook*, *Regional Economic Statistical Yearbook*, *Statistical Yearbook*, and *Environmental Statistical Yearbook* for the corresponding years.

In addition to urban livability, the core explanatory variable, key economic factors, were incorporated, including average wages and housing prices. Wage data were drawn from the *China City Statistical Yearbook* and reflect the annual average salaries of urban employees. It was assumed that workers tend to migrate to cities offering higher wages. Housing prices serve both as a proxy for the cost of living and as an outcome partially shaped by labor inflows. High housing costs may deter migrants; however, labor inflows, improved employment opportunities, and increased livability may also elevate housing prices. Because each individual i faces a choice set of j cities, the conditional logit model estimates a total of $i \times j$ observations.

3. RESULTS AND DISCUSSION

3.1. Basic Regression Results. Table 2 reports the regression results for labor mobility across four survey years, 2000, 2005, 2010, and 2015, based on standardized city-level indicators, including urban livability, average wages, and housing prices. Regression model (1) includes only the three core explanatory variables: average wage, housing price, and urban livability. Regression model (2) introduces additional city-level covariates, including industrial structure, GDP per capita, population density, foreign investment, annual population, provincial capital status, and fixed asset investment. Regression model (2) also incorporates a set of control variables: industrial structure, population density, foreign direct investment (FDI), GDP per capita, fixed asset investment, and provincial capital status. As the analysis employed a conditional logit model and all city characteristics were transformed into positively oriented indicators, the estimated coefficients represent the percentage increase in the probability of city inflow associated with a 1% increase in the respective variable.

The results show that average wage consistently exerted a positive effect on labor inflow across all years, indicating that workers were sensitive to monetary returns. In 2000 and 2005, housing prices positively affected migration decisions, whereas in 2010 and 2015, after controlling for additional variables, the relationship turned negative. This implies an inverted U-shaped effect of housing prices on labor mobility, initially attractive, but eventually inhibitory as costs outweigh benefits.

3.1.1. Individual Heterogeneity of the Impact of Urban Comfort on Labor Mobility. In Table 2, the regression model treated labor as a homogeneous group and reported the overall effect of urban livability on labor mobility. Table 3 extends this analysis by incorporating individual heterogeneities. Specifically, it examines how the impact of urban livability varies by age, gender, educational attainment, and urban–rural *hukou* status. Building on the baseline conditional logit model, interaction terms between individual characteristics and urban livability were introduced. The coefficients of these interaction terms allow for an assessment of how different types of workers responded to variations in urban livability, revealing distinct patterns of migration preferences.

Summary of Regression Findings: (1) *Wages consistently positively influenced labor mobility.* For all four sample years, average urban wages exhibited a stable and significant positive effect on the probability of labor inflow. Higher wages enhanced the attractiveness of a city to migrant workers. (2) *Urban livability is an increasingly important driver.* In 2000 and 2010, higher urban livability was negatively associated with labor inflow, suggesting that during periods of economic uncertainty or slower growth, workers prioritized cities with better income prospects. However, in 2005 and 2015, livability became a strong positive predictor. Notably, in 2005, the influence of livability was less than that

Table 2. *Basic regression results: Impact of urban livability on labor mobility (2000–2015)*

	2000		2005		2010		2015	
Explanatory variable	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Average wage of employees	0.270***	0. 141***	0.375***	0.528***	0.497***	0. 139***	0.315***	0. 152***
	(0.00672)	(0.00842)	(0.00472)	(0.00622)	(0.00373)	(0.00486)	(0.00267)	(0.00499)
Housing price	0.388***	0.218***	0.599***	0.501***	−0.311***	−0.672***	0. 146***	−0. 105***
	(0.00740)	(0.00982)	(0.00458)	(0.00624)	(0.00462)	(0.00801)	(0.00183)	(0.00340)
Urban comfort	−0.455***	−0.687***	0.209***	0.475***	−0.471***	−0.429***	0.509***	0.626***
	(0.0135)	(0.0155)	(0.00675)	(0.0105)	(0.00390)	(0.00458)	(0.00441)	(0.00740)
Industrial structure		−0.0325***		0.0413***		0.0391***		0.0792***
		(0.00584)		(0.00386)		(0.00582)		(0.00336)

	2000	2005	2010	2015
Explanatory variable	(1) (2)	(1) (2)	(1) (2)	(1) (2)
GDP per capital	0. 172***	−0.0780***	0.402***	0.374***
	(0.00404)	(0.00639)	(0.00440)	(0.00394)
Population density	0. 136***	−0.0578***	0.207***	0.234***
	(0.00450)	(0.00332)	(0.00428)	(0.00303)
Foreign investment is ten thousand dollars	0.0568***	0.758***	0.952***	0.0788***
	(0.004500)	(0.00558)	(0.00824)	(0.00278)
Average annual population	0. 113***	0.537***	0. 111***	−0.0852***
	(0.00414)	(0.00551)	(0.00664)	(0.00383)
Whether it is a provincial capital city	0.620***	0.857***	−0.833***	−0.0351***
	(0.0163)	(0.0102)	(0.0222)	(0.0116)
Investment in the fixed assets	−0.021***	−1. 127***	−0.903***	−0.0807***
	(0.0036)	(0.00858)	(0.0106)	(0.00393)
Number of cities	232 232	234 234 93	93	282 282
Number of workers observed	52345 52345	111012 111012 152241	152241	108341 108341
Total observations = Number of cities *Number of workers observed	12144040 12144040	25976808 25976808 14158413	14158413	30552162 30552162

Standard errors in parentheses $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

of wages; by 2015, it had surpassed wage levels in determining migration. (3) *The effect of urban livability on labor mobility is heterogeneous.* The influence varied significantly across demographic groups. Age moderated the positive effect of livability; that is, younger workers were more responsive to urban amenities. Educational attainment enhanced the effect of livability , whereas *hukou* status and sex also introduced measurable differences in sensitivity to urban conditions.

4. CONCLUSIONS AND POLICY SUGGESTIONS

4.1. Research Conclusions. Urban livability has become an increasingly important factor influencing labor mobility, and its impact has intensified over time. However, during periods of slower economic development or economic volatility, improvements in livability alone did not significantly increase the probability of labor inflow. Further, the effect of urban livability on labor mobility did not show individual heterogeneity during those periods. By contrast, younger and more highly educated workers were more inclined to migrate to cities with higher levels of livability. The influence of urban livability on labor mobility corresponded most strongly to the quality of urban resources and the convenience of daily life, suggesting that these components are key drivers in shaping migration preferences. Workers of different ages, educational levels, and *hukou* types responded differently to urban livability. Gender appeared to have little effect on this sensitivity. Younger and more educated workers were more likely to select cities with higher livability, and individuals holding nonagricultural *hukou* were more inclined to migrate to such cities than those with agricultural *hukou*.

https://doi.org/10.64589/juri/209729

Table 3. *Individual heterogeneity of urban comfort affecting labor mobility*

(1) choice VARIABLES		(2) choice VARIABLES		(3) choice VARIABLES		(4) choice VARIABLES	
Salary	0. 152*** (0.00499)	Salary	0. 152*** (0.00499)	Salary	0. 152*** (0.00498)	Salary	0. 152*** (0.00499)
Housing price	−0. 105*** (0.00340)	Housing price	−0. 105*** (0.00340)	Housing price	−0. 106*** (0.00340)	Housing price	−0. 105*** (0.00340)
Urban comfort	0.718*** (0.0109)	Urban comfort	0.638*** (0.0103)	Urban comfort	0.432*** (0.00953)	Urban comfort	0.492*** (0.0104)
Age * City comfort		Gender ∗ Urban Comfort		Educational level ∗ Urban comfort		Household nature ∗ City comfort	
	−0.00253*** (0.000222)		−0.00809 (0.00495)		0.0532*** (0.00164)		0.0911*** (0.00494)
Industrial structure	0.0792*** (0.00336)	Industrial structure	0.0792*** (0.00336)	Industrial structure	0.0790*** (0.00336)	Industrial structure	0.0791*** (0.00336)
GDP per capital		GDP per capital		GDP per capital		GDP per capital	
	0.374*** (0.00394)		0.374*** (0.00394)		0.374*** (0.00394)		0.374*** (0.00394)
Density of population	0.234*** (0.00303)	Density of population	0.234*** (0.00303)	Density of population	0.235*** (0.00303)	Density of population	0.235*** (0.00303)
Actually utilized foreign capital		Actually utilized foreign capital		Actually utilized foreign capital		Actually utilized foreign capital	
	0.0789*** (0.00278)		0.0788*** (0.00278)		0.0792*** (0.00278)		0.0790*** (0.00278)
Year-end population	−0.0855*** (0.00383)	Year-end population	−0.0852*** (0.00383)	Year-end population	−0.0880*** (0.00383)	Year-end population	−0.0861*** (0.00383)
Whether it is a provincial capital city		Whether it is a provincial capital city		Whether it is a provincial capital city		Whether it is a provincial capital city	
	−0.0347*** (0.0116)		−0.0351*** (0.0116)		−0.0318*** (0.0116)		−0.0341*** (0.0116)
Investment in the fixed assets		Investment in the fixed assets		Investment in the fixed assets		Investment in the fixed assets	
	−0.0807*** (0.00393)		−0.0807*** (0.00393)		−0.0804*** (0.00393)		−0.0806*** (0.00393)
Total observations	29,225,916		29,225,916		29,225,916		29,225,916

Standard errors in parentheses $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

4.2. Policy Suggestions. (1) *Enhance urban livability to increase the attractiveness of cities to migrant labor*. As China enters a stage of high-quality economic development, urban livability has become a core determinant of labor migration decisions. This study confirms that livability significantly influences worker mobility. In an era of rising living standards and growing aspirations for improved quality of life, enhancing urban livability is essential to meet labor demands and attract a mobile workforce. (2) *Adopt demand-oriented talent strategies tailored to local conditions.* The regression results indicate significant individual heterogeneity in migration preferences. Younger and more educated workers tend to favor cities with higher livability, and individuals with nonagricultural *hukou* are more likely to migrate to such areas. Talent attraction policies should therefore be demand-driven and locally adapted to align with the preferences of diverse labor groups.

4.3. Directions for Future Research. Although this study provides new insights, several directions remain for future investigation. First, as data availability improves, longer time-series and panel datasets could allow for more precise identification of dynamic effects and causal mechanisms. Second, comparative studies across different countries or within diverse regions of China would help generalize the findings and test their external validity. Third, qualitative research could deepen the understanding of individual migration motivations and identify the specific urban amenities most valued by different labor groups. Finally, the potential mediating roles of digital infrastructure and emerging forms of flexible work in shaping labor mobility merit further exploration.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Lifa Wang – *Department of Economics and Management, Shandong Huayu University of Technology, China;* 0009-0003-0534-1699
Email: 17862931961@163.com

Corresponding Author

Xilin Zhang – *Research Mentor, Professor, Department of Economics and Management, Shandong Huayu University of Technology, China;* 0009-0008-5861-8854
Email: 18905442878@163.com

REFERENCES

(1) Clark, D. (2004). Urban world/global city. Routledge.

(2) Scully, G. W. (1969). Interstate wage differentials: A cross section analysis. The American Economic Review, 59(5), 757–773.

(3) Zhao, W., & Li, F. (2007). The Mobility of Heterogeneous Labor and Regional Income Disparities: An Extended Analysis Based on the New Economic Geography Model. Chinese Journal of Population Science, (01), 27–35, 95.

(4) Sun, X. (2013). Heterogeneous Labor and Labor Mobility in China: An Analysis Based on New Economic Geography. Chinese Journal of Population Science, (03), 36–45.

(5) Wang, X. L., & Fan, G. (2004). Trends and Determinants of Regional Disparities in China. Economic Research Journal, (01), 33–44. Tabuchi T, Thisse JF. Taste heterogeneity, labor mobility and economic geography [J]. Journal of Development Economics, 2002, 69(1): 155–177.

(6) Zhang, H. F., Zhang, J. Z., & Yao, X. G. (2019). Spatial Evolution of Housing Costs and Their Impact on Labor Mobility Decisions in China. Economic Geography, 39(07), 31–38.

(7) Tiebout CM. A Pure Theory of Local Expenditures [J]. Journal of Political Economy, 1956, 64(5):416–424.

(8) Graves, P. E., & Linneman, P. D. (1979). Household migration: Theoretical and empirical results. Journal of urban economics, 6(3), 383–404.

(9) Zhang, Y. L., & Fang, Q. Y. (2019). The Impact of Urban Livability on Labor Mobility. China Population, Resources and Environment, (3), 118–125.

(10) Glaeser, E. L., Kolko, J., & Saiz, A. (2001). Consumer city. Journal of economic geography, 1(1), 27–50.

(11) Diamond, R. (2016). The determinants and welfare implications of US workers' diverging location choices by skill: 1980–2000. American economic review, 106(3), 479–5.

(12) Diamond, R. (2016). The determinants and welfare implications of US workers' diverging location choices by skill: 1980–2000. American economic review, 106(2), 479–524.

(13) Florida, R. (2020). The New Urban Crisis. Basic Books.

(14) Glaeser, E. L., Kolko, J., & Saiz, A. (2001). Consumer city. Journal of economic geography, 1(1), 27–50.

(15) Gennaioli, N., La Porta, R., Lopez De Silanes, F., & Shleifer, A. (2014). Growth in regions. Journal of Economic growth, 19(2), 259–30.

KFUPM
Journal of Undergraduate Research International

Diamond Open Access 8

Article Accounting and Finance/Biological Sciences

Effect of Public Health Expenditure on the Under-Five Mortality Rate in Nigeria

Aliyu Inuwa Kamara[1] and Abubakar Orlando Ijoko[2]*

Cite https://doi.org/10.64589/juri/209727

Submitted: May 19, 2025 **Revised:** July 18, 2025 **Accepted:** August 20, 2025

ABSTRACT

In this study, we examined how public health spending affected Nigeria's under-five mortality rate between 1988 and 2023 using time series data obtained from the Central Bank of Nigeria statistical Bulletin and World Bank World Development Indicators (WDI). Four explanatory variables were used: female literacy rate, technology, HIV prevalence, and public health spending. an Auto-regressive Distributed Lagged (ARDL) model was employed for the analysis after conducting unit root and cointegration tests, which showed a mixed order of integration between the variables. From the short-term results, public health expenditure has a positive and significant impact on the under-five mortality rate, whereas the HIV prevalence rate and technology integration have a negative and significant impact. In contrast, the long-term ARDL results indicate that, public health expenditure has a negative and statistically significant impact on the under-five mortality rate. Furthermore, the HIV prevalence rate and technology integration had a long-term positive and statistically significant relationship with the under-five mortality rate. Based on these findings, to reduce the under-five mortality rate in Nigeria, we recommend that the government through the Ministry of Finance and the Ministry of Health should improve budgetary allocation to the health sector at levels with adequate supervision to prevent wastage and funding misappropriation. By following these measures, Nigeria will move toward achieving United Nations Sustainable Development Goal.3;"Ensure healthy lives and promote well-being for all at all ages."

Keywords: public expenditure, health, under-five mortality rate

1. INTRODUCTION

Health is a critical component of human development that significantly influences quality of life, and productivity, as well as the economic growth of a nation. In Nigeria, health outcomes have been a persistent concern, particularly the under-five mortality rate, which reflects the well-being of children under five years of age. The under-five mortality rate is a vital indicator of the effectiveness of healthcare systems, public health policies, and the socioeconomic conditions of a population. Despite several health interventions and policies, Nigeria has consistently recorded high under-five mortality rates, ranking among the highest in the world[1]. Globally, the under-five mortality rate defined as the probability of dying between birth and five years of age per 1000 live births has decreased by 60%, from an estimated rate of 93 deaths per 1000 live births, in 1990 to 38 deaths per 1000 live births in 2019, The United Nations Sustainable Development Goal (SDG) target is 25 deaths per 1000 live births by 2030. In 2019, an estimated 5.2 million children under-five died globally, approximately 14,000 child deaths each day before their fifth birthday[2].

Public health expenditure plays a crucial role in improving access to healthcare, infrastructure, and service delivery. Government investments in healthcare are expected to address disparities, enhance medical services, and reduce mortality rates, including those of children under five. However, in Nigeria, health financing is characterized by inefficiencies, insufficient allocation, and a high dependence on out-of-pocket expenditure, which limit the impact of public spending on health outcomes. These inefficiencies have contributed to persistently high under-five mortality rates; in 2020 the estimate was 113 deaths per 1000 live births[3]. Given these statistics, Nigeria must reduce early childhood mortality by 70% to meet the SDGs by 2030, urgent attention is required to reduce childhood deaths to below the global average[4]. Public health expenditure in Nigeria is far below the World Health Organization (WHO) recommendation, that is, 15% of the national budget. Furthermore, per capita health allocations in Nigerian states fell from $10.8 in 2020 to $8.5 in 2022, significantly lower than the WHO target of $86 per capita[5]. The gap in public health expenditure has resulted in an Inadequate healthcare infrastructure, insufficiently skilled healthcare professionals, limited access to essential medicines and vaccines, and poor health outcomes, particularly for children under five[6].

However, studies suggest that increased health expenditure does not always translate into better health outcomes because of issues such as corruption, inefficient resource allocation, and poor governance. Moreover, socioeconomic factors, such as

poverty, education, and access to clean water, interact with health expenditure to affect overall health outcomes.

This study investigates the impact of public health expenditure on under-five mortality rates in Nigeria from 1988 to 2023. By analyzing trends over this extended period, we aim to provide insights into how government spending on health has influenced child mortality and to identify potential policy gaps. These findings contribute to the ongoing debate about the effectiveness of public health investments in achieving sustainable health outcomes in Nigeria.

2. LITERATURE REVIEW

Empirical studies on the relationship between public health expenditure and under-five mortality have been conducted both in Nigeria and other countries to assess the impact of public health expenditure on under-five mortality rates.

Concerning Nigeria, Edeme et al. empirically investigated the effect of public health expenditure on health outcomes, as measured by life expectancy at birth and infant mortality rates[7]. This study employed the error correction mechanism (ECM) and Granger causality methods to estimate the models. The results showed that an increase in public health expenditure improved life expectancy and reduced infant mortality rates. Additionally, the urban population and the HIV prevalence rate were found to affect health outcomes significantly, whereas per capita income had no significant effect. These findings suggest that public health expenditure remains a critical component in improving health outcomes in Nigeria.

Later, Zubair studied the impact of governmental health expenditure on health status indicators (life expectancy rate, under-five mortality rate and infant mortality rate) empirically using data from 1985 to 2015[8]. The Engel Granger cointegration method was employed, and three models were developed, with each health outcome as a dependent variable. The study found that private health expenditure, gross domestic product (GDP) per capita, physicians per 1000 population, and population density significantly explained the variations in under-five and infant mortality in Nigeria and were statistically significant across the models. These findings imply that health expenditure contributes to improving health outcomes in Nigeria.

Moreover, with the aim of linking public health expenditure directly or indirectly to infant and under-five mortality in Nigeria, Yaqub et al. employed the ordinary least squares and two-stage least squares (2SLS) methods to study the impact of governmental health expenditure on infant and under-five mortality[9]. They found that public health spending has an inverse relationship with infant and under-five mortality when corruption is involved. Similarly, David used the autoregressive distributed lagged model (ARDL) bounds testing approach and Granger casualty test to examine the connection between governmental health expenditure and infant and under-five mortality empirically[10]. The results showed that public health spending, and immunization have a negative impact in both the short and long-term.

Aderopo and Emmanuel also examined the impact of public health expenditure on the infant mortality rate in Nigeria between 1991 and 2018 using time-series data[11]. The fully modified ordinary least square (FMOLS) method was used to examine these relationships. Further, various robustness checks were conducted to ensure the reliability of the results for policymakers. The findings revealed that all variables employed had a positive impact on infant and maternal mortality, except for diphtheria, pertussis, and tetanus immunization and female literacy rate.

In a recent study, Azuh et al. examined the relationship between public health expenditure and under-five mortality in Nigeria, and employed ARDL to assess the long-term effect of the variables[12]. Data were sourced from the World Bank World Development Indicators (WDIs) for 1985–2017. The results showed that, although public health expenditure was statistically significant, it was positively related to under-five mortality. This implies that a unit increase in public health expenditure leads to a 1.56% increase in the under-five mortality rate. However, in the Nigerian context, this result can be better explained by factors such as the poor coordination of health funds and maternal education. Therefore, Azuh et al. recommended the implementation of proper health fund coordination to ensure that the budget allocated to the health sector is utilized effectively.

Ojo et al. also used the ARDL technique to examine the impact of health expenditure on life expectancy in Nigeria from 1981 to 2018[13]. They found that health expenditure had an insignificant impact on life expectancy in Nigeria. Similarly, Iyakwariet al. employed the ARDL bound test to cointegration methods to investigate the effect of health expenditure on life expectancy in Nigeria empirically using time-series data from 1990 to 2021[14]. The results revealed a negative relationship between health capital expenditure, recurrent health expenditure, and life expectancy in the long-term, whereas out-of-pocket health expenditure had a positive relationship with life expectancy.

Using the under-five mortality per 1000 births and life expectancy as proxies for health outcomes, Orji et al. examined the impact of public health expenditure on health outcomes in Nigeria from 1985 to 2019[15]. The augmented Dickey-Fuller (ADF) technique was used to test the unit root of the variables. Their findings revealed that governmental health expenditure had a significant impact on the under-five mortality rate and life expectancy, as did immunization against measles for the form. In line with earlier findings, the recommendation was that the Federal Government of Nigeria increase its annual allocation to the health sector to improve population health.

Regarding cross country evidence, Kato et al. studied the effect of public health spending on under-five mortality rates in Uganda[16]. A time=series regression analysis was conducted using data obtained from World Bank Indicators covering 1980 to 2012. The findings revealed that recurrent and capital health expenditure were strongly related to the mortality rate of those under-five. Further, they found that increased public health expenditure, along with its effective allocation, would help reduce the mortality rate of people under-five in Uganda. In addition female enrollment in secondary school was significant, suggesting that increased access to education for girls reduces under-five mortality. Overall, they concluded that merely increasing public spending is not sufficient to improve child health outcomes in Uganda, but must be complemented by other factors such as female education.

Rana et. al. studied the association between health expenditure and health outcomes[17]. Heterogeneity and cross-sectional dependence were controlled for in the panel data, which comprised of 161 countries for the period 1995–2014. Infants under-five years of age, maternal mortality, and life expectancy at birth were selected as measures of health outcomes. Cross-sectional augmented Im, Pesaran, and Shin (IPS) unit root, panel ARDL and the Toda Yamamoto approach to the Granger causality test were used to investigate the relationships across four income groups. The results revealed that the link between health expenditure and health outcomes is stronger in low-income countries than in high-income countries. Furthermore, increasing health expenditure was found to reduce child mortality but was not significantly associated with maternal mortality at all income levels. Variations in child mortality were better explained by increased health expenditure than by maternal mortality. They concluded that the influence of health expenditure on health outcomes varies significantly across income levels except for maternal health.

However in contrast, Rahman et al. used fixed and random effects models to examine the relationship between private, public and total health care and three selected health outcomes (life expectancy, under five mortality and crude death rates) in the South Asian Association for Regional Cooperation and Association of Southeast Asian Nations (SAARC ASEAN) regions[18]. Both increased public and private health expenditure improved life expectancy and significantly decreased the mortality rates of the under-five in the region. Moreover, in Iran, Shahraki investigated the causality between public health expenditure and life expectancy in the short and long term from 2000 to 2017 in Iran using the Johansen cointegration approach, findings showed a bidirectional causal relationship between these factors in short and long-term[19].

Using the ARDL estimation technique, Renuka et al. investigated the relationship between public, private, and out-of-pocket health expenditures and the under-five mortality rate in Malaysia[20]. New critical test values were recalculated for a bound testing technique of cointegration using the response surface methodology expanded by Turner, which is based on a small time-series sample of 22 years, from 1997 to 2018. These findings suggest that an increase in out-of-pocket health expenditure increases the under-five mortality rate, indicative of a worsening health situation. In contrast, the effects of public and private health expenditure were found to be statistically insignificant. Thus, these results do not support out-of-pocket payments as an effective way of financing healthcare to improve child health, implying that the cost of out-of-pocket payments is a barrier to seeking medical care. Consequently, reducing out-of-pocket payments is necessary because they contribute to worsening child health outcomes. Therefore, a reliable and effective health financing system, along with the targeted health-related safety net, is essential to protect families.

Ayipe and Tanko examined the relationship between public health expenditure and under-five mortality in low-income sub-Saharan African countries (SSA) using data from the WDIs spanning the year 2000 to 2019[21]. The Breusch Pagan Lagrange test was conducted to investigate the evidence for panel effects across countries. The results specifically indicate that, for every 1% increase in domestic general government health expenditure, there is a reduction of approximately 5.3 under-five deaths per 1000 live births. This indicates a strong relationship between state health spending from domestic funds and under-five mortality rates in low-income SSA countries suggesting that increased public health expenditure in SSA countries is required.

Hosokawa et al. examined the relationship between healthcare spending and healthy life expectancy at birth using descriptive statistics and correlation analyses across all secondary medical areas in Japan. The results revealed significant regional disparities and that the number of medical personnel supporting clinics for home healthcare delivery facilities, home visit treatments, and expenditure per capita (dentistry) have a positive relationship with both life expectancy and healthy life expectancy. This finding is consistent with that of Li et al. who revealed that an increase in the healthy diet score based on food expenditure by 1% will result in a 0.07% increase in life expectancy among men alone, women alone, and men and women combined in the US[22].

2.1. Theoretical Framework. The Grossman production function forms the theoretical foundation of this study. Fayissa and Gutema extended the Grossman, model by incorporating social, economic, and environmental factors as inputs to the health production process[23]. Grossman posited that many health related behaviors, such as seeking medical care, are valued primarily for their contributions to health[24]. Within this framework, the demand for health is modeled using a health production function, in which health is "produced" from various inputs to yield an individual's health status. At the micro-level, the theoretical formulation is represented by:

$$H = F(\mathrm{x}) \tag{1}$$

Here. H denotes health output, and x is a vector of inputs that includes economic variables (e.g., income per capital), social variables (e.g., education), environmental variables (e.g., urbanization), demographic indicators (e.g., population below or above certain age group), health service indicators (e.g., population-doctor ratio, population-hospital ratio etc.), and other factors influencing health status.

3. METHODOLOGY

In this study, we investigated the impact of public health expenditure on the under-five mortality rate in Nigeria between 1988 and 2023 using time series data obtained from the Central Bank of Nigeria (CBN) statistical Bulletin and World Bank WDIs.

3.1. Model Specification. The model used in this study is adapted from the work of Mubarak[23], and is shown in functional and econometric form: in Eqs. (2) and (3), respectively.

$$\mathrm{UMR} = \mathrm{f}(\mathrm{PHE},\ \mathrm{HIVPR},\ \mathrm{TECH},\ \mathrm{FLR}) \tag{2}$$

$$\mathrm{UMR} = \beta_0 + \beta_1\mathrm{PHE} + \beta_2\mathrm{HIVPR} + \beta\mathrm{TECH} + \beta_4\mathrm{FLR} + \mu \tag{3}$$

In the model *UMR* represents the mortality rate of those under-five, *PHE* denotes public health expenditure, and *HIVPR* refers to the *HIV* prevalence rate, *TECH* captures the level of technology, and *FLR* stands for the female literacy rate. The symbol β

represents the coefficients of the independent variables, and μ is the error term.

3.2. Estimation Technique. We used the ARDL model, which is applicable in testing for cointegration when the variables are either purely I(0) or I(1) or have a mixture of stationarity. If variables are found to be nonstationary, the cointegration test, which is a pretest for spurious regression is first carried out. In addition, a pretest for stationarity is necessary in the application of the ARDL model because it does not allow for the inclusion of variables having I(2) and above.

The long-term model is given by Eq. (4).

$$\begin{aligned}\Delta UMRt = {} & \delta_o + \delta_1 UMR_{t-1} + \delta_2 PHE_{t-1} + \delta_3 HIVPR_{t-1} \\ & + \delta_4 TECH_{t-1} + \delta_5 FLR_{t-1} + \sum_{i=0}^{p} \rho_1 \Delta UMR_{t-1} \\ & + \sum_{i=0}^{p} \rho_1 \Delta PHE_{t-1} + \sum_{i=0}^{p} \rho_1 \Delta TECH_{t-1} \\ & + \sum_{i=0}^{p} \rho_1 \Delta FLR_{t-1} + \mu_t \end{aligned} \quad (4)$$

In contrast, the short-term model is given by Eq. (5).

$$\begin{aligned}\Delta UMRt = {} & \varphi_o + \varphi_1 UMR_{t-1} \pm \varphi_2 PHE_{t-1} \\ & \pm \varphi_3 HIVPR_{t-1} \pm \varphi_4 TECH_{t-1} \pm \varphi_5 FLR_{t-1} \\ & + \sum_{i=0}^{p} \Psi_1 \Delta UMR_{t-1} \pm \sum_{i=0}^{p} \Upsilon_1 \Delta \mathrm{PHE}_{t-1} \\ & \pm \sum_{i=0}^{p} \Upsilon_1 \Delta \mathrm{HIVPR}_{t-1} \pm \sum_{i=0}^{p} \Upsilon_1 \Delta \mathrm{TECH}_{t-1} \\ & \pm \sum_{i=0}^{p} \Upsilon_1 \Delta FLR_{t-1} + ECM_{t-1} + \mu_t \end{aligned} \quad (5)$$

The coefficient of ECM_{t-1}, is negative, statistically significant, and has a value less than one in absolute terms.

3.3. Data Sources. The data collection technique for the research is time series in nature, and in the secondary form sourced mainly from the WDIs from World Bank[5] and CBN statistical bulletin[26]. Specifically, data regarding public health expenditure were collected from the CBN Statistical Bulletin, whereas data on mortality rate, HIV prevalence, technology and female literacy were collected from the WDIs.

3.4. Variable Measurement. Five variables are included in this study. The dependent variable was the under-five mortality rate, measured as the number of deaths among children under-five per 1000 live births. The explanatory variables were public health expenditure (PHE), measured as domestic general government health expenditure; HIV prevalence rate (HIVPR), measured as the number of HIV infections per 100,000 individuals; technology (TECH), measured as the number of individuals using the internet; and female literacy rate (FLR), measured as the female primary school enrollment rate.

4. RESULTS AND DISCUSSION

Table 1 lists the descriptive statistics which offer a snapshot of the central tendency, variability, and distribution of the variables, providing an understanding of the dataset. The data indicate that the mean (average value) under-five mortality rate was 2.190859, with a standard deviation of 0.101181. Additionally, for public health expenditure, standard deviation is 0.109207, and its average value is estimated to be 1.207256. The HIV prevalence rate has a mean value of 1.527778 and a standard deviation of 0.377670. Similarly, technology has a standard deviation of 4.321579 and a mean of 14.74815. However, the female literacy rate has a mean value of 83.18652 and a standard deviation of 4.988682. The median values, which represent the midpoint of the distribution, indicate that half of the observations fall above or below these figures. The Kurtosis of under-five mortality, technology and female literacy rates are <3, indicating platykurticity. Whereas, those of public health expenditure and HIV prevalence rate are >3 indicating leptokurticity.

Table 1. *Descriptive statistics*

	UMR	PHE	HIVPR	TECH	FLR
Mean	2.190859	1.207256	1.527778	14.74815	83.18652
Median	2.183708	1.201628	1.55	17.01462	83.64627
Maximum	2.321305	1.429614	2.1	19.13475	92.15672
Minimum	2.021411	0.858742	0.6	8.573574	72.81351
Std. Dev.	0.101181	0.109207	0.37767	4.321579	4.988682
Skewness	−0.045045	−0.332164	−0.525426	−0.351579	−0.324534
Kurtosis	1.555252	4.695917	3.026495	1.300186	2.344129
Jarque-Bera	3.14312	4.976199	1.657488	5.075697	1.277182
Probability	0.207721	0.083068	0.436597	0.079036	0.528036
Sum	78.87093	43.46121	55	530.9333	2994.715
Sum Sq. Dev.	0.358318	0.417415	4.992222	653.6615	871.0432
Observations	36	36	36	36	36

Table 2. *Results of the unit root test*

Variables	ADF statistics	Critical value	Stationarity status
LogUMR	−9.79295	−3.54849	1(0)
LogPHE	−3.613128	−3.544284	1(0)
HIVPR	−6.595028	−3.544284	1(0)
TECH	−3.724449	−3.54849	1(1)
LogFLR	−4.294966	−3.557759	1(1)

4.1. Stationarity. The ADF test was employed to assess the present of unit root in the data. Each variable was initially tested for the presence of both a trend and an intercept. Because all five variables exhibited both, the ADF test was conducted, including both a trend and an intercept, in the regression (Table 1).

The results of the unit root test in Table 2 show that three variables (UMR, PHE, and HIVPR) were stationary, whereas two variables (TECH and FLR) were stationary at the first difference. The mixed order of stationarity of the variables enables the use of ARDL for analysis.

4.2. Cointegration Results. The results of cointegration based on the ARDL model are listed in Table 3 and are based on three maximum lag selections.

The cointegration test results in Table 3 were obtained using the ARDL bounds test. This is necessary to avoid spurious regression results. Table 3 lists the result of the long-term cointegration test between the dependent and independent variables in the model. The results reveal a long-run relationship between the dependent and independent variables. This conclusion is supported by the F-statistic obtained from the F-bound test, which is greater than both the lower and upper critical values at the 5% significance level.

4.3. Long-term Coefficients. Table 4 list the long-term ARDL coefficient results of PHE, HIVPR, TECH and FLR in Nigeria. The results show that PHE has a negative and significant impact on the under-five mortality, having a coefficient of −0.20224848 and a p-value of 0.0000 at the 1% significance level. This implies that a 1% increase in public expenditure on health will lead to a 0.202248% decrease in the under-five mortality rate in the long-term. The result is consistent with the theoretical expectation of this model. Furthermore, the HIVPR and TECH are statistically significant and have a positive impact on the under-five mortality rate, having p-values of 0.0236 and 0.0291 respectively. In the long-term, this implies that a 1% increase in HIVPR and TECH leads to 0.057894% and 0.002766%, respectively, increase in the under-five mortality rate. Furthermore, the results show that FLR has no significant relationship with the under-five mortality rate in the long-term, as indicated by its p-value of 0.4340.

4.4. Error Correction Mechanism. Because the variables were found to be cointegrated, implying that they have long-term equilibrium relationships, short-term relationships were investigated.

The Table 5 lists the ARDL short-term coefficients obtained with an ECM. The results show that in the short-term, PHE has a positive and statistically significant impact on under-five mortality rates with a coefficient of 0.178165 and a p-value of 0.0277. This implies that a 1% increase in PHE results in a 0.179165% increase in the mortality rate of those under the age of five. This may be because, unlike the long-term period, the short-term

Table 3. *ARDL bounds testing for Cointegration*

F-bounds test		Null hypothesis: No levels of relationship		
Test Statistic	Value	Significance.	I(0)	I(1)
			Asymptotic: n=1000	
F-statistic	3.53215	10%	2.2	3.09
K	4	5%	2.56	3.49
		2.50%	2.88	3.87
		1%	3.29	4.37

Table 4. *Estimated ARDL long-term coefficients*

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PHE	−0.202248	0.026498	7.632648	0.0000
HIVPR	0.057894	0.021279	2.720685	0.0236
TECH	0.002766	0.001067	2.591946	0.0291
FLR	0.054116	0.066083	0.818912	0.434
CointEq(-1)	2.374781	0.145051	16.37201	0.0000

Table 5. *Error correction mechanism*

Variable	Coefficient	Std. Error	t-Statistics	Prob.
D(UMR(1))	1.109142	0.156586	7.083295	0.0001
D(PHE)	0.178165	0.066354	2.536555	0.0277
D(HIVP)	−0.040584	0.016899	−2.401619	0.0431
D(TECH)	−0.006708	0.002733	−2.454632	0.0396
D(FLR)	0.001246	0.000919	1.355756	0.2122
CointEq(-1)	−0.800535	0.183688	−4.358114	0.0024

R^2 = 0.996238, adjusted R^2 = 0.9916694, F-statistic = 3.532151 probability = 0.0003, and Durbin-Watson = 2.971880

Table 6. *Breusch Pagan Godfrey test for heteroskedasticity*

F-statistic	1.855698	Prob. F(22,9)	0.1695
Observed R^2	26.21981	Prob. x^2 (22)	0.2424
Scaled explained sum of squares	2.262211	Prob. x^2 (22)	1.0000

period is insufficient for public expenditure to have a meaningful impact on the health sector. HIVP had a negative and statistically significant impact on the under-five mortality rate, having a coefficient of -0.040584 and a corresponding p-value of 0.0431. This implies that a 1% increase in HIV prevalence results in a 0.040584% decrease in the under-five mortality rate at the 5% significance level. This finding is contrary to a priori expectations and may be attributed to medical advances that prevent HIV transmission from mothers to unborn babies. Furthermore, the result of TECH in the long-term shows a negative and statistically significant relationship, having a coefficient of -0.006708 and a p value of 0.0396. This means that a 1% increase in technology usage results in a 0.006708% decrease in the under five mortality rate at the 5% level of significance. This finding is consistent with the a priori expectations of this study, indicating that improvements in technology use in the health sector reduce the under-five mortality rate in Nigeria. The result in Table 5 also show that the FLR has a positive but statistically insignificant effect on under-five mortality, having a coefficient of 0.001246 and a p value of 0.2122. The results further indicate an error correction term (ECT) value of −0.800535 with a p-value of 0.0024, which is consistent with the theoretical expectation that the ECT should be negative. This implies that the speed of adjustment from the short-term disequilibrium to the long-term equilibrium is approximately 8% annually. The model also had a good fit, as indicated by the R^2 and adjusted R^2 values of 0.996238 and 0.991669 respectively, indicating that the variations in the dependent variable are well explained by the explanatory variables in the model.

4.5. Heteroscedasticity Test on the ARDL Model. As the test result in Table 6 shows, the F-statistic, observed R^2, and the scaled explained sum of squares obtained from the Breusch Pagan Godfrey test shows that the ARDL model is free from heteroscedasticity because their probability values are greater than 0.05 (5% level of significance)

4.6. Cumulative Sum Test for ARDL Model Stability. The cumulative sum (CUSUM) test (Figure 1) shows that CUSUM falls within the critical region, indicating that the model parameters are stable over the sample period (1988 – 2023).

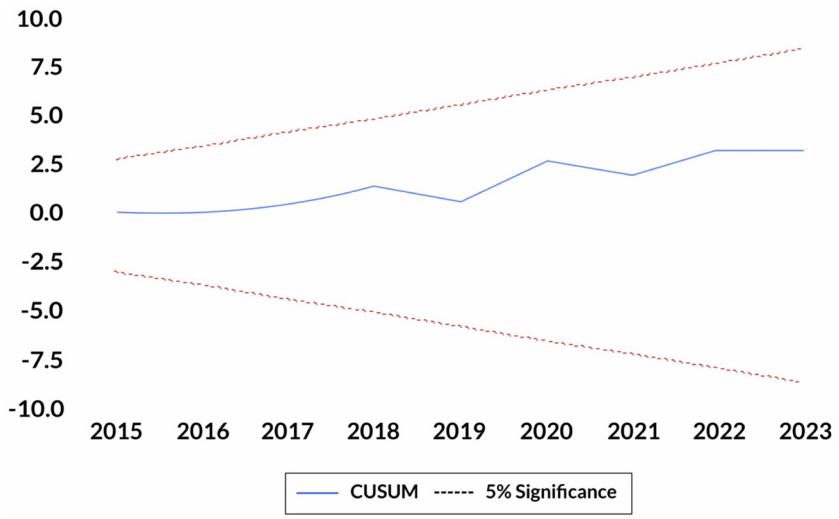


Figure 1. Cumulative sum test results

4.7. Discussion. The results of the analysis of the relationship between under-five mortality and public health expenditure revealed that in the long-term, public expenditure on health has a negative and significant impact on the under-five mortality rate at the 5% significance level. This result aligns with the findings of Azuh et al. and Kato et al. [12,16], but contrasts with the findings of Yaqub et al. [9], who reported that an increase in public health expenditure leads to an increase in under-five mortality rate.

However, the results of the analysis of the relationship between the female literacy rate and under-five mortality rate show that the female literacy rate has a positive but statistically insignificant impact on under-five mortality at the 5% significance level in both the short-term and long-term periods. This finding contradicts that of Aderopo and Emmanuel [11]. However, the results suggest that female literacy did not significantly improve the well-being of children under five years of age in Nigeria during the period under review.

5. CONCLUSIONS

In this study we examine the impact of public health expenditure on the under-five mortality rate in Nigeria from 1988 to 2023 using time-series data obtained from the CBN Statistical Bulletin and World Bank WDIs. The findings indicate that increased public health expenditure significantly reduces under-five mortality in the long-term, although the short term effect was found to be contrary. This underscores the importance of sustained public health investments in improving child health outcomes. Additionally, the study found a negative and statistically significant relationship between technology and under-five mortality in the short-term, suggesting that technologicaladvancements contribute to better child health outcomes in Nigeria. Based on these findings, the following recommendations are proposed:

(1) The demonstrated long-term effect of public health expenditure reducing under-five mortality highlight the need for the Federal Government of Nigeria to increase and sustain funding in the health sector. Priority areas should include infrastructure, health workforce development, and public health education campaigns. Transparent disbursement mechanisms must be implemented, involving effective oversight from relevant agencies such as the Economic and Financial Crimes Commission, to prevent the diversion or misuse of funds

(2) The government, through the Ministry of Communications and Digital Economy, should continue promoting the use of innovative health technologies to strengthen healthcare delivery, particularly in pediatric care. Mobile health applications should be deployed to empower parents and caregivers by offering tools to track children's health metrics, access educational content, and receive timely reminders for vaccinations and checkups.

(3) The government should intensify its investment in health research and development, especially to prevent mother-to-child transmission of HIV during pregnancy. The ministry of Education should prioritize female education, recognizing the pivotal role women play in family health decision-making. Programs aimed at educating girls not only improve their health but also enhance their capacity to advocate for their families' well-being.

ACKNOWLEDGEMENTS

First and foremost, the authors would like to express their profound gratitude to Almighty Allah for giving them the strength, wisdom and good health to carry out this research work. We also praise Him for his grace, guidance, blessings and infinite mercy on us. We are most grateful to acknowledge the support provided by the staffs and students of Economics Department, Nigerian Army University Biu. The authors also extend their appreciation to the reviewer of this research for their outstanding support toward ensuring this research is in line with the standard, and finally to family and friends thank you for your individual contributions.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Aliyu Inuwa Kamara – *Department of Economics, Nigerian Army University Biu (NAUB), Nigeria;*
0009-0008-4291-2629
Email: aliyuinuwa30@gmail.com

Corresponding Author

Abubakar Orlando Ijoko – *Research Mentor, Department of Economics, Nigerian Army University Biu (NAUB), Nigeria;*
0000-0001-5632-2365
Email: abuijoko@gmail.com

REFERENCES

(1) Ijoko, A. O.,Magaji, S. & Gombe, B. M. An empirical analysis of the impact of public expenditure on health infrastructure in primary healthcare centres in FCT. *Proc. First Int.Conf. Dept. Econ., Gombe State univ., Nigeria, 74-81,(2021)*

(2) United Nations. *World Population Prospects 2019*. (2020).

(3) World Health Organization. Children: Reducing mortality and improving well-being. *WHO Fact Sheet*. https://www.who.int/fr/news-room/fact-sheets/detail/children-reducing-mortality (2020)

(4) UNICEP. WHO levels and trends in child mortality. https://data.unicef.org/country/nga (2020).

(5) World Bank. Nigeria overview. https://www.worldbank.org/en/country/nigeria/overview. (2022).

(6) Ijoko, A. O. Impact of public expenditure on health services delivery in Federal Capital Territory, Nigeria. (Unpublished PhD thesis, 2023).

(7) Edeme, R. K., Emecheta, C. & Omeje, M. O. O. Public health expenditure and health outcomes in Nigeria. *Am. J. Biomed. Life Sci.* 5(5), (2017).

(8) Zubair, K. T. Impact of public health expenditure on health status in Nigeria. (PhD thesis, Kwara State University, Nigeria, 2018).

(9) Yaqub, J. & Ojapinwa, T. V. Public health expenditure and health outcome in Nigeria: The impact of governance. *Eur. Sci. J.* 8(13), 189–201 (2018). https://eujournal.org/index.php/esj/article/viewFile/206/248

(10) David, J. Infant mortality and public health expenditure in Nigeria: Empirical explanation of the nexus 1980 – 2016. *Timisoara J. Econ. Bus.* 11(2), 149–164. (2018).

(11) Aderopo, R. A., & Emmanuel, E. Can public health spending and maternal education predict future under-5 mortality rate in Nigeria. *Indian J. Econ. Dev*. 7 (12), 1–9. (2019).

(12) Azuh, D. E., Osabohien, R., Orbih, M. & Godwin, A. Public health expenditure and under-five mortality in Nigeria: An overview for policy intervention. *J. Med. Sci.* 8(E), 353–362 (2020).

(13) Ojo, O. O. et al. Health expenditure and life expectancy in Nigeria. *Lead City J. Soc. Sci.* 5(1), 66–71 (2020).

(14) Iyakwari, A. D. B., Awujola, A., & Ogwuche, D. D. Effect of health expenditure onlife expectancy in Nigeria. *Lafia J. Econ. Manage. Sci.* **8**, 105–118 (2023).

(15) Orji, A. et al. Are wealthy countries always healthy? Health outcomes and public health spending nexus in Nigeria. *SAGE Open* **11**(3), 21582440211040793 (2021).

(16) Kato, K. et al. Exploring effect of public health spending on under-five mortality rate in Uganda. *Afr. J. Econ. Rev.***6**(1), 1–14 (2018).

(17) Rana, R. H., Alam, K. & Gow, J. Health expenditure, child and maternal mortality nexus: A comparative global analysis. *BMC Int. Health Hum. Rights* **18**(29), 1–15 (2018).

(18) Rahman, M. M., Khanam, R. & Rahman, M. Health care expenditure and health outcome nexus: New evidence from the SAARC-ASEAN region. *Global Health* **14**(1), 1–11 (2018).

(19) Shahraki, M. Public and private health expenditure and life expectancy in Iran. *Payesh (Health Monitor)* **18**(3), 221–230 (2019).

(20) Logarajan, R. D., Mohamed Nor, N., Sirag, A., Said, R., & Ibrahim, S. The impact of public, private, and out-of-pocket health expenditures on under-five mortality in Malaysia. *Healthcare (Basel),* **10**(3), 589 (2022).

(21) Ayipe, F. I. & Tanko, M. Public health expenditure and under-five mortality in low-income sub-Saharan African countries: A panel data analysis. *SSRN* 4389168 (2023).

(22) Li, Y. et al. Impact of healthy lifestyle factors on life expectancies in the US population. *Circulation* **138**(4), 345–355 (2021).

(23) Fayissa, B. & Gutema, P. Estimating a health production function for sub-Saharan Africa (SSA). *Appl. Econ.* **37**(2), 155–164 (2005).

(24) Grossman, M. On the concept of health capital and the demand for health. *J. Polit. Econ.* **80**(2), 223–255 (1972).

(25) Mubarak, A. Impact of public expenditure and health outcomes (maternal mortality rate and under-five mortality rates) in Nigeria. *Dept. Econ., Nigerian Army Univ., Biu*, 13–14 (2023).

(26) Central bank of Nigeria. *Annual Statistical Bulletin*. https://www.cbn.gov.ng/documents/Statbulletin.html (2023).

KFUPM
Journal of Undergraduate Research International

Diamond Open Access 16
Article Chemistry

Effects of Organic and Inorganic Hardeners on Properties of Foamed Gypsum-Cement Composites

Hassan M. H. Muhammad[1] and M. A. Tantawy[2]*

Cite https://doi.org/10.64589/juri/209724

Submitted: June 04, 2025 **Revised:** July 14, 2025 **Accepted:** August 20, 2025

ABSTRACT

This study investigated the effects of organic and inorganic hardeners on the microstructure, hydration, and physical properties of foam gypsum–cement composites for lightweight construction applications. Raw materials, including gypsum plaster and white ordinary Portland cement (OPC), were first characterized using X-ray fluorescence, X-ray diffraction (XRD), Fourier transform infrared (FTIR) spectroscopy, thermogravimetric analysis (TGA), and particle size distribution (PSD) analysis. Three hardeners were studied: two organic (styrene–butadiene rubber [SBR] and polyvinyl acetate [PVA]) and one inorganic (sodium metasilicate [SMS]). The impact on the hydration of the foamed-gypsum pastes was assessed via XRD, FTIR spectroscopy, TGA, and scanning electron microscopy (SEM). XRD and FTIR spectroscopy confirmed the partial hydration of hemihydrate to dihydrate. The residual hemihydrate was most pronounced in the PVA- and SBR-modified samples, indicating inhibited hydration. TGA revealed the presence of multiple $CaCO_3$ phases, indicating carbonation of OPC and residual gypsum carbonate. SEM images highlighted the differences in pore structure and crystal morphology among the mixes, with denser matrices and smaller crystallites observed in the sodium silicate-modified pastes. PVA significantly reduced the bulk density of the foamed-gypsum pastes by stabilizing the foam, resulting in highly porous structures with reduced compressive strength and increased water absorption. SBR achieved moderate porosity with increased strength owing to its film-forming properties. SMS significantly improved the early setting, compressive strength, and shrinkage control, indicating accelerated hydration and matrix densification. While PVA delayed setting and moderately reduced shrinkage, SBR offered balanced performance in terms of strength and dimensional stability.

Keywords: foamed gypsum, hardeners, physicochemical properties, density, porosity, microstructure

1. INTRODUCTION

Foamed-gypsum is a lightweight and versatile construction material that has attracted increasing research and industrial interest because of its desirable combination of reduced density, ease of processing, environmental compatibility, and enhanced thermal and acoustic insulation properties. Compared with standard gypsum, its density is 30%–35% lower, depending on the foaming technique and additives employed[1,2]. The increased porosity resulting from the foaming process significantly reduces thermal conductivity, from approximately 0.57 W/m·K in standard gypsum to as low as 0.07–0.25 W/m·K in foamed variants, enhancing its effectiveness for thermal insulation[1]. Additionally, foamed-gypsum exhibits increase in sound absorption and reduction in noise coefficients[3,4]. Although the compressive strength diminishes with reduced density, the incorporation of additives such as microfibers or expanded vermiculite can aid in maintaining or enhancing the strength and stability[2,5]. Furthermore, foamed gypsum, particularly when combined with specific additives, exhibits excellent fire retardancy and retains structural integrity at elevated temperatures[4,6].

Foamed-gypsum serves a variety of purposes, including construction panels and boards such as lightweight plasterboards, thermal insulating plasters, and core materials for gypsum boards[1]. It is also used in non-load-bearing walls for interior partitions and hollow blocks, where high insulating performance and low weight are important[4]. In sustainable construction, the integration of waste materials (e.g., stone dust and rice straw) enhances environmental compatibility and cost-effectiveness[7]. Additionally, foamed gypsum has specialized applications in enhancing fire resistance and cleaning of delicate surfaces in conservation work[6,8].

Introducing stable foam into gypsum creates a lightweight, porous material with excellent insulation but also leads to reduced mechanical strength and increased water absorption. Several strategies involving supplementary materials and modifiers have been developed to address these drawbacks and to expand the use of foamed-gypsum in demanding environments. Methods for enhancing the mechanical properties of foamed-gypsum include the addition of synthetic polymers such as sulfur polycarboxylates and other synthetic polymers[9]; the addition of water-reducing agents such as naphthalene and

polycarboxylic acid [10]; and blending with cement, fly ash, lime [11], and nanosilica [12]. Methods to reduce the water absorption of foamed-gypsum include internal mixing and external coating with waterproofing agents such as hypromellose [13]; the addition of composites such as a nanosilica–silicone oil–paraffin emulsion [12]; the addition of sulfoaluminate cement, latex powder, and stearic acid [14,15]; and the inclusion of plastic waste [16]. Despite the abundance of current studies, there is a lack of evaluation of the impact of economic hardener additives, despite their effectiveness and importance, particularly in industry. There is a lack of systematic and comparable comparison between different additives (organic and inorganic hardeners) in the same foam system to evaluate their relative performance. Furthermore, the mechanisms by which hardeners interact with the foam to stabilize or destabilize air voids remain unclear. There are no quantitative hydration kinetic models for dihydrate and C–S–H formation in the presence of each hardener. There has been no exploration of dosage ranges or synergistic blends to optimize cost–performance tradeoffs. By addressing these gaps (through multiscale mechanistic studies, time-resolved hydration analyses, extended durability testing, and dosage optimization), foam gypsum–cement composites can be better tailored to meet insulation, structural, and durability requirements.

Styrene-butadiene rubber (SBR) has attracted increasing attention as a polymer modifier for improving the performance of cementitious materials, including foamed gypsum. However, research on the use of SBR as a hardener to improve the properties of foamed-gypsum paste is limited. SBR significantly improves the interfacial bonding properties of gypsum-based repair mortars and modifies the microstructure of the gypsum matrix, resulting in a denser and more cohesive material [17]. The addition of SBR imparts greater flexibility to the matrix, which can help accommodate stress and reduce the risk of cracking, thereby contributing to long-term durability [17]. SBR forms a flexible polymer film within the gypsum matrix, bridging microcracks and enhancing the overall cohesion of the material. This film also improves the resistance of the material to water ingress and chemical attacks [17].

Polyvinyl acetate (PVA) is used as a hardener and modifier to enhance the performance of foamed-gypsum and address key challenges such as mechanical weakness and water absorption. Studies have demonstrated that PVA significantly improves the mechanical strength, water resistance, and microstructural characteristics of gypsum-based materials. Adding PVA (typically approximately 2–3 wt%) to gypsum composites increases the flexural strength by approximately 3%–19% and the compressive strength by up to 19% in water-cured samples. The combination of PVA with other additives such as carbon fibers can increase the flexural strength by >60% [18]. The presence of PVA changes the gypsum crystal morphology from long needle-like to short compact forms, further improving the strength and water resistance [19].

Sodium metasilicate (SMS) is used as a hardener and activator to enhance the properties of gypsum-based and foamed-gypsum materials. Its addition can improve mechanical strength, influence shrinkage, and promote beneficial microstructural changes. Increasing the dosage of SMS as an activator in alkali-activated slag-gypsum mortars leads to a higher compressive strength. This is due to the formation of additional binding phases that strengthen the matrix. However, the addition of SMS may also lead to increased drying shrinkage [20]. SMS reacts with gypsum to form C–S–H gels, which are layered, flocculent, and fibrous. These gels fill pores, refine the microstructure, accelerate cement hydration, and significantly improve the early-age compressive strength of cement-based materials, making them denser and stronger [21].

The present experimental investigation explored the effects of three hardener solutions on the functional, mechanical, and durability characteristics of a baseline foamed-gypsum mix. The additives considered included SBR latex, a PVA emulsion, and an aqueous SMS solution. These materials were selected according to their known roles in improving bond strength, water resistance, shrinkage control, and durability in cementitious and gypsum-based systems. Each additive introduces distinct chemical interactions with the gypsum–cement matrix and is hypothesized to contribute uniquely to the foam stabilization, setting kinetics, and performance.

The foamed-gypsum mix design used in this study consisted of 95 wt% commercial-grade gypsum plaster ($CaSO_4 \cdot ½H_2O$) and 5 wt% Type I Portland cement, which slightly modified the setting behavior and enhanced the early-age strength. The water/binder mass ratio was set as 1.05:1 to ensure sufficient dispersion of the foam and additives. The foaming agent used was alpha-olefin sulfonate (AOS)—a surfactant solution that was mechanically aerated before mixing and added at approximately 3 wt% relative to the total dry binder content to produce a stable and uniform air-void system. Each additive solution was introduced separately at a selected dosage, and fresh mixes were cast into standardized molds. The specimens were cured under controlled temperature and humidity conditions (35 °C, 50% relative humidity) for 1 d to assess both early-age and mature properties.

By providing a side-by-side comparison of SBR, PVA, and SMS, this research addresses a critical gap in the field of foamed gypsum composites. The findings will not only clarify the distinct mechanisms by which these additives interact with the foam and gypsum-cement matrix but also offer a foundation for optimizing material properties. The insights from this study will pave the way for creating more durable, high-performance foamed gypsum composites that can be tailored for a wider range of construction applications, ultimately contributing to more sustainable and energy-efficient building practices.

2. METHODOLOGY

2.1. Materials. Commercial gypsum plaster and Type I white ordinary Portland cement (OPC) served as the binder components. Distilled water was used to prepare all the foamed-gypsum pastes. The AOS surfactant was mechanically foamed and added at 2 wt% relative to the total binder mass. Three hardener solutions were used: SBR latex (50 wt%), PVA emulsion (40 wt%), and SMS solution (Na_2SiO_3, 50 wt%). Table 1 presents the properties of the hardener solutions.

2.2. Mix preparation. Four mixes containing 95% plaster and 5% white OPC were mixed in a 1:1 water-to-mix ratio with

Table 1. *Properties of the hardener solutions*

	Hardener solution		
Property	**SBR**	**PVA**	**SMS**
Description	Milky white liquid with a strong, pungent odor	White liquid with a mild odor	Colorless, transparent, odorless liquid
Density, g/cm^3	1.029	1.097	1.035
Solid content, wt%	39	68	50

a 2% hardener solution. The mixes were labeled as follows: mix without a hardener (G0), mix containing the SBR hardener (G1), mix containing the PVA hardener (G2), and mix containing the SMS hardener (G3). Table 2 presents the compositions of these mixes. Figure 1 shows the dry mixture of plaster and cement, water mixed with the added hardeners, the foaming agent (AOS), and the hardeners.

The dry ingredients (plaster and OPC) were first homogenized in a low-speed mixer for 1 min. Distilled water containing the designated hardener solution (replacing an equivalent mass of water) was added, followed by mixing for 1 min. The pre-foamed AOS was then introduced, followed by mixing for 30 s at 300 rpm to ensure uniform foam distribution without collapse. The fresh foamed slurry was cast into steel molds (2.5 ×2.5 ×2.5 mm^3). The specimens were demolded after 30 min, cured at 35 ±1 °C with 50% ±5% relative humidity for 1 d, and dried in an oven at 80 °C overnight. Figure 2 presents a flowchart of the mixing process. Figure 3 illustrates the preparation of foamed-gypsum cubes G0–G3 and their curing in an incubator.

2.3. Testing physicochemical characteristics of foamed gypsum. The setting time of the foamed-gypsum paste was measured in a simple yet accurate manner using a practical hand-press (touch test) approach. After starting the timer at the beginning of the mixing, the paste behavior was observed by gently touching the surface with a finger every 2 min. The initial setting time was recorded as the time at which the paste began to resist finger pressure and felt hard but still plastic. The final setting time was recorded as the time when the paste became firm to touch and no longer yielded to pressure. All tests were completed in triplicate (three cubes) unless otherwise mentioned, and data are provided as the mean ±standard deviation.

The bulk density of the foamed-gypsum paste was measured using Archimedes' principle of buoyancy in accordance with ASTM C188-23[22]. The compressive strength of the foamed-gypsum was estimated using a compressive-strength apparatus in accordance with ASTM C109-80.

The percentage of gypsum hydration was calculated using the balanced equation $CaSO_4{\cdot}2H_2O = CaSO_4{\cdot}0.5H_2O + 1.5H_2O$. Applying the law of conservation of matter and taking into account the percentage of gypsum in the mixture (95%), the theoretical weight of the dehydrated gypsum sample, after heating 1 g of hydrated gypsum at 160 °C in a drying oven until the weight was constant, was $W_{theoretical} = 0.803$ g. If the actual weight of the dehydrated gypsum sample is W_{actual}, the percentage of unhydrated gypsum is calculated as follows:

$$\text{Unhydrated gypsum}\,\% = 100 \times (W_{actual} - 0.803)/0.953. \quad (1)$$

Accordingly, the percentage of hydrated gypsum is given as

$$\begin{aligned}\text{Hydrated gypsum}\,\% &= 100 - \text{Unhydrated gypsum}\% \\ &= \text{Unhydrated gypsum}\,\% = 100 \times (W_{actual} - 0.803).\end{aligned} \quad (2)$$

The total porosity of the foamed-gypsum pastes was measured in accordance with ISO 5018:1983[23]. The water absorption of the foamed-gypsum pastes was measured by immersing specimens in water at 25 °C for 24 h, followed by oven-drying to a constant weight. Absorption was represented as a percentage of the mass loss relative to the initial dry mass[24]. The linear shrinkage of the foamed-gypsum paste was calculated from the change in cube length after drying. The lengths before drying (L1) and after drying (L2) were measured using a stainless-steel electronic digital vernier caliper (Dpl, 150 mm) with an accuracy of +/- 0.02 mm[25].

$$\text{Linear shrinkage}\,\% = 100 \times (L1 - L2)/L1 \quad (3)$$

The distributions of the surface and internal pores of the foamed-gypsum cubes were examined via photography using a

Table 2. *Compositions of the foamed-gypsum mixes*

	Compositions, wt%			
Material	G0	G1	G2	G3
Gypsum plaster, g	95	95	95	95
OPC, g	5	5	5	5
Water, mL	105	103	103	103
Foam, g	2	2	2	2
Hardener solution, mL	-	2	2	2
Hardner type	-	SBR	PVA	SMS

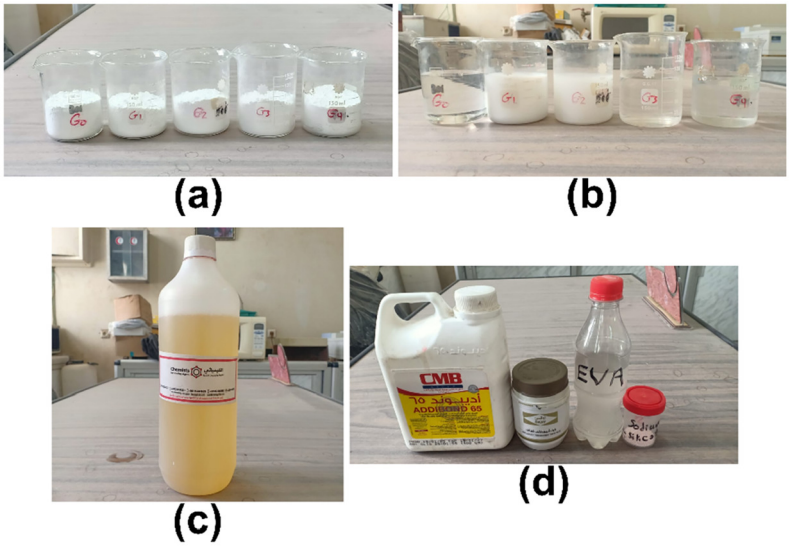


Figure 1. (a) Dry mixture of gypsum plaster and cement, (b) water mixed with the added hardeners, (c) the foaming agent (AOS), and (d) the hardeners

VIBOTON 1000 × WiFi digital microscope. Figure 4 illustrates the testing methods for the foamed gypsum.

2.4. Analysis of foamed-gypsum pastes. The chemical compositions and microstructures of the raw materials and selected foamed-gypsum pastes were explored using X-ray fluorescence (XRF), X-ray diffraction (XRD), Fourier transform infrared (FTIR) spectroscopy, thermogravimetric analysis (TGA), and scanning electron microscopy (SEM). The XRF analysis was performed using a Philips spectrometer (PW1606). XRD analysis was performed using a Philips diffractometer (PW1370) with a Ni-filtered Cu Kα radiation source. FTIR analysis was performed using a PerkinElmer System Spectrum X spectrometer in the range of 400–4000 cm^{-1}. TGA/differential thermogravimetric analysis (DrTGA) was performed using a Shimadzu thermal analyzer (DTG-60 H) at a heating rate of 10 °C/min up to 900 °C under a N_2 atmosphere. The SEM analysis was performed using a Jeol-Dsm 5400 LG instrument (Central Laboratory for Microanalysis and Nanotechnology, Minia University).

2.5. Characterization of raw materials.

2.5.1. XRF. The XRF oxide analysis results for the plaster are presented in Table 3. The CaO/SO_3 mass ratio (~0.697) was close to the theoretical value for pure bassanite ($CaSO_4 \cdot ½H_2O$) (~0.700), indicating the purity of the sample. The loss on ignition (LOI ≈ 0.697) was slightly higher than the theoretical value for well-formed hemihydrate (6.2). The small excess LOI is attributed to adsorbed moisture that caused the transformation of the bassanite into dihydrate ($CaSO_4 \cdot 2H_2O$). Minor impurities of SiO_2, Al_2O_3, Fe_2O_3, and MgO were attributed to the presence of residual carbonate, silicates/clays, and traces of associated magnesium sulfate or dolomitic inclusions in the natural gypsum source.

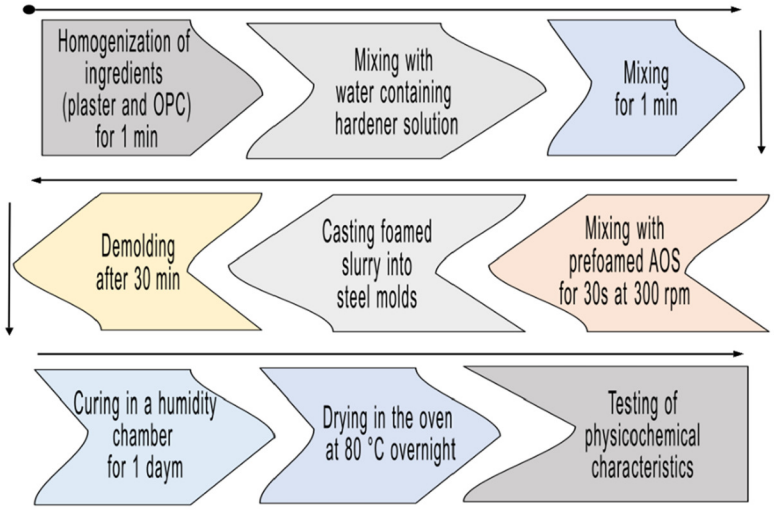


Figure 2. Flowchart of the mixing process

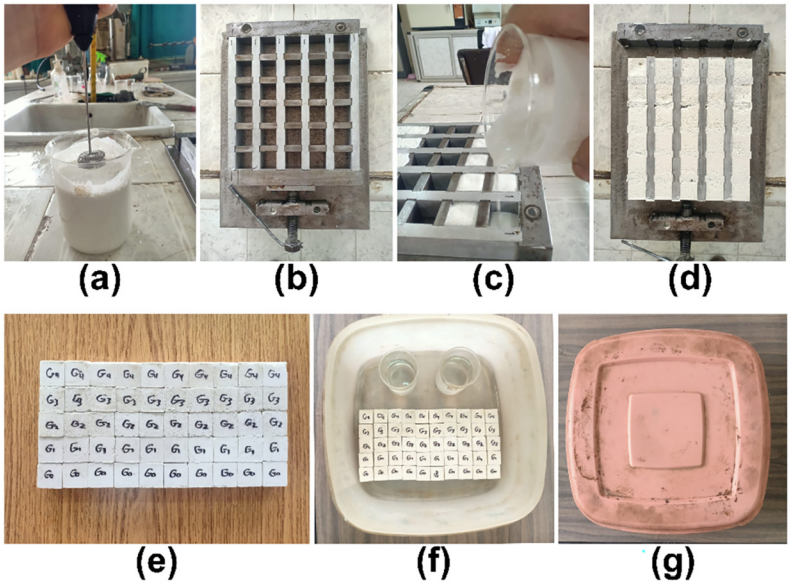


Figure 3. Preparation and curing of foamed-gypsum pastes G0–G3: (a) foaming process; (b) mold; (c) casting process; (d) finishing the cast, (e) demolding; (f) immersion in the humidity chamber; (g) closing the humidity chamber

The XRF analysis of white OPC confirmed a well-engineered OPC with excellent whiteness, strength potential, and low impurity content, which is suitable for architectural and aesthetic applications where light color is essential.

2.5.2. XRD. Figure 5 presents the XRD pattern of plaster, showing that the gypsum sample contains β-hemihydrate (Bassanite), as indicated by the sharp and intense peaks appearing at $2\theta \approx 14.7°, 25.4°, 29.1°, 31.3°, 34.4°, 47.3°$, and $50.6°$, matching the International Center for Diffraction Data (ICDD) card for β-$CaSO_4 \cdot 0.5H_2O$ (Powder Diffraction File of Bassanite (PDF) 00-041-0224) and indicating the presence of trace dihydrate. The sharp peaks imply good crystallinity, whereas the minimal sloping "hump" underneath indicates the presence of minimal amorphous or poorly crystalline material[26]. The XRD pattern of OPC revealed the following phases: The monoclinic and rhombohedral forms of C_3S were indicated by the strongest peak at $2\theta = 32.2°$ ($d \approx 2.78$ Å) with accompanying reflections at approximately 25.3° and 50.8°. The β-C_2S polymorph was indicated by the medium-intensity peaks at $2\theta = 29.4°, 34.1°$, and 53.9°. The C_4AF (ferrite) was indicated by the small shoulder peak at $2\theta = 35.1°$, and a weak feature at 31.0° can be assigned to cubic C3A. Minor peaks at $2\theta = 11.6°, 20.7°$, and 23.4° corresponded to the gypsum dihydrate added as a setting regulator. The absence of extraneous peaks confirmed that secondary-phase impurities were negligible[27].

2.5.3. FTIR spectroscopy. The FTIR spectra of plaster and OPC shown in Figure 6 indicate that hemihydrate (Bassanite) was the dominant constituent in the plaster, as evidenced by the appearance of a strong and sharp peak corresponding to the stretching vibration (ν_4) of SO_4^{2-} at 1140 cm^{-1}, the sharp peak corresponding to the asymmetric bending vibration (ν_3) of SO_4^{2-} at 660 cm^{-1}, and the broad O–H stretching envelope centered at 3400–3240 cm^{-1} corresponding to the structural water (H_2O) vibrations due to less ordered water molecules. Residual dihydrate was also present, as indicated by the peak corresponding to strong H–O–H bending vibration (δ) at 1620 cm^{-1} due to crystalline water, two sharp bands at 653 and 593 cm^{-1} associated with a combination of O–H out-of-plane bending modes and S–O bending vibrations specific to dihydrate, and the shoulder at 1093 cm^{-1} characteristic of the ν_3 splitting of SO_4^{2-}[28]. The FTIR spectrum of OPC exhibited the following absorption

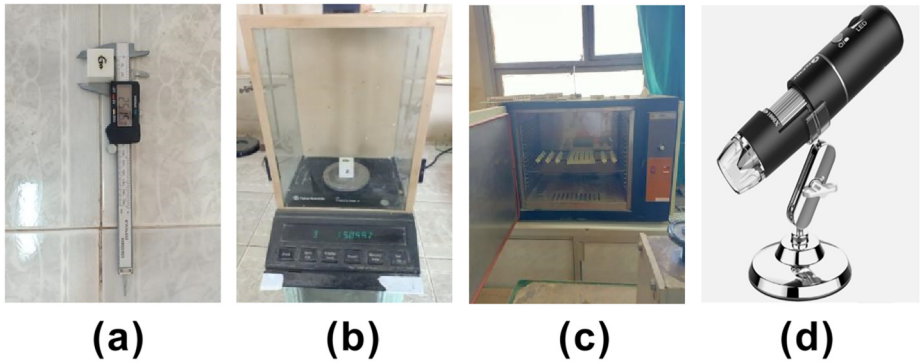


Figure 4. Testing methods for the foamed gypsum: (a) digital vernier caliper; (b) electronic balance; (c) drying oven; (d) 1000 × WiFi digital microscope

Table 3. *Chemical compositions of the raw materials*

Oxide wt%	Gypsum plaster	OPC
SiO_2	1.39	22.53
Al_2O_3	0.53	4.83
Fe_2O_3	0.17	0.68
CaO	36.2	64.1
MgO	1.49	3.15
Na_2O	0.02	0.13
K_2O	0.18	0.42
P_2O_5	0.09	0.22
SO_3	51.96	2.06
Cl^-	0.07	0.06
LOI*	7.55	1.36
Total	99.65	99.54

bands. H–O–H hydrogen-bonded stretching vibration was indicated by peaks at 3601 and 2506 cm^{-1}, and bending vibration appeared at 1797 cm^{-1}. The ν_2 out-of-plane bending vibration of the carbonate group appeared at 874 cm^{-1}, and the ν_3 asymmetric stretching vibration appeared at 1417 cm^{-1}. The ν_3 SO_4^{2-} stretching vibration of dihydrate and ettringite appeared at 1093 cm^{-1}, and the bending vibration (δ) of sulfate (SO_4^{2-}) of dihydrate and ettringite appeared at 665 cm^{-1}. The Si–O bending and M–O–Si lattice vibrations of belite appeared at 468 cm^{-1} [29].

2.5.4. TGA. Figure 7 shows the TGA/DrTGA thermograms of the plaster. The first weight loss of approximately 1.4% was due to moisture loss up to 100°C. The second weight loss of approximately 6.0 wt%, accompanied by a sharp DrTGA peak, was due to the transformation of hemihydrate ($CaSO_4 \cdot ½H_2O$) into anhydrite ($CaSO_4$) at 100–140° C. The third weight loss, i.e., a residual decline of approximately 2 wt% up to 700 °C, accompanied by the weak, broad DrTGA signal at approximately 500–700 °C, was due to decomposition of carbonate impurities [30].

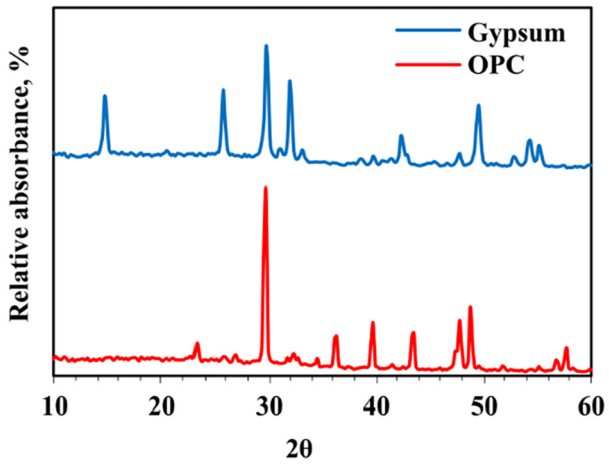


Figure 5. XRD patterns of gypsum plaster and OPC

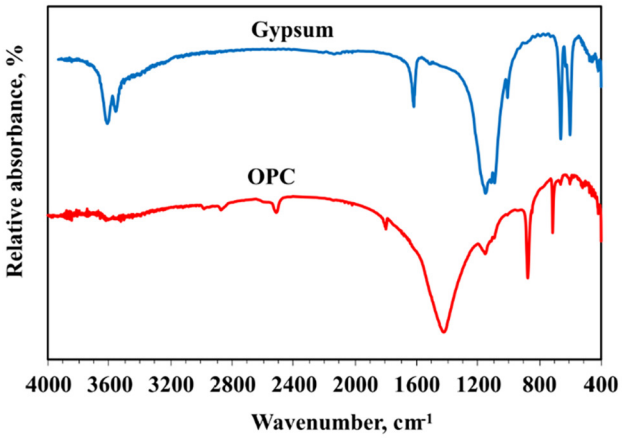


Figure 6. FTIR spectra of gypsum plaster and OPC

2.5.5. PSD. The particle size distributions (PSD) of plaster and OPC presented in Figure 8 indicate that the dominant particle size of OPC was approximately 20–30 μm, which is typical for common cement to optimize hydration and strength development. While the dominant particle size of plaster was approximately 10–20 μm, corresponding to the fine fraction, the particle size of the remaining portion was approximately 80–100 μm, corresponding to the coarse fraction. The suboptimal bimodal PSD of plaster with excessive coarse particles (>50 μm), likely due to incomplete grinding or agglomeration during the production process, may reduce early reactivity by slowing hydration and affect the setting time and strength [31,32].

3. RESULTS AND DISCUSSION

3.1. Phase identification of foamed gypsum.

3.1.1. XRD. Figure 9 shows the XRD patterns of G0–G3. The plaster is primarily composed of hemihydrate (bassanite), exhibiting characteristic peaks at 2θ = 14.8°, 14.7°, 25.8°, 29.8°, 32.0°, 42.6°, and 49.6°. The foamed-gypsum consisted of dihydrate, exhibiting characteristic peaks at 2θ = 11.6°, 14.7°, 20.7°, 23.6°, 25.5°, and 29.4°. Calcite was detected in the foamed-gypsum mixes, as indicated by characteristic peaks at 2θ = 29.4°, 47.6°, and 48.6°, owing to carbonation of portlandite produced from hydration of OPC. Residual hemihydrate was detected in the foamed-gypsum mixes, as indicated by characteristic peaks at 2θ = 14.8°, 25.8°, 32.0°, and 49.6°, owing to incomplete

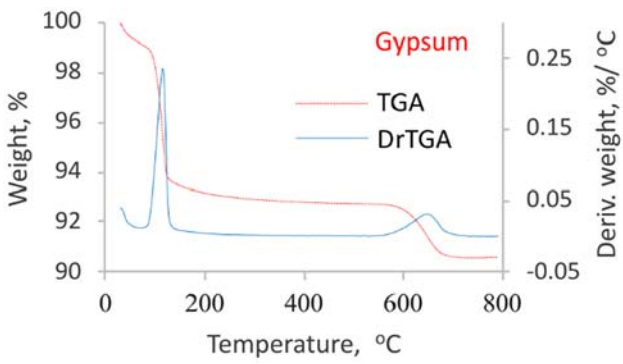


Figure 7. TGA/DrTGA thermograms for gypsum plaster

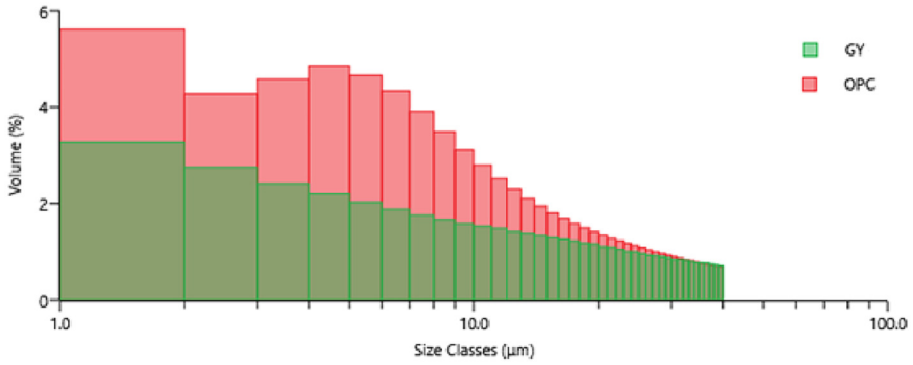


Figure 8. PSD graph of gypsum plaster and OPC

hydration of the plaster. The amount of remaining hemihydrate increased with the addition of organic hardeners (SBR and PVA), which bound the hemihydrate granules and prevented them from hydrating into the dihydrate. For the G3 paste containing sodium silicate, the intensities of the dihydrate and calcite peaks increased, whereas the intensity of the hemihydrate peak decreased, confirming the improved hydration of plaster in the presence of sodium silicate compared to other hardeners [26,27].

3.1.2. FTIR spectroscopy. Figure 10 shows the FTIR spectra of G0–G3. The multiple peaks at 3605 and 3557 cm^{-1} are characteristic of structural water (H_2O) vibrations of the hemihydrate. The sharp, strong peak at 1620 cm^{-1} is due to H–O–H bending vibration (δ) of crystalline water in dihydrate. The presence of this peak in the case of the plaster indicates partial hydration of the hemihydrate. The two sharp bands at 653 and 593 cm^{-1} are due to the combination of O–H out-of-plane bending modes and S–O bending vibrations specific to dihydrate. The weak sharp band at approximately 1005 cm^{-1} is due to the ν_3 asymmetric stretching of SO_4^{2-} in hemihydrate and dihydrate. The strong sharp band at approximately 1640 cm^{-1} is due to the H–O–H bending vibration of water molecules in dihydrate. The strong, broad band in the range of 1050–1220 cm^{-1} corresponds to the asymmetric stretching vibrations of the sulfate (S–O) groups in hemihydrate and dihydrate. The strong broad band at 1423 cm^{-1} and strong sharp band at 874 cm^{-1} are attributed to the ν_3 asymmetric stretching vibration and ν_2 in-plane bending vibration of CO_3^{2-} respectively, confirming that the carbonation of portlandite resulted from cement hydration. The intensity of the ν_2 peak was highest for G0, followed by G3, and was lower for G1 and G2 [28,29].

3.1.3. TGA. Figure 11 shows the TGA/DrTGA thermograms of G0–G3. The first weight loss of approximately 13.0 wt% was due to the dehydration of dihydrate ($CaSO_4 \cdot 2H_2O$) to anhydrite ($CaSO_4$) at 100–140 °C. This weight loss was lower than the theoretical value for dihydrate because of the low plaster content in the mixtures (95 wt%), presence of calcium carbonate impurities, and incomplete hydration of the plaster. The second weight loss of approximately 3.6 wt% up to 700 °C was accompanied by the distinct DrTGA shoulder and peak at approximately 600–800°C. The shoulder was at approximately 500–600°C, and

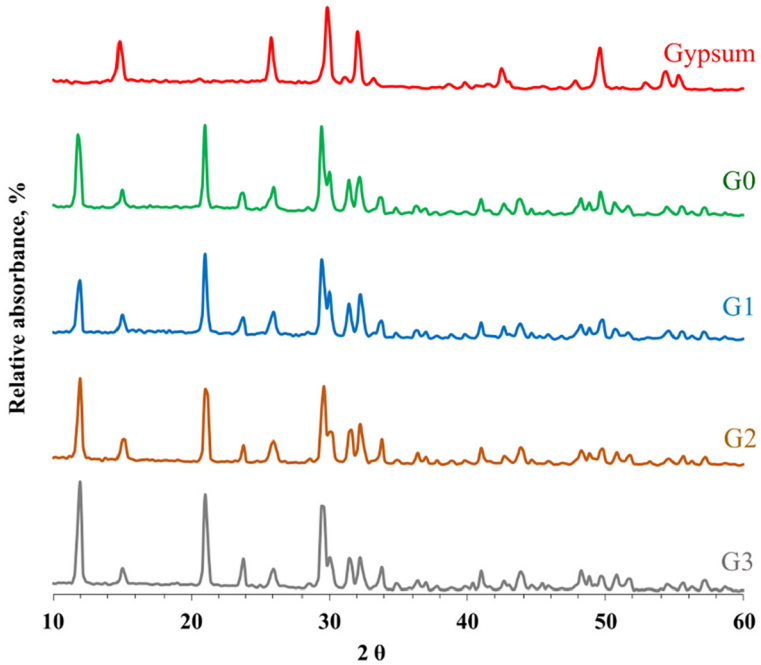


Figure 9. XRD patterns of gypsum plaster and foamed-gypsum pastes G0–G3

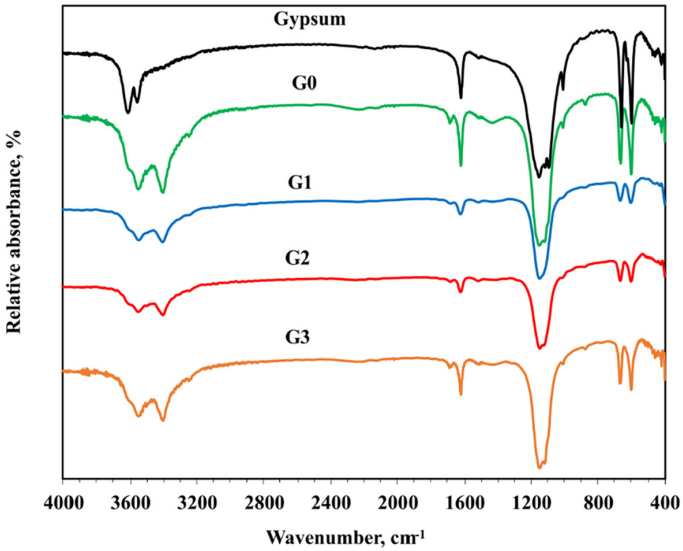


Figure 10. FTIR spectra of gypsum plaster and foamed-gypsum pastes G0–G3

the peak was centered around 700°C. These results indicated two origins of $CaCO_3$. The shoulder was due to vaterite or poorly crystalline $CaCO_3$ that originated from the carbonation of portlandite produced from the hydration of OPC. In contrast, the peak was due to the well-crystallized calcite of the natural $CaCO_3$ impurity already present in the plaster. The weight-loss ratio up to 200°C indicated the degree of hydration of the plaster. The weight-loss ratios were approximately 13.5, 13, 13.7, and 15.4 wt% for foamed-gypsum pastes G0–G3, respectively. This implied that the degree of hydration of the plaster was reduced by the addition of SBR and PVA and increased by the addition of SMS[30].

3.1.4. SEM. Figure 12 presents SEM images of G0–G3 at different magnifications (1000×, 2500×, and 5000×) and provides information about the microstructural evolution of foamed-gypsum pastes with different hardeners. The low-magnification images were used to examine the distribution and homogeneity of pores within the foamed-gypsum pastes. G0 exhibited a highly porous structure with large, irregular pores from the foam. The dihydrate crystals appeared as coarse interlocked plates/needles. The crystals were relatively large because of their unmodified growth. The pore walls were thin and fragile. The pore structures of both G1 and G2 were generally more uniform, and the pores were slightly smaller and better defined than those of G0. The pore walls were thicker and more robust. The pore structures in G3 differed significantly. The pores were smaller and more irregular. The matrix appeared denser and less crystalline compared to that in G0–G2[33]. High-magnification images were used to examine the crystal morphology of the dihydrate within the foamed-gypsum pastes. G0 contained large, thin, plate-like dihydrate crystals (often characteristic of unmodified gypsum pastes). In G1, the SBR polymer films were clearly visible, coating dihydrate crystals and forming continuous membranes/bridges between crystals[34]. In G2, the PVA films appeared as thicker, more localized coatings or globs concentrated at crystal junctions and surfaces, in contrast to the ultrathin, pervasive membranes of the SBR. G3 exhibited a significant change in microstructure. Individual dihydrate crystals were difficult to distinguish. Instead, a dense microcrystalline matrix was dominant.

3.2. Physicomechanical properties of foamed gypsum. Figure 13 presents the setting times of the foamed gypsum. The plain gypsum mix exhibited setting behavior typical of a plaster–cement mix with no additives. SBR delayed setting by forming a film on the plaster–cement grains, slowing early hydration. The SBR produces a flexible polymer matrix that slows ion diffusion, increasing the setting time. PVA has a stronger retarding effect than SBR owing to its higher water retention and surface-coating efficiency[17]. The PVA stabilized the foam and created more entrained air, extending the workability time. SMS functions as an accelerator, enhancing the hydration of plaster particles and promoting the early formation of binding gels (C–S–H and ettringite)[21].

The bulk density chart (Figure 14) reveals that not all hardeners effectively stabilize the foam in a foamed-gypsum matrix. The bulk densities of the SBR and SMS were nearly identical to those of the unmodified mix. This indicates that they do not appreciably improve the foam stability or air entrainment under these conditions. In contrast, the PVA hardener reduced the bulk density by approximately 35%–40%. The PVA hardener was effective in stabilizing air bubbles, resulting in a lighter and more porous structure. A lower density often results in better thermal and acoustic insulation but also implies lower compressive strength and probably more shrinkage. Therefore, the PVA hardener appears to be the most promising option for preparing lightweight plaster or insulating panels. Although SBR and SMS are desirable, they can be excluded to maintain high strength[17,21].

The unmodified mix G0 had the highest drying shrinkage (approximately 0.5%) owing to gypsum-paste contraction as water was lost, which can cause microcracking or adhesion loss in the plaster. The SBR hardener (G1) significantly reduced shrinking (~65%) but still resulted in ~0.2% contraction. The PVA hardener (G2) resulted in moderate expansion (approximately –0.08%), possibly because of its capacity to absorb water and produce flexible films in the pore walls, counteracting the capillary pull. In contrast, the SMS hardener (G3) provided almost full shrinkage control (~0.04%), likely owing to early stiffening of the set matrix and less plastic behavior. Thus, PVA is preferred for crack-free, dimensionally stable foamed-gypsum (e.g., thin panels or intricate moldings). The SMS hardener is particularly effective for managing shrinkage and fire resistance. Meanwhile, SBR provides moderate improvement, offering enhanced performance but to a lesser extent than the other hardeners.

Figure 15 shows the total porosity, compressive strength, water absorption, and degree of hydration of the plasters. PVA contributed significantly to the overall porosity by stabilizing the air voids during mixing and setting, resulting in a total porosity of 66.8%. In comparison, SBR and SMS slightly increased the porosity but not to the same extent, with total porosity values of 58.5% and 56.5%, respectively. Despite its high porosity, G1 had a high compressive strength of 686.5 kPa due to SBR's film-forming and toughening properties, which improved the matrix cohesion and flexibility[17]. G2 exhibited a substantial reduction in strength owing to its high porosity; its compressive strength was only 196.1 kPa. G3 had the highest strength of 833.6 kPa, likely due to its denser and more mineral-rich matrix[20].

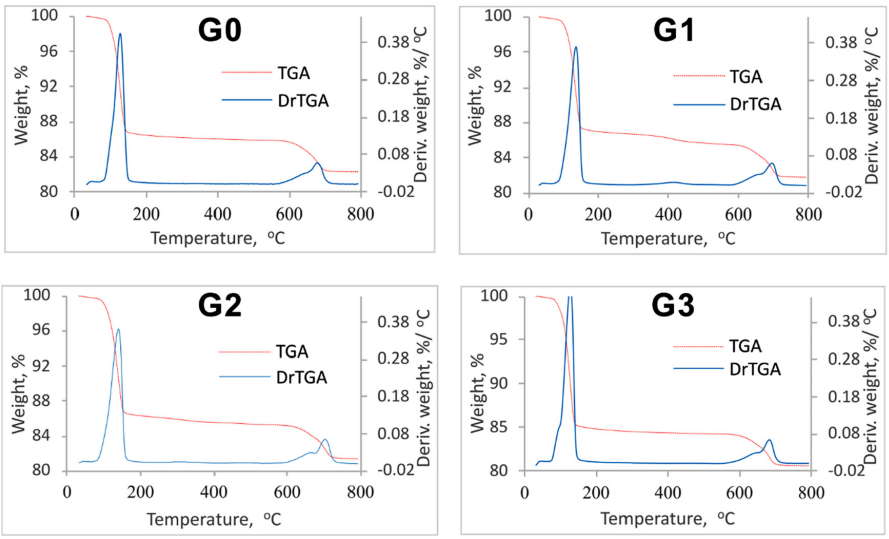


Figure 11. TGA/DrTGA thermograms of foamed-gypsum pastes G0–G3

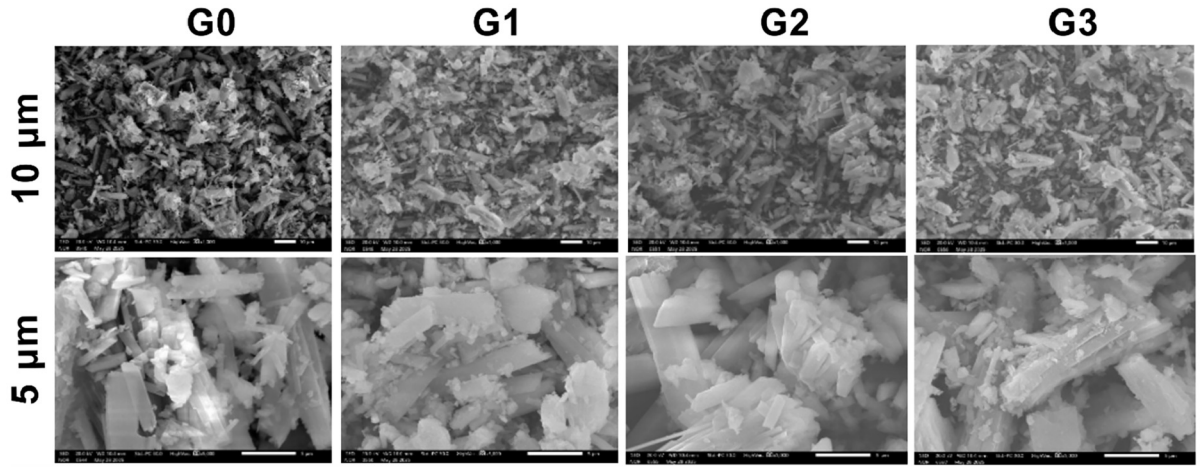


Figure 12. SEM images of foamed-gypsum pastes G0–G3 (10µm, 5µm)

The capacity of a material for water absorption exhibits a clear, direct relationship with its total porosity, such that increases in porosity result in proportionally greater water uptake. G0 exhibited moderate water absorption (101%), which is typical for foamed gypsum. G1 exhibited higher water absorption (113%) owing to the larger holes introduced by the SBR. G2, which had the highest porosity, exhibited significant water uptake (173%), which may have negatively impacted its durability. G3's mild water absorption (110%) was impressive given its adequate porosity, indicating a refined pore structure. G0 exhibited a high degree of hydration (94.4%), which is a characteristic of unmodified plaster. G1 exhibited a lower degree of hydration (92.6%) because SBR may reduce water availability for hydration. Furthermore, G3 maintained a high degree of hydration (93.9%) despite densification, indicating that SMS promotes the early hydration of plaster. G2 demonstrated reduced levels of hydration (93%), presumably because the air-void system interfered with water delivery. Excess porosity and the presence of polymers may hinder full hydration [19].

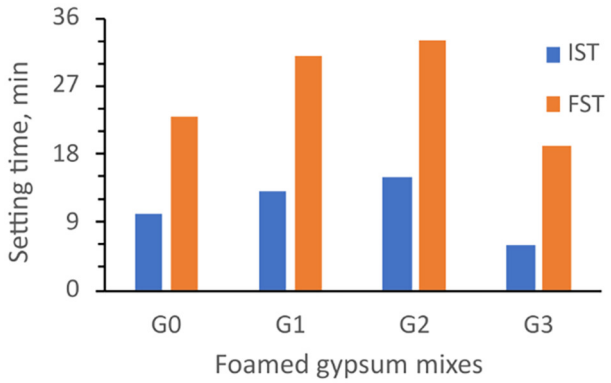


Figure 13. Initial and final setting times of foamed-gypsum pastes G0–G3

Thus, PVA is an ideal additive for lightweight insulating materials because of its high porosity and low density; however, it requires protective coatings owing to its poor strength and high water absorption. SMS additives are the best option for improving mechanical performance because they provide high strength and low shrinkage. Meanwhile, for balanced performance, the SBR additive provides a compromise among weight

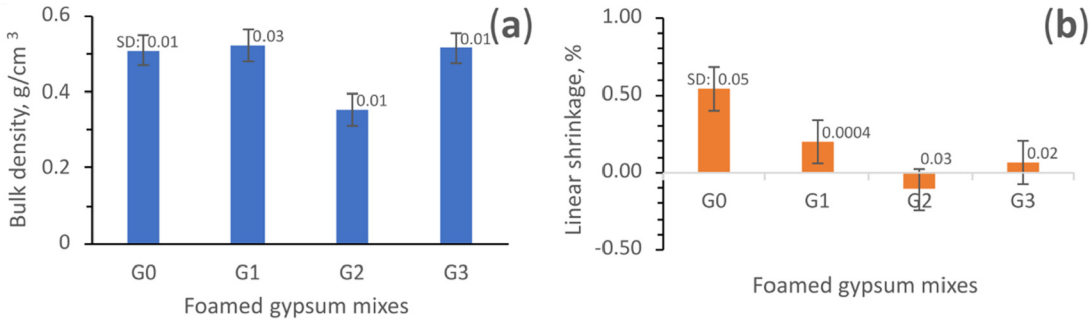


Figure 14. (a) Bulk density and (b) linear shrinkage of foamed-gypsum pastes G0–G3; SD represents the standard deviation

reduction, acceptable porosity, high strength, and dimensional stability.

The standard error ranges for the total porosity, compressive strength, and water absorption measurements of the foamed-gypsum samples were wider than those of the other measurements. This typically indicates that there is considerable scatter in the individual measurements (i.e., a large standard deviation due to the testing of few specimens). The main factors that tend to inflate the variability in foamed-gypsum systems are the heterogeneous pore structure and surface irregularities. The heterogeneous pore structure arises from differences in foam distribution, pore collapse, and pore connectivity during setting. The heterogeneous pore structure creates regions of high and low porosity, which decrease and increase the strength, respectively, thereby widening the spread of values. The larger connected pores soak water quickly, whereas the finer closed pores limit water ingress. Surface irregularities originate from nonparallel faces that are not present in larger, more uniform blocks.

Figure 16 shows how the surface and internal pores were spread out in the foamed-gypsum samples, as observed through digital optical microscopy. G0, which did not contain hardeners, had the fewest surface pores per unit surface area, with pore diameters of ≤125 µm, among the four mixtures. G1, which contained the SBR hardener, had the largest number of surface pores and the widest pores, with diameters exceeding 125 µm. G2, which contained the PVA hardener, had slightly larger surface pores than G0, with diameters no greater than 125 µm. G3, which contained the SMS hardener, had approximately the same number of surface pores as G2, with diameters no greater than 125 µm. G0 contained irregularly distributed pores, the largest of which exceeded 150 µm in diameter. G1 contained uniformly distributed pores, the largest of which did not exceed 150 µm in diameter. G2 contained irregularly distributed pores, the largest of which exceeded 500 µm in diameter. G3 contained uniformly distributed pores, the largest of which did not exceed 150 µm in diameter. The surface and internal pore distributions of the foamed-gypsum samples revealed that although G2 had less surface pores than the other mixtures, its internal structure contained the largest number of united pores. The internal structure of the paste appeared to be a highly porous spongy fabric, which

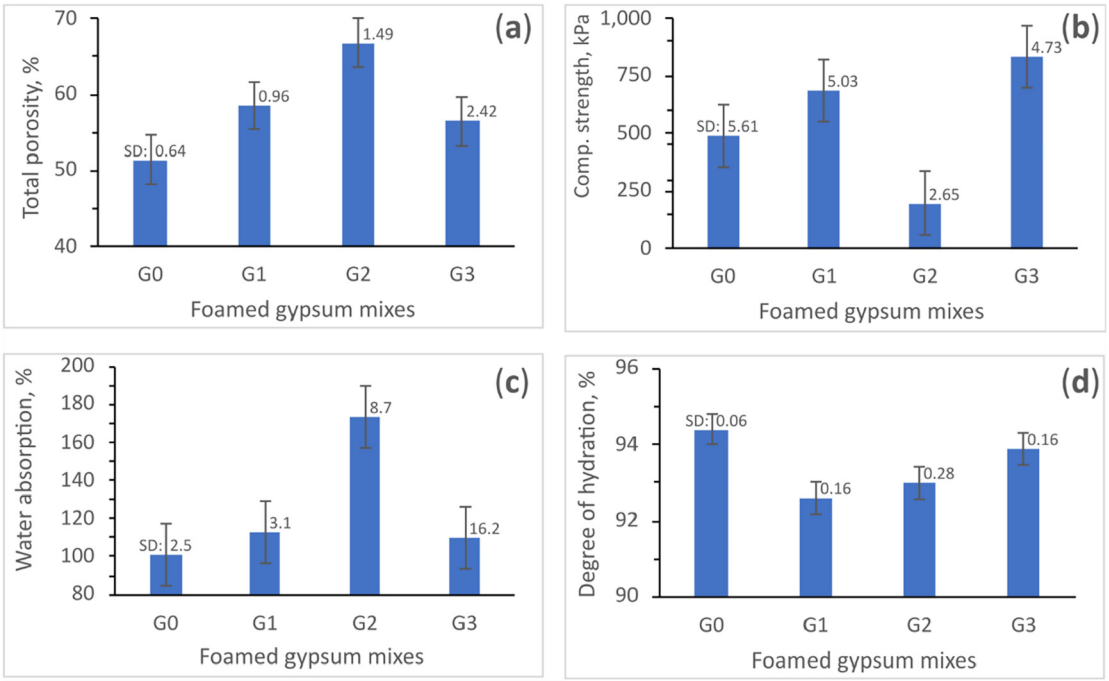


Figure 15. (a) Total porosity, (b) compressive strength, (c) water absorption, and (d) degree of hydration of foamed-gypsum pastes G0–G3; SD represents the standard deviation

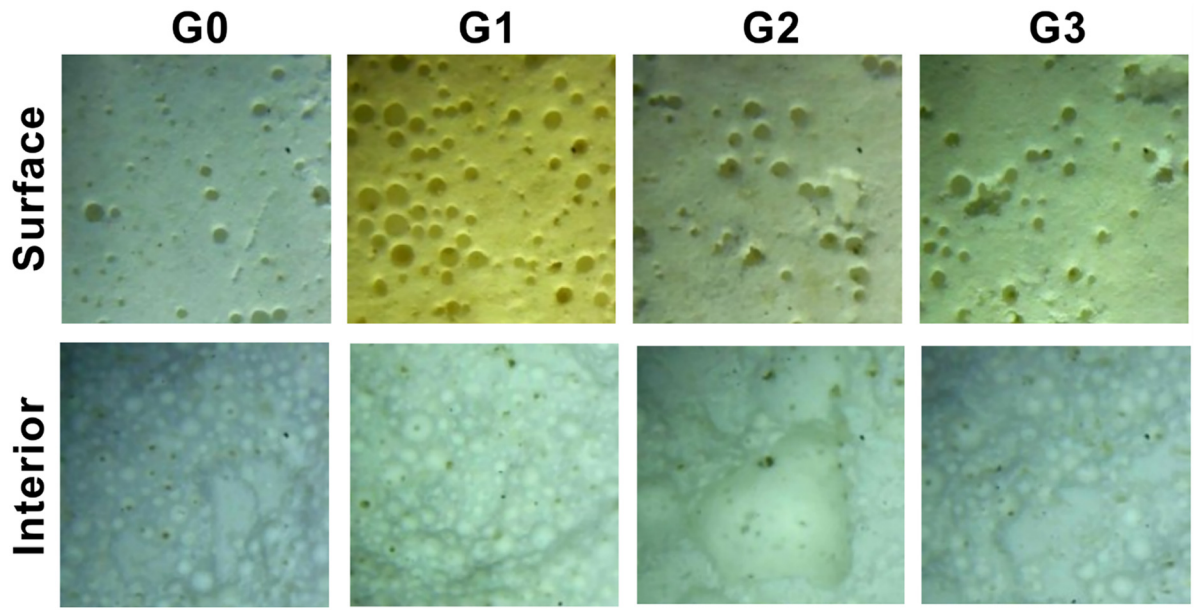


Figure 16. Pore distributions of the surface and interior of foamed-gypsum G0–G3, photographed by a digital microscope

explains its high total porosity and water absorption and lower density, compressive strength, and degree of sintering compared to the other mixtures. These imaging results may also help to explain the expansion of the G2 paste.

4. CONCLUSIONS

This study investigated the effects of organic (SBR and PVA) and inorganic (SMS) hardeners on the microstructure, hydration, and physical properties of foamed gypsum–cement composites. The following conclusions are drawn:

1. The gypsum plaster used was predominantly β-hemihydrate (bassanite) with excellent purity and crystallinity, as confirmed by XRF, XRD, FTIR spectroscopy, and TGA. The OPC also exhibited high quality, with minimal impurities and a suitable phase composition for gypsum–cement applications.
2. Organic hardeners (SBR and PVA) (G1 and G2) tended to reduce the degree of hydration owing to the coating effects on gypsum grains, whereas SMS (G3) enhanced hydration by functioning as an accelerator. SEM revealed that the organic polymers modified the pore structure and crystal morphology, with PVA (G2) producing finer, more uniform pore networks and SBR (G1) enhancing matrix cohesion.
3. Organic hardeners (particularly PVA (G2)) significantly retarded setting because of their water retention and surface-coating effects. SMS (G3) accelerated setting by promoting C–S–H gel formation.
4. PVA (G2) significantly reduced the bulk density and increased the total porosity owing to its strong foam-stabilizing effects, making it ideal for lightweight applications. In contrast, SBR (G1) and SMS (G3) did not significantly alter the density and maintained a denser and stronger structure.
5. SMS (G3) achieved the best shrinkage control, followed by PVA (G2), which reduced shrinkage or even induced slight expansion owing to the polymer flexibility and water absorption. SBR (G1) also improved the dimensional stability compared with the unmodified mix.
6. PVA (G2) is best suited for lightweight, insulating applications but requires surface protection owing to its low strength and high water absorption.
7. SMS (G3) is ideal for high-strength and dimensionally stable applications, offering excellent hydration, shrinkage control, and mechanical performance.
8. SBR (G1) offers a balanced solution with moderate improvements in porosity, strength, and shrinkage and is suitable for general-purpose foamed-gypsum products.
9. The surface and internal pore distributions of G2 indicate its highly porous internal structure, which is related to the degradation of mechanical properties.

This study systematically evaluated the effects of organic (SBR, PVA) and inorganic (SMS) hardeners on the hydration, microstructural, and macroscopic properties of foamed-gypsum paste, advancing the mechanistic understanding of foam stabilization. Key structure–property relationships were identified, yet the fundamental physical and chemical interactions between polymer membranes and AOS foam remain unresolved. Interpretation of these findings is limited by: (i) the one-day curing period, restricting long-term durability assessment; (ii) the use of laboratory rather than industrial-scale foaming, which may alter foam morphology, stability, and scalability; and (iii) the absence of advanced time-resolved characterization, constraining insight into dynamic foam–binder–hardener interactions. To address this in future work, we plan to resolve interfacial morphology by applying cryogenic SEM and atomic-force microscopy, and probe chemical dynamics in real time with in situ FTIR spectroscopy, will provide molecular-level insights critical for the rational design of durable, high-performance, and industrially scalable foam-stabilizing agents for next-generation gypsum materials. This integrated high-resolution and real-time characterization approach represents a decisive step toward

unlocking the molecular blueprint for durable, scalable foam-stabilizing agents, propelling gypsum-based materials into a new era of industrial performance.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Hassan M. H. Muhammad – *Chemistry Department, Faculty of Science, Minia University, El-Minia 61519, Egypt;* 0009-0008-7624-9140
Email: h.mu7ammad@outlook.com

Corresponding Author

M. A. Tantawy – *Research Mentor, Chemistry Department, Faculty of Science, Minia University, El-Minia 61519, Egypt;* 0000-0001-8401-6402

REFERENCES

(1) Capasso I, Pappalardo L, Romano R, Iucolano F. Foamed gypsum for multipurpose applications in building. *Constr. Build. Mater.* https://doi.org/10.1016/j.conbuildmat.2021.124948 (2021).

(2) Çolak A. Density and strength characteristics of foamed gypsum. *Cem. Concr. Compos.* 22, 193–200 (2000).

(3) Liu J, Xie H, Wang C, Han Y. Preparation of multifunctional gypsum composite with compound foaming process. *Powder Technol.* https://doi.org/10.1016/j.powtec.2023.118289 (2023).

(4) Singh S et al. Elevated temperature and performance behaviour of rice straw as waste bio-mass based foamed gypsum hollow blocks. *J. Build. Eng.* https://doi.org/10.1016/j.jobe.2023.106220 (2023).

(5) Sekavová H, Prošek Z, Tesárek P. Dependence of mechanical and thermal properties on the composition of lightweight gypsum composites. *Acta Polytech. CTU Proc.* https://doi.org/10.14311/app.2023.40.0088 (2023).

(6) Zhenxing D et al. Foamed gypsum composite with heat-resistant admixture under high temperature. *Cem. Concr. Compos.* 108, 103549 (2020).

(7) Krejsová J, Heralová S, Doleželová M, Vimmrová A. Environmentally friendly lightweight gypsum-based materials with waste stone dust. *Proc. Inst. Mech. Eng. Part L: J. Mater. Des. Appl.* 233, 258–267 (2019).

(8) Lascurain P et al. Agar Foam: Properties and Cleaning Effectiveness on Gypsum Surfaces. *Coatings.* https://doi.org/10.3390/coatings13030615 (2023).

(9) Wu H, Xia Y, Hu X, Liu X. Improvement on mechanical strength and water absorption of gypsum modeling material with synthetic polymers. *Ceram. Int.* 40, https://doi.org/10.1016/j.ceramint.2014.06.085 (2014).

(10) Yu J et al. Effects of water-reducing agents on the mechanical properties of foamed phosphogypsum. *Appl. Sci.* https://doi.org/10.3390/app14188147 (2024).

(11) Li Z, Wang X, Hou Y, Wu Z. Optimization of mechanical properties and water absorption behavior of building gypsum by ternary matrix mixture. *Constr. Build. Mater.* https://doi.org/10.1016/j.conbuildmat.2022.128910 (2022).

(12) Li J et al. Effect of nano-silica and silicone oil paraffin emulsion composite waterproofing agent on the water resistance of flue gas desulfurization gypsum. *Constr. Build. Mater.* 287, https://doi.org/10.1016/j.conbuildmat.2021.123055 (2021).

(13) Wang L et al. A novel approach for improving the water resistance of gypsum plaster by internal mixing hypromellose and external coating waterproofing agent. *Constr. Build. Mater.* https://doi.org/10.1016/j.conbuildmat.2023.132940 (2023).

(14) Yang S et al. A sustainable foamed material preparation via ettringite-targeted mineral transition of industrial solid wastes. *J. Clean. Prod.* https://doi.org/10.1016/j.jclepro.2022.134029 (2022).

(15) Li L, Li G. Research on the Waterproof Properties of Lightweight Gypsum Materials. *Appl. Mech. Mater.* 548–549, 1655–1659 (2014).

(16) Vidales-Barriguete A et al. Analysis of the improved water-resistant properties of plaster compounds with the addition of plastic waste. *Constr. Build. Mater.* https://doi.org/10.1016/j.conbuildmat.2019.116956 (2020).

(17) Shi C et al. Interfacial bonding properties of styrene-butadiene rubber and ethylene vinyl acetate emulsion-modified OPC–CAC–G repair mortar. *Constr. Build. Mater.* https://doi.org/10.1016/j.conbuildmat.2023.130747 (2023).

(18) Raouf Z, Mahdy N, Obaidi H. Improving the Properties of Gypsum By Using Additives. *J. Eng.* https://doi.org/10.31026/j.eng.2012.01.11 (2023).

(19) Khalil A, Tawfik A, Hegazy A. Plaster composites modified morphology with enhanced compressive strength and water resistance characteristics. *Constr. Build. Mater.* 167, https://doi.org/10.1016/j.conbuildmat.2018.01.165 (2018).

(20) Chen H et al. Effect of fly ash and gypsum on drying shrinkage and mechanical properties of one-part alkali-activated slag mortar. *Struct. Concr.* https://doi.org/10.1002/suco.202400006 (2024).

(21) He G et al. Preparation of C-S-H gels by mechanochemistry and its influences on properties of super-retarded cement-based materials with sucrose. *Cem. Concr. Compos.* https://doi.org/10.1016/j.cemconcomp.2024.105734 (2024).

(22) ASTM. Standard test method for density of hydraulic cement (C188-23). https://doi.org/10.1520/c0188-16 (2023).

(23) ISO. Determination of True Density of Refractory and Other Raw Materials. ISO 5018:1983 (1983).

(24) Li L, Li G. Research on the Waterproof Properties of Lightweight Gypsum Materials. *Appl. Mech. Mater.* 548–549, 1655–1659 (2014).

(25) Miao K et al. Effects of Boehmite on the Calcination Shrinkage and Mechanical Properties of Gypsum-Bonded Molds. *Adv. Eng. Mater.* 24, https://doi.org/10.1002/adem.202100683 (2021).

(26) Schmidt H, Paschke I, Freyer D, Voigt W. Water channel structure of bassanite at high air humidity: crystal structure of $CaSO_4 \cdot 0.625H_2O$. *Acta Crystallogr. Sect. B* 67, 467–475 (2011).

(27) Tobón JI, Mendoza Reales OA, Payá Bernabeu JJ. Performance of white Portland cement matrixes blended with nanosilica and limestone for architectural applications. *Adv. Cem. Res.* 28, 602–613 (2016).

(28) Al-Jobouri HA. FTIR Spectroscopy for Gypsum after Treatment with Steam Pressure. *J. Al-Nahrain Univ. Sci.* 14, 123–130 (2011).

(29) Shrestha SL. Characterization of Some Cement Samples of Nepal Using FTIR Spectroscopy. *Int. J. Adv. Res. Chem. Sci.* 5(7), 19–23 (2018).

(30) Kyono A, Ikeda R, Takagi S, Nishiyasu W. Structural evolution of gypsum ($CaSO_4 \cdot 2H_2O$) during thermal dehydration. *J. Mineral. Petrol. Sci.* 117, 015 (2022).

(31) Al-Mukhtar M, Al-Jabari M. The Effect of Particle Size Distribution on some Properties of Gypsum. *Key Eng. Mater.* 857, 145–152 (2020).

(32) Kim D. Effect of Adjusting for Particle-Size Distribution of Cement on Strength Development of Concrete. *Adv. Mater. Sci. Eng.* (2018).

(33) Rajković MB, Tošković DV. Phosphogypsum Surface Characterisation Using Scanning Electron Microscopy. *APTEFF* 34, 61–70 (2003).

(34) Rodríguez J et al. Flexural behavior and microstructure analysis of a gypsum-SBR composite material. *Mater. Lett.* 59(2–3), 230–234 (2005).

KFUPM
Journal of Undergraduate Research International

Diamond Open Access 28

Article Chemical Engineering/Chemistry

Effects of Reaction Temperature and Catalyst Type on Fluid Catalytic Cracking (FCC) of Crude Oil Feeds: A Microactivity Test Unit Study

Osama Wael Aljohani[1] and Abdullah M Aitani[2]*

Cite https://doi.org/10.64589/juri/207996

Submitted: February 25, 2025 **Revised:** July 28, 2025 **Accepted:** August 20, 2025

ABSTRACT

This study investigates the effects of reaction temperature and catalyst type on the fluidized catalytic cracking (FCC) of crude oil feeds in a micro-activity test (MAT) unit. It explores the variations in the reaction conditions that impact changes on product yields, specifically, the gas, liquid, and coke of the paraffinic feeds. A fixed-bed MAT setup featuring precise temperature settings and catalyst-to-oil (C/O) ratios is used. The experiment included gas chromatography, simulated distillation, and carbon analysis to analyze product production, composition, and yield. The results indicate that at higher temperatures and C/O, as opposed to thermal cracking, FCC employs a catalyst that cracks large hydrocarbons into smaller, valuable hydrocarbons. These ratios increased the overall conversion, and they also fostered the formation of olefins, such as propylene and ethylene. The results revealed the need to optimize the FCC parameters to balance conversion efficiency, product selectivity, and catalyst longevity while reducing coke formation and catalyst deactivation. This balance is a direct input for maximizing product value and efficiency in today's refineries.

Keywords: micro-activity test (MAT), fluid catalytic cracking (FCC), conversion efficiency, products' distribution, experimentation, catalyst longevity

1. INTRODUCTION

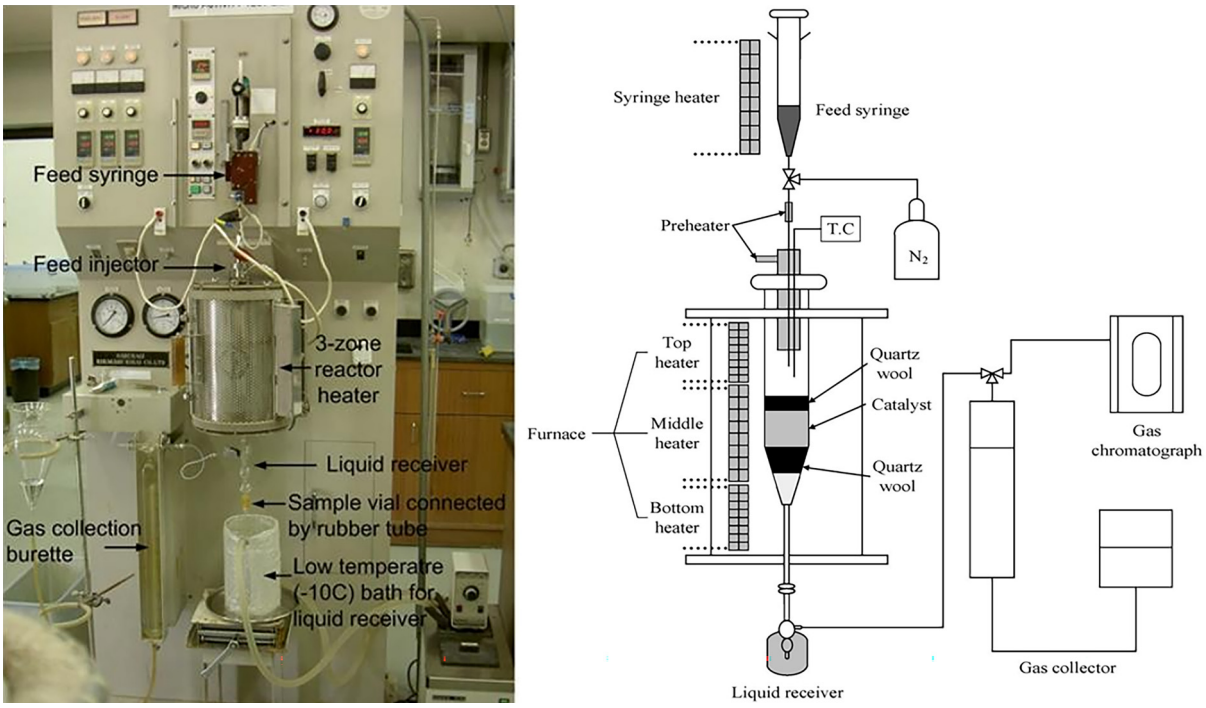
The transformation of crude oil into useful chemical products has been a major challenge for the petrochemical industry. Until now, crude-oil refining has primarily focused on gasoline and diesel production. However, the growing market for olefins, such as propylene and ethylene, has driven the development of new refining methods. One such approach—direct oil-to-light olefins—has led to significant market development.

Fluid catalytic cracking (FCC) is an essential technology instrumental to this transition. The share of FCC conversion process is approximately 16% of the total crude distillation capacity in oil refining worldwide. The initial objective of its development was to improve the production of gasoline from heavy oil. However, currently, FCC has become a widely adopted technology for refining all types of crude oil, such as gas oil and heavier residues, into higher value products [1–6]. The FCC process involves a powdered catalyst that becomes a fluid when added to hydrocarbons and then vaporizes it. In contrast to thermal cracking, FCC employs a catalyst to crack large hydrocarbons into small, valuable hydrocarbons, such as light olefins and aromatics, while also producing gases, liquid fuels, and coke as byproducts [7].

Several researchers have investigated the effect of temperature and catalytic properties on the FCC performance. Rahimi and Karimzadeh asserted that the reactor temperature was a conducive factor for conversion; however, it was associated with overcracking and coke formation [8]. Corma et al. explored the potential of zeolite-based catalysts and demonstrated that the catalyst acidity and pore size were key factors influencing product selectivity of the products [9]. Furthermore, Ancheyta et al. elaborated on the kinetics aspect and showed that the FCC process is conditioned by the type of feed used and the operation variables [10]. Microactivity test (MAT) units, as evaluated by Bjørgen et al., further verify catalyst activity and the positive effect that the supported materials and regeneration have in a lab-scale setup [11].

At King Fahd University of Petroleum and Minerals (KFUPM) Center for Refining & Advanced Chemicals, various reaction conditions on the performance of the FCC process were investigated with a MAT unit. The MAT unit is a suitable instrument for testing the products of FCC catalysts through a method in which the user controls the reaction on a small scale and has parameters such as temperature, pressure, and catalyst/oil (C/O) ratio finely tuned [12]. The FCC tests in this study utilize the standard procedure of placing the catalyst in the bed of the reactor, injecting a measured amount of crude oil, and heating using a syringe pump. As the crude feed contacts the catalyst, it cracks into gaseous, liquid, and solid products, which are directed into separate containers for quantification and conversion efficiency analysis.

Numerous studies have explored the optimization of the FCC for model feedstocks or under generalized catalytic conditions. However, few studies have systematically investigated the combined impact of temperature and catalyst type on crude oils with different compositions. This study evaluates FCC performance of light crude oil feeds such as Arabian Extra Light (AXL) using



(a) Physical MAT Components (b) Flow Diagram of MAT

Figure 1. MAT unit components and its schematic [15,16]

different catalyst systems under controlled MAT conditions. This approach evaluates the impact of feed properties and catalyst characteristics on product distribution and coke deposition.

2. METHODOLOGY

2.1. MAT Unit. A fixed-bed MAT unit was employed to test catalyst with light feedstocks, using a quartz or stainless-steel tubular reactor (22 mm inner diameter and 380 mm in length). Figure 1 shows a the layout of the MAT setup and its components. An identical bench-scale reactor was used to experimentally examine the naphtha cracker feed. However, instead of standard iced water cooling, a low-temperature chiller that was kept at −10 °C was directly connected to the reactor jacket to chill the incoming feed. The −10 °C quenching temperature, emphasized by Altamira Instruments [13], was selected to rapidly condense volatile products, minimizing light component loss and ensuring accurate mass balance and yield analysis. This adjustment ensured more stable temperature control throughout the reaction. The receiving system was also modified to accommodate high product volatility. A wide-mouth 2 mL sample vial capped with a tightly fitting rubber hose was held at the bottom of the receiver to ensure a sealed, gas-tight connection of the two parts of the system to minimize losses and prevent contamination.

A 30 s reaction time was used in each experiment with the MAT unit. This aligns with the American Society for Testing & Materials Test Method ASTM D5154M-25 [14], which recommends short contact times to simulate FCC riser conditions and prevent excessive secondary cracking. The test commenced by first adding a fixed amount of the catalyst to the reactor. Subsequently, the system was flushed with nitrogen gas for 15 min before the feed was introduced. Once the liquid receiver and gas collection system were set up and the liquid receiver was cooled for the gas products to condense, a leak test was performed before the experiment began. The reactor was filled with 0.9 g of crude oil for 30 s. At the end of the reaction, the catalysts were stripped from hydrocarbons using nitrogen for 5 min.

The product gas was collected in a burette and analyzed using a gas chromatograph, and the liquid products were analyzed via simulated distillation. The exact amount of feed oil was measured from the weight difference of the feed micro-syringe before and after the MAT run. Table 1 presents the operating conditions of the MAT unit. The data obtained were compiled into an Excel sheet for further interpretation.

2.2. Analysis of Gas, Liquid Products, and Coke . The cracked output comprises gaseous products, liquid hydrocarbons, and solid coke. To assess the feedstock performance, a comprehensive gas chromatographic analysis was conducted on the products obtained from the MAT unit, providing detailed yield distributions.

2.2.1. Analysis of gaseous products. Gaseous components were analyzed using a Fusion® Inficon MicroGC unit having three thermal conductivity detectors , enabling precise quantification of light hydrocarbons up to C4, including C5 paraffins, hydrogen, and fixed gases. This system offers a rapid analysis with high sensitivity, making it suitable for refinery gas applications. The analysis accurately measured hydrocarbons ranging from C1 to C4 and C5 paraffins. The masses of each identified gas component were summed, and the masses of compounds heavier than C4 were combined with the liquid product weight to determine the overall product distribution. Comparative studies have demonstrated the effectiveness of such analytical approaches for

Table 1. *Operating conditions of the MAT.*

Activity	Operating Conditions
Injected feed weight	1 gram
Weight of catalyst weight	A multiple of the catalyst weight to achieve the required C/O ratio
Time-on-stream (time of feed injection)	30 s
Temperature of feed syringe	60 °C
Temperature of feed injector	5 °C higher than reaction temperature
Temperature of liquid receiver	-9 °C
Liquid/catalyst stripping time with receiver in cold bath with cold bath removed	Total 8 min 5 min 3 min

evaluating FCC product yields across different reactor configurations [17].

2.2.2. Analysis of liquid products. Unlike conventional physical distillation methods, gas chromatography with a flame ionization detector (FID) was used to simulate the distillation of the cracked liquid product, providing a rapid and precise analysis of boiling point distributions in petroleum fractions [18]. The liquid products contained three different fractions, namely, gasoline of naphtha (C_5-221 °C), heavy cycle oil (343+°C) and light cycle oil (221–343 °C).

2.2.3. Analysis of solid coke product. The amount of coke deposited on catalyst after the MAT run was examined using a Horiba carbon analyzer (Model EMIA-220V). One gram of the spent catalyst (with tungsten added as a combustion promoter) was combusted in a furnace at elevated temperature, at approximately 760 °C. The CO_2 resulting from the combustion was passed through an infrared analyzer, and the amount of carbon was estimated (as a percentage of catalyst weight).

Figure 2 shows the conventional FCC schematic and catalyst processes used for the analysis of the gas and liquid products and coke operations.

3. EXPERIMENT

The distillation and physical properties of crude oil feeds (AXL, AL, and AH) are shown in Table 2. For AXL, the simulated distillation results showed 40% gasoline, 26% middle distillate, and 33% heavy oil. Per structured multiphase guidelines, the experiment for catalyst optimization in the FCC unit was performed. First, a trial run was implemented with the present FCC catalyst for comparison to improve the operation and understand the bottlenecks of the unit. The choice of the catalyst was the first step in the physical and chemical analyses, followed by steam deactivation to reach the equilibrium state.

Tests in the MAT unit were performed by adjusting the steaming conditions and C/O ratios. Subsequently, the results were analyzed and screened using a kinetic model to establish a definitive economic practicality. Using the results of evaluating economics, the best-known solution was established, qualitative analysis was conducted, and references were checked. The post-choice comprised monitoring the change following a test run, ensuring that the correct computer model was in action, and collecting vital signs. Finally, a report was created that required knowledge of changes in performance and the impact of methodological measures.

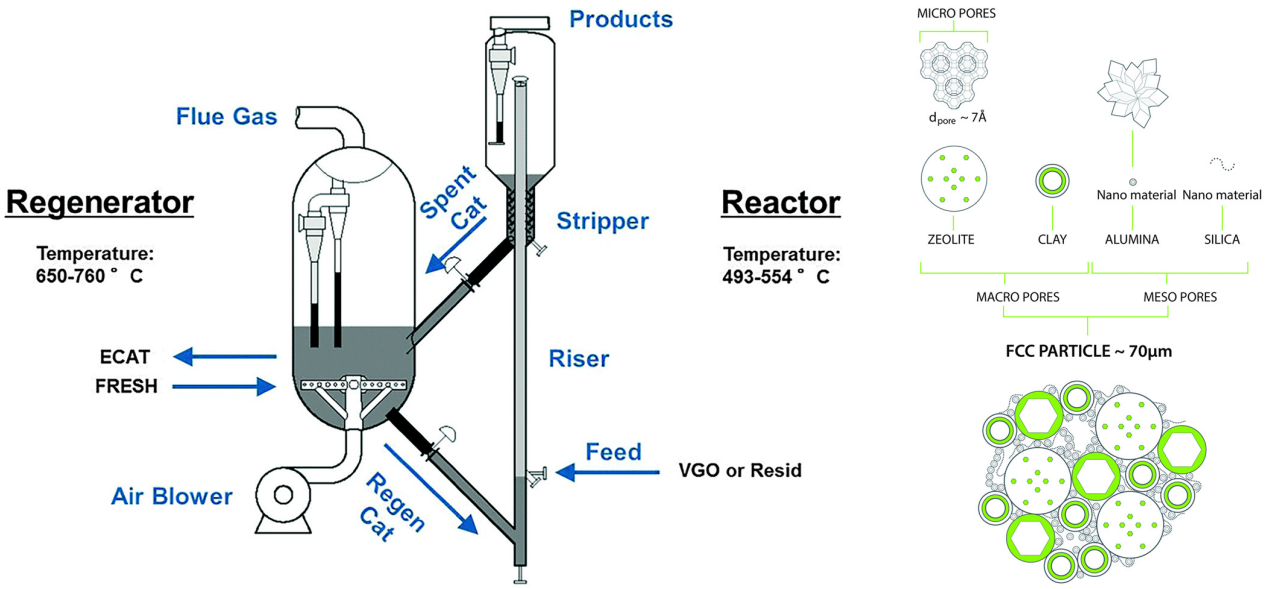


Figure 2. Conventional FCC Schematic and catalyst processes [12]

Table 2. *Physical and distillation properties of crude oil feeds.*

Property	AXL	AL	AH
API Gravity (degree)	39	34	21
Sulfur (wt. %)	1.6	2.3	2.9
Vanadium (ppm)	2.7	16	30
Nickel (ppm)	<1	3.3	9
Elemental composition (wt. %)			
Carbon	84.3	84.3	–
Hydrogen	12.6	12.2	–
Nitrogen	0.7	0.64	0.15
Simulated distillation-SimDist (°C)			
Initial boiling point	25	22	22
50%	287	307	350
Final boiling point	577	580	590
Distillation fractions (wt. %)			
Gasoline (C_5–221 °C)	41	35	20
Middle distillates (221–343 °C)	26	26	22
Heavy oil (343 °C+)	33	39	58

The general cracking reaction mechanism depends on the intrinsic zeolite acidic characteristics of the catalyst. The high-acidic and wide mesopores of the FCC catalyst facilitated the cracking of large hydrocarbon molecules (heavy oil) to light fractions, mainly light olefins and aromatics. The deposition of coke was more related to thermal and not catalytic cracking with the increase in temperature. The MAT tests provided catalytic activity and selectivity curves, which mean yield curves depending on conversion, defined as the sum of coke and total gas yields.

4. RESULTS AND DISCUSSION

Following the experimental procedures, the obtained data were compiled using Microsoft Excel for visualization and detailed analysis. The results were used to demonstrate the influence of different temperatures and catalyst parameters on the FCC performance of several crude oil feeds. The evaluation focused on determining significant trends, recognizing unexpected trends, and comparing results with the existing literature. This section discusses the efficiency of conversion, distribution of the product (vapor, liquid, and coke), and performance of the catalyst. These factors provide valuable information for identifying the most efficient method for increasing yield while minimizing byproducts.

4.1. Effects of Temperature and Catalyst Observed.

The experimental results showed that the FCC process is highly dependent on temperature and catalyst characteristics. The rise in temperature improves thermal cracking and catalytic activity, thereby enabling β-scission reactions that are in accordance with

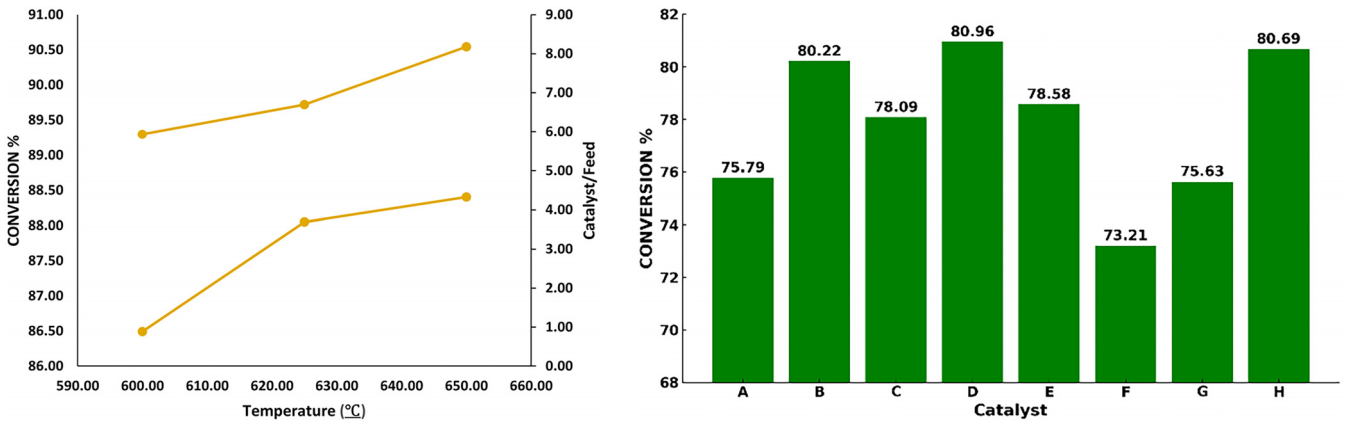


Figure 3. Effect of temperature and catalyst on overall conversion: (a) Conversion percentage of AXL with respect to temperature and C/O ratio; (b) Effects of different catalysts on the conversion percentage of AXL feed at 650 °C

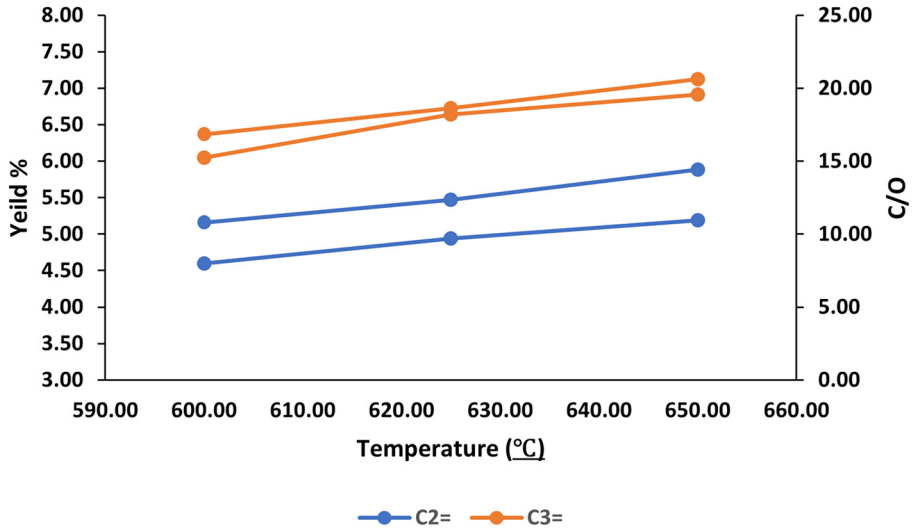


Figure 4. Yield percentage of C2 and C3 from AXL Feed and Catalyst with respect to temperature and C/O ratio

the formation of short-chain olefins, such as propylene and ethylene. Conversely, high temperatures lead to the over-cracking of intermediates, such as gasoline-range hydrocarbons, resulting in less liquid and more gas. This aligns with the endothermic characteristics of olefin synthesis and the thermodynamic compatibility of the gas-phase products at elevated temperatures.

Generally, an increase in both the temperature and C/O quotient is favorable for crude feed conversion. However, high values can spur overcracking, which can reduce the formation of usable products, such as gasoline, and increase light gas production. Gas products such as ethylene and propylene showed only a slight increase in their yields, indicating that only a small amount of gas was generated under the FCC standard conditions.

Liquid yields, particularly of gasoline, showed a more significant response and declined when the conditions were above the optimum level. Additionally, heavier and more aromatic feedstocks increased coke formation, particularly as temperature and catalyst activity were considered. Therefore, controlled and stable catalysts must be developed to achieve satisfactory results. Consequently, maximizing temperature and catalyst parameters in FCC operations helps maintain process efficiency, product selectivity, and catalyst stability.

4.1.1. Effects of temperature and catalyst on overall conversion. The overall conversion obtained from FCC of the oil feed is related to the temperature and amount of catalyst. Theoretically, increasing the temperature and C/O ratio would increase the overall conversion of the crude feed in the FCC. Figure 3(a) shows a 2% increase in the overall conversion with increasing the temperature by 50 °C and the amount of catalyst by 3 in the FCC of AXL oil feed. This indicates the significant impact of temperature and catalyst amount on the overall conversion of the crude feed in FCC. As the type of catalyst is essential for overall conversion, it must be carefully selected. Figure 3(b) shows the effect of different catalyst types on the AXL oil feed conversion under identical conditions. Catalysts

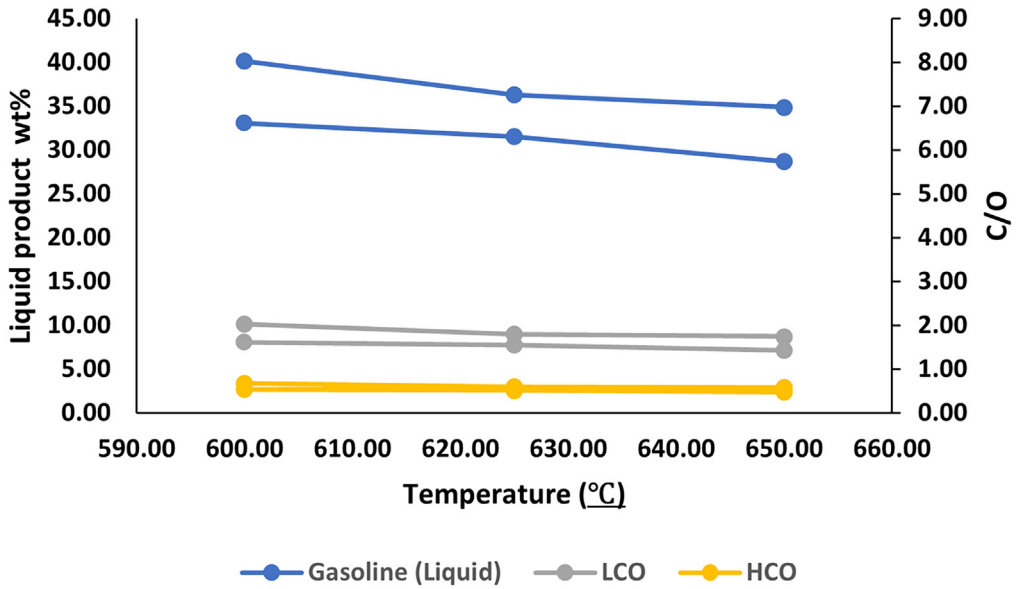


Figure 5. Amount of liquid products in wt% from AXL feed with catalyst with respect to temperature and C/O ratio

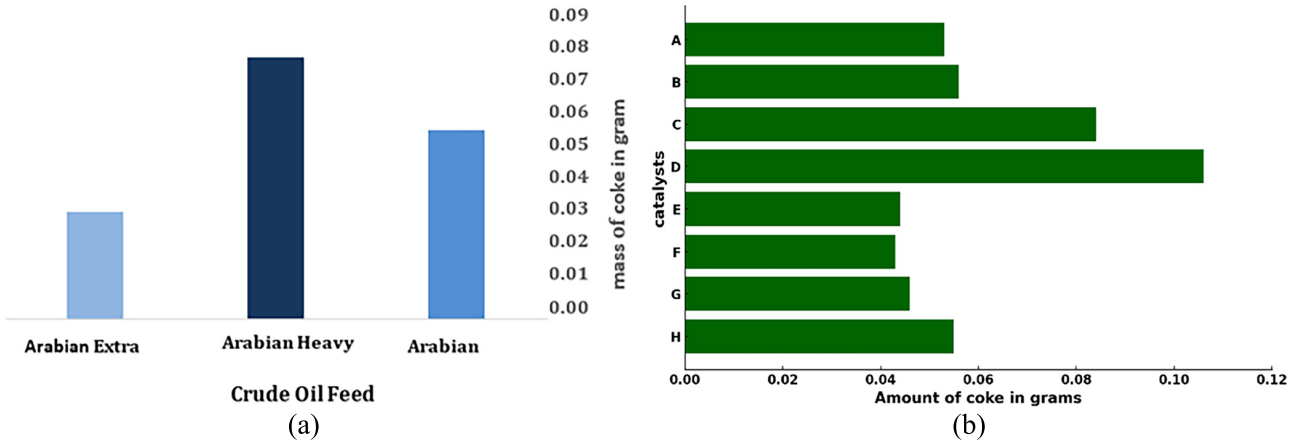


Figure 6. Effect of temperature and catalyst on coke: (a) Amount of coke in grams produced by different types of crude feeds; (b) Effect of different catalysts on the amount of coke produced from the FCC of AXL crude oil at 530 °C and 7 C/O ratio

A–H are proprietary formulations with undisclosed compositions, evaluated through a screening approach.

4.1.2. Effects of temperature and catalyst on gas products yield. Ethylene (C2=) and propylene (C3=) gases are two of the main products obtained from the FCC process and are essential in the production of various plastics and petrochemical products; their yields are also influenced by the temperature and catalyst amount. Additionally, optimizing these parameters can increase the efficiency and profitability of the production process. As shown in Figure 4, increasing the temperature by 50 °C and C/O by 1 would increase the yields of propylene and ethylene by 0.63% and 1%, respectively. This indicates the minor effects of the variables on the gas products because the FCC produces lower gas yields than liquid products and requires extremely high temperatures to overcome gas product condensation.

4.1.3. Effects of temperature and catalyst on liquid products yield. FCC generates three main types of liquid products: gasoline, low-boiling point light cycle oil (LCO), and high-boiling-point heavy cycle oil (HCO). Gasoline is particularly valuable, and most FCC units aim to maximize the conversion of their feeds to fuels and other liquefied petroleum gas (LPG). The increase in temperature and C/O ratio has a more significant impact on liquid products, such as gasoline, LCO, and HCO, compared to gaseous products, such as ethylene and propylene. FCC produces significant amounts of gasoline and other distillates. However, excessive conversion may reduce gasoline yield and increase LPG generation[19,20]. Thus, the gasoline yield has an inverse relationship with temperature and C/O; reducing the temperature would increase the gasoline yield (Figure 5).

4.1.4. Effects of temperature and catalyst on coke. The FCC process produces a solid product known as coke, which is combusted to CO_2 during the regeneration cycle, and the heat generated is used to heat the catalyst and feed undesirable byproducts of the condensation of hydrocarbon vapors on catalyst surface during cracking. This deactivates the catalyst and reduces efficiency. To restore activity, coke is burned off the catalyst through regeneration. These observations indicate that feedstocks with higher boiling points and greater aromaticity, such as crude feed, yield larger amounts of coke[21]. This relationship is exemplified in Figure 6(a), which shows that AH produces a higher amount of coke than AL or AXL crude oil feeds. This is because heavier feeds such as AH have larger quantities of polyaromatic hydrocarbons and compounds with high boiling temperatures, which are more susceptible to polycondensation and carbon deposition. These components are resistant to catalytic cracking and form refractory species that generate coke upon decomposition. Additionally, the elevated metal and asphaltene contents in AH contribute to catalyst fouling and further accelerate coke formation during cracking.

As shown in Figure 6(b), catalyst D produces the highest coke yield for the AXL crude feed due to its higher conversion. As conversion depends on both temperature and C/O ratio, coke formation also correlates with these parameters. This relationship aligns with recent studies showing that increased C/O ratios enhance conversion and coke yield, as greater catalyst availability promotes secondary cracking and condensation reactions that deposit carbonaceous species on the catalyst surface. Specifically, at C/O ratios between 1 and 7 and temperatures of 510–550 °C, coke yields increased sharply from 1.3% to 6.3% in parallel with conversion gains[22]. Thus, the FCC conditions require balancing the desired conversion with manageable coke levels, guided by C/O and operating temperature strategies.

5. CONCLUSIONS

This study explored the impact of various process conditions , such as reaction temperature, catalyst type, and crude oil feed, on the performance of FCC using crude oil feeds. It analyzed product yields, namely gas, liquid, and coke, under set-up conditions to identify the most suitable conversion routes. Based on the results of MAT experiments, the following conclusions were drawn:

- High temperatures and C/O ratios lead to faster conversion rates and light olefin (C2=, C3=) formation.
- The gasoline yield is highest at moderate temperatures and catalyst conditions, whereas extremely high values cause overcracking, and LPG becomes the dominant component.
- Coke formation mostly occurs during feed conversion, primarily for heavier feedstocks, such as AH crude.

- The choice of catalyst type has the most significant effect on oil conversion and coke deposition , indicating the crucial role of tailored catalyst use.
- Pretreatment methods such as hydrotreating and demetallization are crucial for reducing catalyst deactivation and coke formation, thereby improving the overall process efficiency.

The results align with current attempts to improve FCC operations and produce both fuels and chemicals. Catalyst selection can be performed through formulations that limit coke precursors, particularly if the feed is heavy. Future studies may validate these results at pilot and industrial scales, examining the effects of different feed pretreatment methods (such as hydroprocessing) and testing the efficiency of catalyst regeneration for high coke loads.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Osama Wael Aljohani – *Department of Chemical Engineering, King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, Saudi Arabia; Phone: +966542425403;* 0009-0006-3741-7864
Email: s202168630@kfupm.edu.sa

Corresponding Author

Abdullah M Aitani – *Research Mentor, Interdisciplinary Research Center for Refining & Advanced Chemicals, King Fahd University of Petroleum & Minerals (KFUPM), Saudi Arabia;* 0000-0001-5071-4034
Email: maitani@kfupm.edu.sa

ACKNOWLEDGEMENTS

I express my deepest gratitude to my supervisor, Dr. Abdullah Aitani, for his guidance, support, and valuable insights throughout this research. I am also grateful to the Center for Refining and Advanced Chemicals at King Fahd University of Petroleum and Minerals (KFUPM) for providing access to the resources and equipment that supported the successful completion of this research. I also appreciate the support provided by the Undergraduate Research Office (URO) under the Uxplore program, which played a crucial role in facilitating this study.

CONFLICTS OF INTEREST

The authors report no conflicts of interest.

FINANCIAL DISCLOSURE

The support provided by KFUPM Center for Refining and Advanced Chemicals is appreciated.

REFERENCES

(1) Zhao, Z., Guo, Z., He, Y. & Wang, Y. Recent advances in fluid catalytic cracking for olefin production: Catalyst design and reaction pathway optimization. *Fuel Process. Technol.* **246**, 107039 (2023).

(2) Al-Khattaf, S., Saeed, M. R., Aitani, A. & Klein, M. T. Catalytic cracking of light crude oil to light olefins and naphtha over E-Cat and MFI: Microactivity test versus advanced cracking evaluation and the effect of high reaction temperature. *Energy Fuels* **32**, 6189–6199 (2018).

(3) Jermy, B. R. *et al.* Crude oil conversion to chemicals over green synthesized ZSM-5 zeolite. *Fuel Process. Technol.* **241**, 107610 (2023).

(4) Qureshi, Z. S. *et al.* Steam catalytic cracking of crude oil over novel hierarchical zeolite–containing mesoporous silica–alumina core–shell catalysts. *J. Anal. Appl. Pyrolysis* **166**, 105621 (2022).

(5) Tanimu, A., Tanimu, G., Alasiri, H. & Aitani, A. Catalytic cracking of crude oil: Mini review of catalyst formulations for enhanced selectivity to light olefins. *Energy Fuels* **36**, 5152–5166 (2022).

(6) Chen, Y., Duan, H., Li, Q. & Wei, X. Crude-to-chemicals via catalytic cracking: Advances and industrial insights. *Chem. Eng. J.* **453**, 139820 (2023).

(7) Wang, L., Yang, Y., Wu, S. & Li, C. Catalytic strategies to improve light olefin yield from direct crude oil cracking. *J. Anal. Appl. Pyrolysis* **159**, 105413 (2022).

(8) Rahimi, N. & Karimzadeh, R. Catalytic cracking of hydrocarbons over modified zeolites: A review. *Appl. Catal. A Gen.* **398**, 1–17 (2011).

(9) Corma, A., Martínez, A., Martínez, C. & Sauvanaud, L. Design and operation of an FCC unit with high propylene yield. *J. Catal.* **240**, 318–324 (2006).

(10) Ancheyta, J., Sánchez, S. & Rodríguez, M. A. Kinetic modeling of hydrocracking of heavy oil fractions: A review. *Catal. Today* **109**, 76–92 (2005).

(11) Bjørgen, M., Joensen, F., Lillerud, K. P., Olsbye, U. & Svelle, S. The mechanism of coke formation in catalytic cracking. *J. Catal.* **249**, 195–207 (2007).

(12) Vogt, E. T. C. & Weckhuysen, B. M. Fluid catalytic cracking: recent developments on the grand old lady of zeolite catalysis. *Chem. Soc. Rev.* **44**, 7342–7370 (2015).

(13) Altamira Instruments. Evaluation of FCC catalysts using a microactivity test (MAT) reactor. Technical Note, *Altamira Instruments.* Number **24,** 1–3 (2021).

(14) ASTM International. *ASTM D5154/D5154M–25* Standard test method for determining activity and selectivity of fluid catalytic cracking (FCC) catalysts by microactivity test (2025).

(15) Akah, A., Al-Ghrami, M., Saeed, M. & Siddiqui, M. A. B. Reactivity of naphtha fractions for light olefins production. *Int. J. Ind. Chem.* **8**, 221–233 (2017).

(16) ASTM International. ASTM D 3907-19, Standard test method for testing fluid catalytic cracking catalysts (2019).

(17) Meng, X., *et al.* A Review on Production of Light Olefins via Fluid Catalytic Cracking. *Energies* **14**, 10–89 (2021).

(18) Marques Jorge, J. Getting Better Value from Petroleum Products using Simulated Distillation Gas Chromatography. *Petro Industry News* (2020).

(19) Moorehead, E., McLean, J. & Cronkright, W. Microactivity evaluation of FCC catalysts in the laboratory: Principles, approaches and applications. *Stud. Surf. Sci. Catal.* **76**, 223–255 (1993).

(20) Speight, J. G. Catalytic cracking. In *The Refinery of the Future*, 181–208 (Elsevier, 2011).

(21) Sawarkar, A. N., Pandit, A. B., Samant, S. D. & Joshi, J. B. Petroleum residue upgrading via delayed coking: A review. *Can. J. Chem. Eng.* **85**, 1–24 (2007).

(22) Trantham, K. et al. Unravelling vacuum gas oil catalytic cracking: The influence of the catalyst-to-oil ratio on FCC catalyst performance. *Catalysts* **15**, 170 (2023).

KFUPM
Journal of Undergraduate Research International

Diamond Open Access 35

Article Computer Engineering/Computer Science

Predicting Smoking Status with Graph Neural Networks and Transformers: A Data-Driven Approach

Sk. Md Abir Hasan Imran[1], Arupa Barua[2] and Md. Osama[3*]

Cite https://doi.org/10.64589/juri/209730

Submitted: June 05, 2025 **Revised:** July 19, 2025 **Accepted:** August 20, 2025

ABSTRACT

Smoking remains a major global public health issue, contributing to millions of preventable deaths annually and placing a significant burden on healthcare systems, particularly in low- and middle-income countries. Despite widespread awareness of its harmful effects, tobacco use continues to increase in certain populations, regardless of level of education. Traditional self-report methods for identifying smokers often suffer from inaccuracies; therefore, there is need for reliable data-driven alternatives. In this study, we explored a predictive modeling approach to classify individuals as smokers or nonsmokers based on a range of demographic, behavioral, and psychological factors. A custom dataset was developed comprising 223 instances and 17 features related to personal background and smoking-related influences. To address the complexity of feature interactions, we propose a hybrid deep-learning architecture that integrates a (GNN) and a feature-tokenizer-based transformer. This model leveraged both relational and contextual information to improve the identification of smoking patterns. The findings highlight the potential of advanced machine learning methods to support early intervention strategies and enhance public health planning.

Keywords: smoking detection, public health, Graph Neural Network (GNN), transformer encoder, deep learning

1. INTRODUCTION

Smoking remains a significant global public health challenge that contributes to a wide range of preventable diseases and premature deaths. Tobacco use causes more than 8-million premature deaths each year: over 7-million deaths because of direct use and 1.3-million caused by second-hand smoke exposure [1]. Alarmingly, even educated individuals aware of these severe health risks continue to smoke. In addition to its devastating health consequences [2], smoking imposes a heavy economic burden on healthcare systems worldwide. Low- and middle-income countries are especially vulnerable and are experiencing an increase in smoking-related illnesses. To address this issue, the factors that lead individuals to smoke must be understood, and strategies for prevention and cessation should be developed [3]. However, traditional self-reporting methods for identifying smokers often lack accuracy because individuals may underreport or misrepresent their smoking habits. To overcome these limitations, researchers have focused on data-driven solutions and emerging technologies [4]. Mobile health (mHealth) applications and wearable sensors have been explored to support smokers in quitting smoking by automatically detecting smoking behavior through patterns such as hand-to-mouth gestures. These solutions assist in smoking reduction or cessation programs and include systems for automatic smoking detection and smoking cessation support. Automatic smoking detection systems use technology to infer the number of cigarettes smoked with minimal user intervention, unlike approaches that rely on self-reporting. Wearable devices such as armbands or smartwatches, which can interact with smartphone apps, are used to detect smoking events by analyzing movements such as hand-to-mouth motion [4].

Additionally, deep-learning (DL)-based computer vision methods are being explored for smoker classification and detection; these are particularly useful for monitoring cigarette use in designated non-smoking indoor and outdoor environments. Vision-based methods can recognize smokers from a distance, although challenges remain in smoke detection against certain backgrounds and in identifying cigarettes because of their size [5]. Furthermore, machine learning (ML) and DL technologies have been used to develop predictive models for identifying potential smokers and classifying smoking status based on individual health metrics. Analyzing modern healthcare data, including parameters such as blood pressure, cholesterol levels, and hemoglobin concentrations, provides insight into the associations between these metrics and smoking behavior [6]. Identifying significant relationships enables the tailoring of interventions and targeting of high-risk individuals. The development of robust models for the automatic classification of smoking status based on relevant features can significantly enhance early intervention efforts and inform public health strategies. In this study, we focused on classifying individuals as smokers or nonsmokers using a structured dataset containing demographic, behavioral, and health-related attributes collected through a survey. The goal

Table 1. *Literature summary*

Author	Objective	Feature	Data Type	Method
Khan et al. (2025)	Review and evaluate deep learning models for smoker classification and cigarette detection.	YOLOv9, YOLO11, CNN/transformer features	Image datasets (CigDet)	Deep learning (CNNs, transformers), YOLO benchmarking
Ali et al. (2024)	Classify smokers/non-smokers using demographic and health metrics	BP, cholesterol, hemoglobin, etc.	Tabular health data (1338 and 38,984 samples)	Logistic regression, extra trees, random forest
Ortis et al. (2020)	Review wearable/mobile tech for smoking detection and cessation	IMU sensors, app-based metadata	Sensor signals (accelerometer, gyroscope), app logs	Literature review, empirical analysis
SmokeSift – Zameer et al.	Develop ML model for real-time classification on Raspberry Pi	27 features: age, BP, BMI, blood data	Clinical dataset (55,692 instances)	Random forest (best), XGBoost, AdaBoost, LDA, GUI integration
Senyurek et al. (2020)	Detect smoking gestures using wearable IMU sensors	Hand-to-mouth gesture regularity	IMU time-series data from 35 smokers	Signal processing, autocorrelation, threshold-based binary classification
Jarvis et al.	Compare biomarker-based and self-reported smoker identification	Cotinine, CO, nicotine, thiocyanate	Biochemical samples (211 patients)	Sensitivity/specificity comparisons across biomarker thresholds
This study (2025)	Classify smokers vs. non-smokers using demographic, behavioral, and psychological features	Age, gender, education, income, mental health, peer/family influence, etc.	Survey dataset (223 responses, 17 features)	ML classification achieving 93% accuracy

of this study was to develop an accurate predictive model that could aid in early detection and inform targeted public-health interventions. Our key contributions are as follows:

- We developed a custom dataset consisting of 223 samples and 17 features covering personal, social, and behavioral indicators related to smoking habits.
- We built a ML model capable of classifying individuals as smokers or nonsmokers, achieving an accuracy of 93%.

2. RELATED WORK

Several studies have investigated smoking detection and classification using various techniques[7], including behavioral analysis, physiological data, and ML algorithms. Early research has highlighted the significant influence of family smoking habits on adolescents, suggesting that children with parents or siblings who smoke are more likely to adopt this habit[8,9]. Similarly, peer influence is a powerful factor: Ali and Bradley[10], reported that having a close friend who smokes greatly increases the likelihood of smoking. Mental health has also emerged as a key variable. Fluharty et al.[11] and Lemstra et al.[3] found a strong bidirectional link between smoking and conditions such as anxiety and depression, underlining the importance of psychological variables in predictive models. Carroll et al.[1] emphasized the protective role of physical activity because it is often inversely associated with smoking behavior.

Recent advances in ML have led to the development of real-time smoker-classification tools. The SmokeSift system[12] uses *random forest* classifiers deployed on embedded systems to perform real-time classification. Ali et al.[13] employed various classifiers such as *logistic regression* and *extra trees* on tabular health datasets, similar to the approach used in this study. Other researchers have explored image-based and sensor-based methods that use convolutional neural networks (CNNs), transformers, and wearable device data for gesture recognition and cigarette detection. Jarvis et al.[14] compared biomarker-based and self-reported smoker identification using 211 biochemical samples, and evaluated cotinine, CO, nicotine, and thiocyanate using sensitivity and specificity thresholds. Table 1 summarizes the key studies in the field.

3. METHODOLOGY

3.1. Dataset Description. The dataset was specifically curated to analyze and predict smoking behavior based on sociodemographic, psychological, and behavioral factors. Data were collected via a structured Google Form from a diverse group of individuals, including university students, working professionals, and general community members, to ensure varied representation. The dataset comprised 223 valid individual records, each with a combination of numerical, categorical, and binary attributes. These features were chosen based on prior research indicating their relevance to smoking initiation and maintenance. Table 2 lists the features of this dataset.

Table 2. *Feature description*

FEATURE	TYPE	DESCRIPTION
AGE	Numerical	Current age of the respondent.
GENDER	Categorical	Biological sex: male or female.
MARITAL STATUS	Categorical	Marital status, such as single, married, etc.
EDUCATION LEVEL	Categorical	Highest degree or academic level achieved.
OCCUPATION	Categorical	Employment type (e.g., student, employed, unemployed).
MONTHLY INCOME	Numerical	Approximate monthly income in Bangladeshi taka (BDT).
FAMILY SMOKING HISTORY	Binary	Indicates whether any family member smokes (yes/no).
PEER SMOKING INFLUENCE	Ordinal	Degree of smoking among peers: none, Few, most, or all.
MENTAL HEALTH STATUS	Categorical	Self-reported mental health status over recent months.
EXERCISE FREQUENCY	Categorical	Physical activity frequency: Regular, occasional, or never.
PRESSURE TO SMOKE	Binary	Has the respondent has been pressured to smoke.
REASON FOR SMOKING	Text (for smokers)	Open-ended response on why the respondent smokes.
AGE STARTED SMOKING	Numerical (for smokers)	Age of smoking initiation.
TARGET VARIABLE	Binary	"Do you smoke?" (yes = 1, no = 0).

Table 3. *Comparison of DC Motor types in EV applications*

Feature	Mean/Mode	Std/Frequency
Age	23.8 years	SD = 4.5
Monthly income (BDT)	4066	SD = 12,126
Gender	Male (62.3%)	Female (37.7%)
Family smoking history	Yes (54.7%)	No (45.3%)
Peer smoking influence	Few (38.2%)	None (26.5%), most (21%), all (14.3%)
Mental health	Good (41.2%)	Fair (35.6%), poor (23.2%)
Exercise frequency	Occasional (46%)	Regular (28%), never (26%)
Pressure to smoke	Yes (31%)	No (69%)
Smokers (target=1)	79 respondents	35.4% of total

Table 3 summarizes the key variables in the dataset, presenting central tendencies such as the mean age and monthly income, as well as the frequencies of categorical attributes such as sex, family smoking history, and mental health status. The proportion of smokers in the sample is also shown. Figure 1 shows the distribution of the students' monthly incomes. Because most students had no income and only a few earned modest amounts, the distribution was highly right-skewed, as indicated by the skewness score of 4.39.

Together, these statistics provided a comprehensive overview of the respondents' characteristics and the prevalence of smoking-related factors in the study population.

3.2. Model Development. The proposed model follows a hybrid deep learning pipeline that integrates graph neural networks (GNNs) with transformer encoders to classify smoking behavior. This architecture was designed to capture both relational interactions between features and complex dependencies within high-dimensional data representations. The model processes a combination of categorical and continuous features, embeds them into a shared latent space, and learns local and global interactions through a sequence of neural modules. The overall pipeline was optimized for tabular data

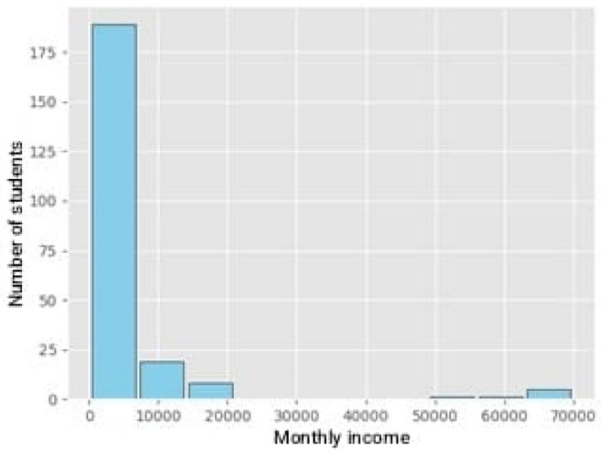


Figure 1. Distribution of students' monthly income

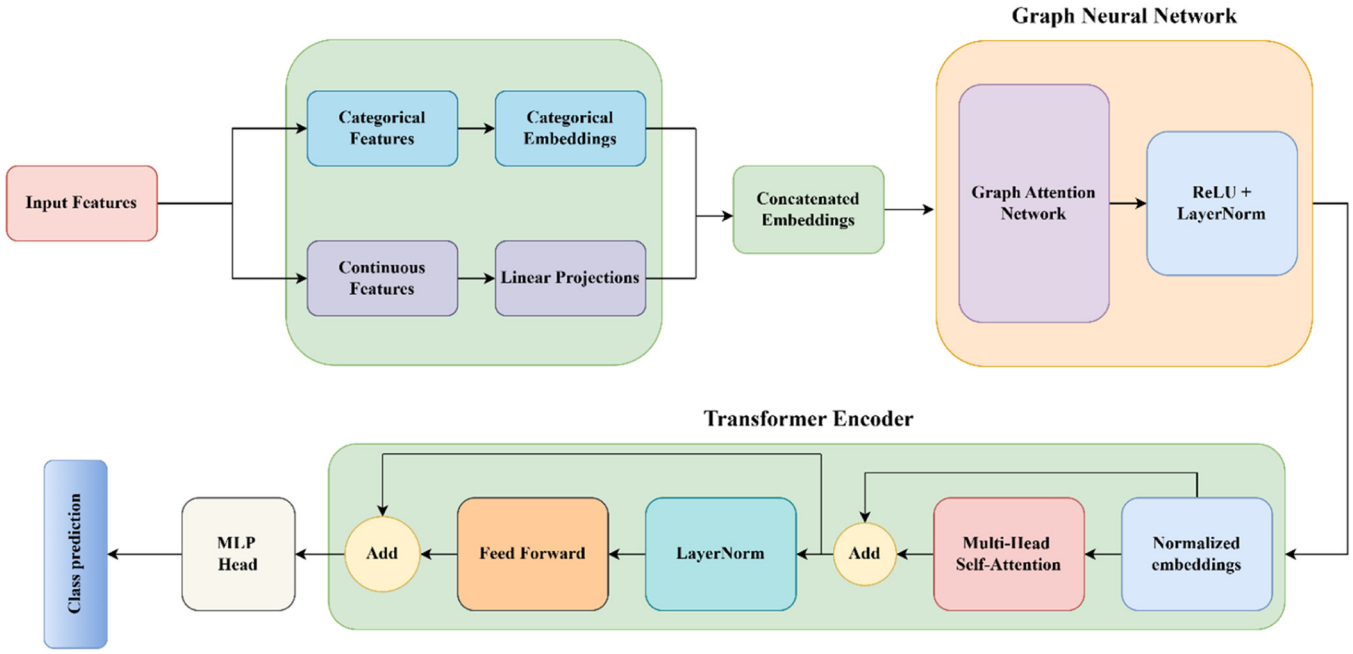


Figure 2. Proposed methodology

with heterogeneous feature types and aimed to enhance both feature representation and decision accuracy through joint attention mechanisms. Figure 2 shows the overall development process schematically.

3.2.1. Input Features. The model begins by separating raw input into two main categories: categorical and continuous. The categorical features included variables such as sex, marital status, education level, and peer influence, whereas the continuous features included age, monthly income, and age. This distinction is important because these feature types require different treatments during preprocessing and embedding. By processing them independently at the beginning, the model ensures an optimal representation before combining them for downstream tasks.

3.2.2. Embedding. For categorical features, embedding layers were used to transform each discrete value into a dense continuous vector representation. This enabled the model to learn the similarities between categories such as different levels of peer-smoking influence. The continuous features were processed through linear projection layers that scale and transform them into the same latent space as the categorical embeddings. When both types are encoded, they are concatenated into a single unified vector that represents the entire feature set of an individual instance. This embedding process ensures that all features, regardless of their original format, are compatible with the neural computations.

3.2.3. Graph Neural Network . In our model, each input instance is represented as a fully connected graph in which each

Table 4. *Performance comparison on test set*

Model Type	Model	Precision	Recall	Macro-F1	Accuracy
ML	Logistic regression	0.77	0.61	0.68	0.78
	Decision tree	0.68	0.81	0.73	0.80
	Random forest	0.76	0.79	0.77	0.82
	Multinomial NB	0.81	0.74	0.77	0.84
	SVM	0.73	0.58	0.58	0.78
DL	LSTM	0.84	0.71	0.75	0.84
	GRU	0.78	0.70	0.72	0.82
	RNN	0.78	0.62	0.64	0.80
	Bi-LSTM	0.82	0.82	0.82	0.87
	Bi-GRU	0.86	0.83	0.84	0.89
Proposed	**GNN+FT-transformer**	**0.89**	**0.96**	**0.92**	**0.93**

Table 5. *Performance comparison on five-fold cross validation*

Seed	Model	Precision	Recall	Macro-F1	Accuracy
0	GNN	0.8538 ±0.0775	0.8709 ±0.1684	0.8564 ±0.1060	0.9328 ±0.0463
	FT-transformer	0.8713 ±0.1267	0.9073 ±0.1237	0.8806 ±0.0752	0.9416 ±0.0384
	GNN + FT-transformer	0.8909 ±0.0452	0.9255 ±0.0465	0.9074 ±0.0375	0.9552 ±0.0179
42	GNN	0.9183 ±0.0949	0.7927 ±0.1488	0.8466 ±0.1037	0.9329 ±0.0459
	FT-transformer	0.9088 ±0.1073	0.8145 ±0.1393	0.8519 ±0.0880	0.9330 ±0.0387
	GNN + FT-transformer	0.8841 ±0.0969	0.8909 ±0.1659	0.8771 ±0.0845	0.9420 ±0.0371
123	GNN	0.8629 ±0.1019	0.9255 ±0.0852	0.8908 ±0.0768	0.9461 ±0.0378
	FT-transformer	0.8347 ±0.1376	0.8455 ±0.1265	0.8376 ±0.1171	0.9233 ±0.0528
	GNN + FT-transformer	0.8614 ±0.1487	0.9255 ±0.0852	0.8856 ±0.0870	0.9415 ±0.0495

node corresponds to an individual feature (categorical or continuous). The graph structure was constructed manually and uniformly for all instances. We defined a fully connected topology between the feature nodes, meaning that every feature interacts with every other feature, including itself. This was implemented using a fixed complete-edge index that was repeated for each batch during training. The concatenated feature embeddings obtained from the learned categorical embeddings and a projection of continuous variables served as the node representations. These are passed to a graph attention network (GAT) layer that computes the attention scores between all feature pairs. This mechanism allows the model to weigh the feature interactions of the tabular data dynamically based on their learned relevance in the prediction task. For example, it can prioritize the relationship between mental health and family smoking history over less-informative feature combinations. The output of the GAT layer undergoes rectifier (ReLU) activation followed by layer normalization to promote training stability and ensure consistent feature scaling.

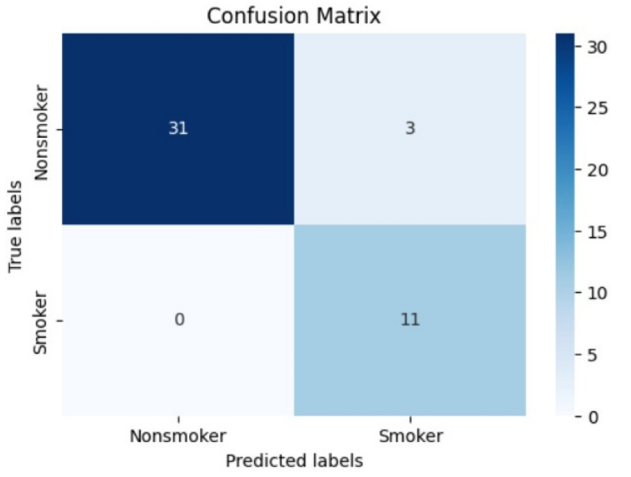


Figure 3. Confusion matrix of GNN + FT-transformer on test set

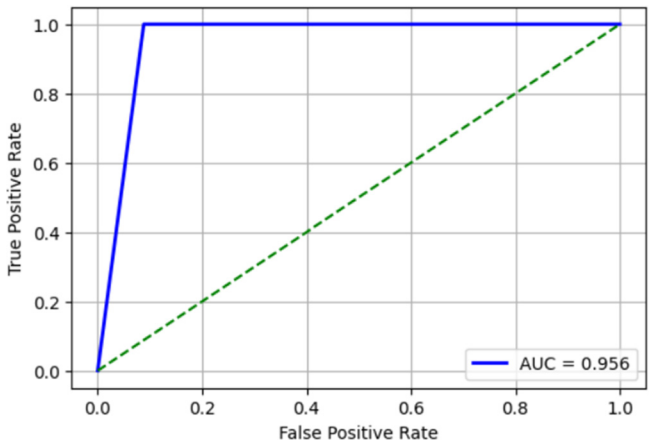


Figure 4. ROC curve of GNN + FT-Transformer on the test set

3.2.4. Transformer Encoder. After GNN processing, the normalized feature embeddings were input into a transformer encoder composed of four stacked blocks. Each block includes residual connections, layer normalization, and a feedforward network. The purpose of this encoder is to refine feature representations further by capturing higher-order dependencies and patterns that may not be apparent using simple attention or projection methods. The transformer depth enables the model to develop a more detailed understanding of how different features interact across multiple levels of abstraction.

3.2.5. Multihead Self-Attention. The core of each transformer block is a multihead self-attention mechanism. This module allows the model to attend to different feature subsets simultaneously and learn various types of relationships in parallel. Each head processes a separate projection of the input embeddings and their outputs are concatenated and linearly transformed. This enables the model to capture diverse perspectives on the data; for example, one attention head may learn about the relationships between demographic factors, whereas another may focus on behavioral aspects. Layered attention ensures robust feature learning, particularly for datasets with complex interactions.

4. RESULTS

To benchmark the performance of the proposed hybrid GNN + feature tokenizer (FT) + FT transformer model, we conducted a series of experiments using traditional ML classifiers, DL architectures, and the proposed method. The experiments were aimed

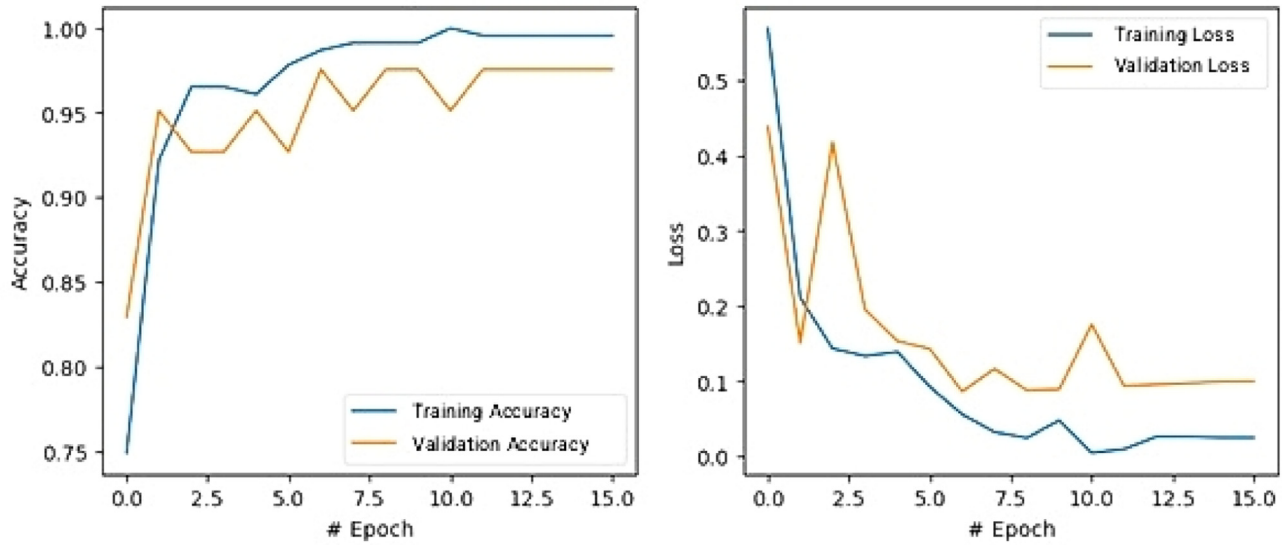


Figure 5. Accuracy and loss curve of GNN + FT-transformer on the test set

at validating the effectiveness of combining graph-based attention with transformer encoders for complex tabular classification.

4.1. Experimental setup. The dataset comprising 223 valid instances was split into training and testing sets in an 80:20 ratio. Stratified sampling was used to ensure a balanced distribution of smokers and nonsmokers in both sets. After splitting the data into training and testing sets, the training set consisted of 136 samples of class 0 and 42 samples of class 1, indicating class imbalance. To address this, a synthetic minority oversampling technique for nominal and continuous features was applied, resulting in a balanced training set with 136 samples for each class. Subsequently, 15% of the augmented training data were reserved for validation during the model training process.

4.2. Experiments.

4.2.1. ML Models. ML models serve as foundational baselines for smoking classification tasks. *Logistic regression* offers a simple linear approach, whereas *decision tree* and *random forest* introduce nonlinearity and feature interaction handling through tree structures. By contrast, multinomial *naïve Bayes* utilizes probabilistic assumptions, making it well-suited for datasets with categorical dominance. *Support vector machine* (SVM), although powerful in high-dimensional spaces, showed limitations with the size and nature of the data. These models were relatively fast to train and interpret but lacked the representational depth required to exploit the complex dependencies between features fully.

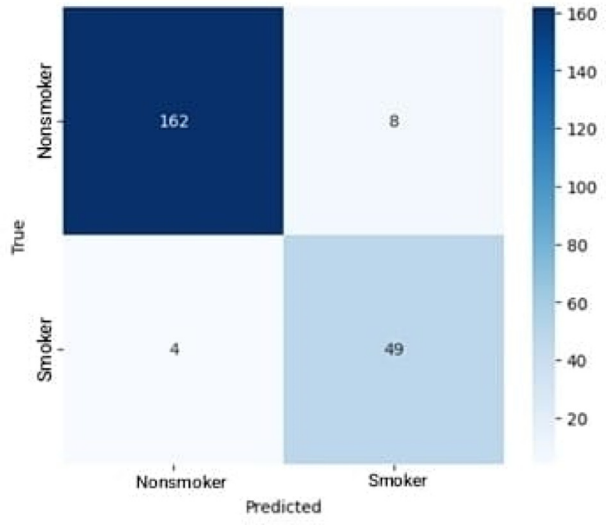


Figure 6. Confusion matrix of the GNN (seed 123)

4.2.2. DL Models. Several DL architectures have been explored to capture temporal and sequential patterns among features. Recurrent models such as *long short-term memory* (LSTM), *gated recurrent unit* (GRU), and *vanilla recurrent neural network* (RNN) have been employed because of their capacity to model dependencies across sequences of embedded features. *Bidirectional variants* (Bi-LSTM and Bi-GRU) enhance this capability by processing input from both forward and backward directions, allowing the models to learn context-aware representations better. These models demonstrate significant improvements over classical ML models, particularly in terms of recall and F1 scores, owing to their ability to learn complex feature relationships over multiple training epochs.

4.2.3. Proposed Architecture. The proposed model integrates a GNN with a FT-based transformer to capture both relational and contextual dependencies in tabular data. The GNN component identifies interfeature interactions by modeling features as graph nodes with attention-based connections, whereas the transformer encoder refines these representations through multihead self-attention mechanisms. This hybrid approach leverages both the structure- and sequence-aware modeling capabilities, enabling the network to learn more expressive and robust feature embeddings. The proposed model outperformed all baselines in terms of precision, recall, F1 score, and accuracy, confirming its superior ability to model complex behavioral patterns.

4.3. Evaluation.

4.3.1. Metrics. The performance of all models was evaluated using four metrics.

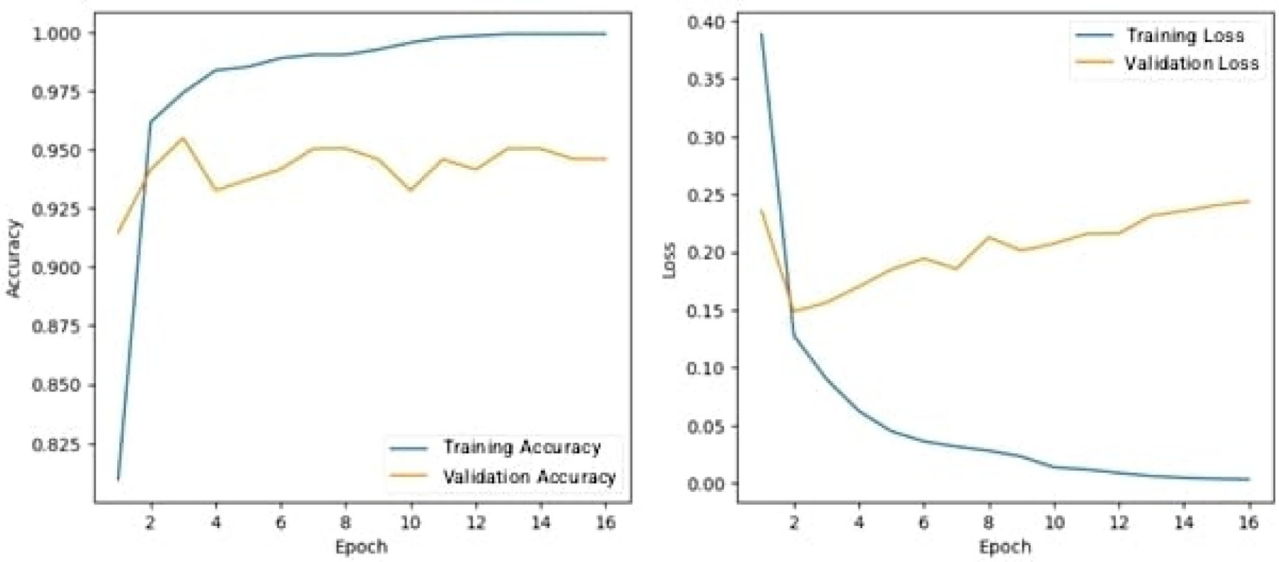


Figure 7. Accuracy and loss curve of GNN (seed 123)

1. **Precision**: The ratio of true positives to predicted positives.

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

Here, TP = true positives and FP = false positives.

2. **Recall**: The ratio of true positives to actual positives.

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

Here, FN = false negatives.

3. **Macro-F1 Score**: Harmonic mean of precision and recall across both classes.

$$Macro\ F1\ Score = 2 \times \frac{Precision + Recall}{Precision \times Recall}$$

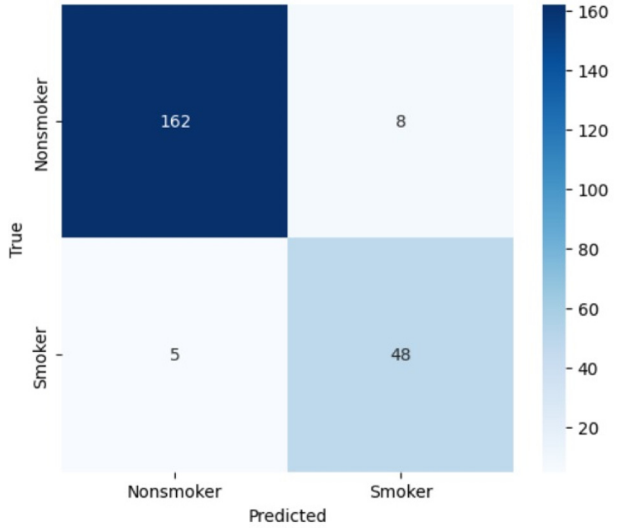


Figure 8. Confusion matrix of FT-transformer (seed 0)

For macro averaging, the F1 Score was computed independently for each class and then averaged.

$$Macro\ F1 = \frac{1}{N}\sum_{i=1}^{N} F1_i$$

Here, N = number of classes and $F1_i$ = *Macro F1* score for class i

4. **Accuracy**: The overall proportion of correct predictions.

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

4.3.2. Model Comparison (Train-Test Split). Table 4 presents the overall evaluation results of the test sets of the explored models.

As shown, the proposed GNN + FT transformer outperformed all other models across all evaluation metrics. The Bi-GRU and Bi-LSTM models also achieved high scores but still outperformed in both precision and recall. Traditional ML models, such as *random forest* and *naïve Bayes*, performed moderately well but could not fully capture complex feature interactions.

4.3.3. K-Fold Cross Validation and Ablation Study. To assess the generalization capability of the proposed model more reliably, we applied five-fold cross-validation. This technique helps to mitigate overfitting and provides a better estimate of the model performance on unseen data by ensuring that each data point is in the validation set. To enhance the reliability of our results and account for randomness in model initialization and data splitting further, we repeated the entire five-fold cross-validation procedure using three random seeds: 0, 42, and 123. This setup ensures that our performance metrics are not dependent on a single data split and allows for a more robust and statistically sound evaluation. For each seed, we computed the mean performance across five folds, and reported the overall mean and 95% confidence intervals across all three seeds.

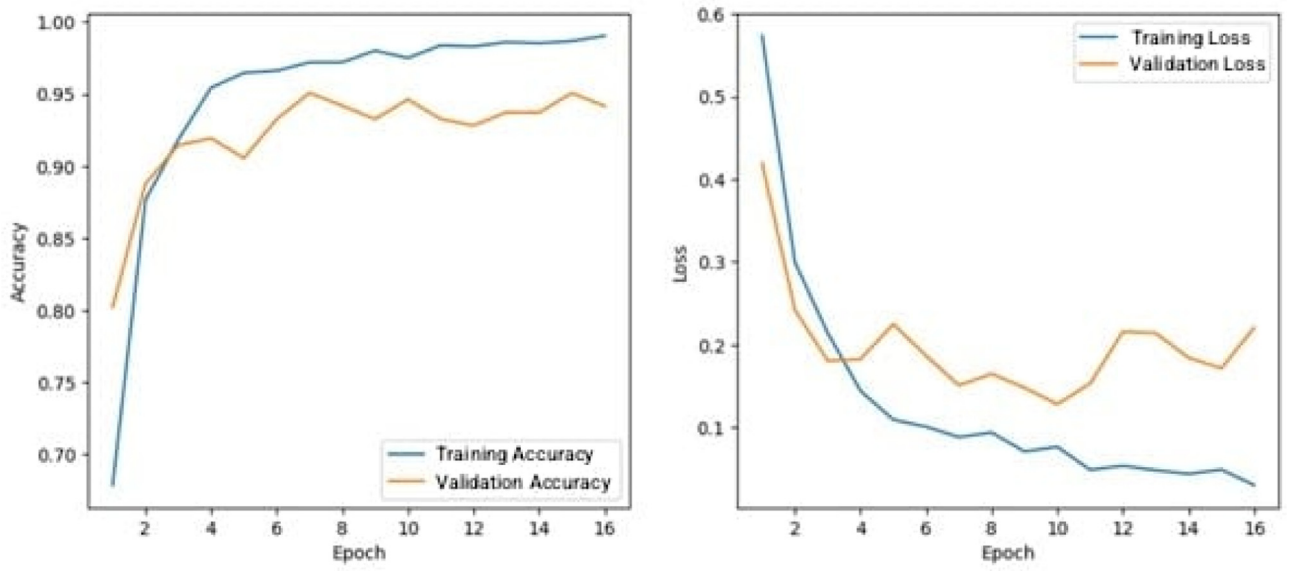


Figure 9. Accuracy and loss curve of the FT-transformer (seed 0)

In addition, an ablation study was conducted to understand the contributions of each component of the proposed architecture. We evaluated and compared the performances of three configurations: GNN, FT-transformer, and GNN + FT-transformer. Table 5 summarizes the results of cross-validation and ablation studies for the three seeds.

5. DISCUSSION

5.1. Train-Test Split. The proposed GNN + FT-Transformer model achieved the highest performance across all key metrics, including 93% accuracy, 0.89 precision, 0.96 recall, and a macro-F1 score of 0.92. Traditional ML models, such as *random forest* and *naïve Bayes*, performed reasonably well but had lower recall, indicating reduced sensitivity to the smoker class.

DL models such as Bi-GRU and Bi-LSTM showed improvements: Bi-GRU achieved up to 89% accuracy; however, they were still outperformed by the proposed hybrid model. As shown in Figure 3, the confusion matrix confirms a balanced classification of the model with fewer false positives and negatives. The receiver operating characteristics (ROC) curve in Figure 4, which has an area under the curve (AUC) of 0.97, further highlights the strong discriminative capability.

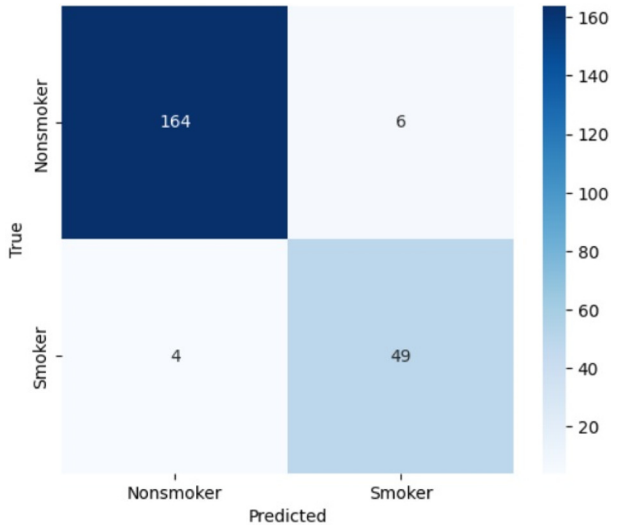


Figure 10. Confusion matrix of GNN + FT-transformer (seed 0)

Qualitatively, the model benefits from the complementary strengths of the GNN and FT-transformer. The GNN component captures the structural relationships among features, such as family and peer influence or mental health, by modeling them as a fully connected graph.

The FT-transformer adds contextual depth by focusing on the most informative features in each instance. This combination helps to manage class imbalances and irregular feature distributions. As shown in Figure 5, the training-validation loss curve shows smooth convergence with no major overfitting. Additionally, the attention weights from both the GNN and transformer modules offer interpretability, consistent with known behavioral patterns of smoking.

5.2. Cross Validation (Multiple Random Seed).

Graph Neural Network (GNN). To evaluate the standalone effectiveness of the GNN, we trained the model using five-fold cross-validation with three different random seeds: 0, 42, and 123. Of these configurations, the model initialized with seed 123 achieved the highest performance.

The confusion matrix in Figure 6 for seed 123 shows how well the GNN could distinguish between classes with minimal misclassifications. Figure 7 shows the training-validation accuracy and loss curve for seed 123. This standalone GNN model served as the baseline for ablation analysis. Its performance validates the ability of the GNN to capture topological relationships and structural dependencies within the data effectively.

FT Transformer. To isolate the contribution of feature-based attention mechanisms, we evaluated the FT-transformer independently without incorporating any graph-structural information. We conducted five-fold cross-validation with three different random seeds: 0, 42, and 123. Of these, the model trained

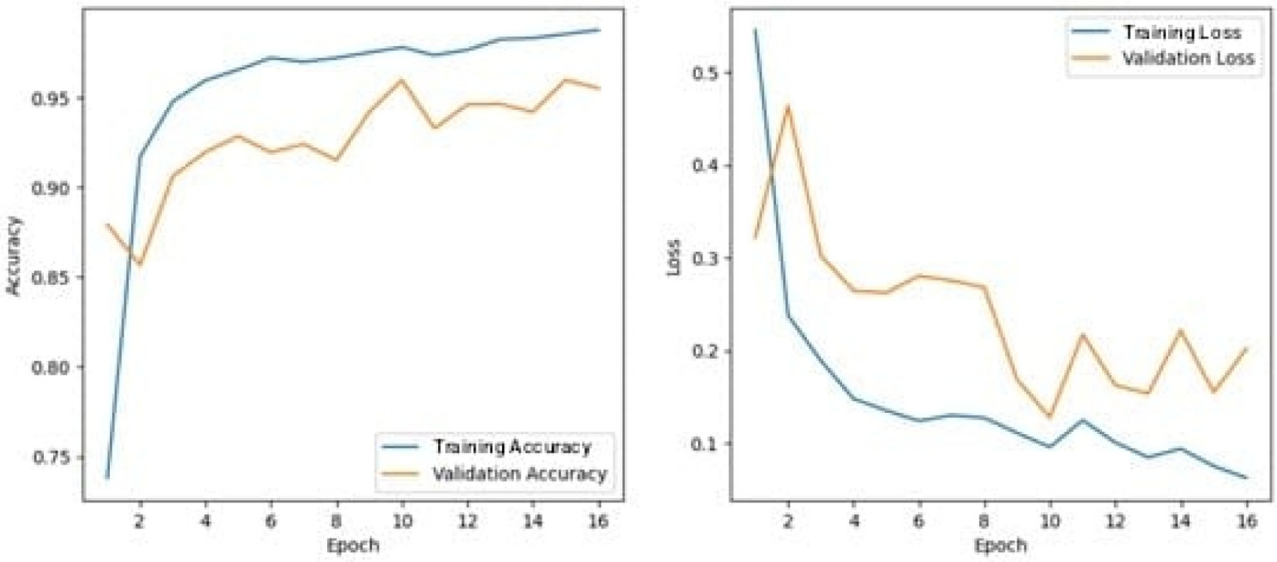


Figure 11. Accuracy and loss curve of GNN +FT-transformer (seed 0)

with seed 0 consistently achieved the highest average accuracy during cross-validation. The confusion matrix shown in Figure 8 further demonstrates the ability of the FT-transformer to model complex patterns in tabular data effectively using its feature-wise attention mechanism. Figure 9 shows the training-validation accuracy and loss curves for seed 0, reflecting the stable convergence of the model. The FT-transformer effectively captures complex feature interactions in the tabular data using its attention-based mechanism, offering strong standalone performance. However, it cannot model inter-entity relationships, limiting its effectiveness in tasks for which structural dependencies are critical.

GNN + FT-Transformer. In the final stage of our ablation study, we evaluated the hybrid GNN + FT-transformer model, which integrated the strengths of both GNNs and FT-transformers to capture relational and feature-level patterns simultaneously. We conducted five-fold cross-validation with three random seeds: 0, 42, and 123. The hybrid model achieved the best performance with seed 0, attaining the highest mean accuracy and showing strong generalization across the validation folds. The confusion matrix in Figure 10 for seed 0 shows highly accurate predictions with minimal misclassifications, indicating that the hybrid model effectively learned from both the graph structure and tabular features. The training–validation curves illustrated in Figure 11 show smooth convergence with no significant overfitting, further validating the stability and robustness of the hybrid approach.

6. CONCLUSIONS

Herein, we have proposed a hybrid GNN + FT-transformer model to classify smoking behavior using sociodemographic, psychological, and behavioral features. The model outperformed traditional ML and DL baselines, achieving an accuracy of 93% and 0.92 macro-F1 score. The success of the model stems from the ability of the GNN to capture feature relationships and the strength of the transformer in the contextual representation. To ensure robustness and reliability, the model was evaluated using a stratified five-fold cross-validation, which consistently demonstrated stable performance across cross-validation runs, thereby mitigating the risk of overfitting and variance arising from data splits. Further evaluation using a confusion matrix, ROC curve, and loss analysis confirmed the robustness and generalization of the model. The approach also offers interpretability and was consistent with known behavioral factors. This study highlights the advantages of hybrid architectures for behavioral prediction and public health applications. A key limitation of the current hybrid model is its dependence on predefined graph structures, which may not accurately capture complex or evolving relationships in the data. Future studies will focus on enabling automatic graph construction and exploring adaptive architectures to improve generalization, particularly in real-world settings with irregular or domain-shifted data.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Sk. Md Abir Hasan Imran – *Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology, Saidpur 5311, Bangladesh;* 0009-0008-3572-3897
Email: researchabir15@gmail.com

Author

Arupa Barua – *Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Bangladesh;* 0009-0004-5261-225X
Email: arupab362@gmail.com

Corresponding Author

Md. Osama – *Research Mentor, Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology, Saidpur 5311, Bangladesh;* 0009-0001-5307-9993
Email: osama_cse@baust.edu.bd

ACKNOWLEDGEMENTS

The authors express their sincere gratitude to the Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology (BAUST), Saidpur Cantonment, Nilphamari, Bangladesh, for providing continuous support, valuable guidance, and necessary resources throughout this research.

REFERENCES

(1) Ali, N., Paul, Y., Husain, A. & Hussain, H. in *The International Conference on Recent Innovations in Computing*. 93–116 (Springer).

(2) Carney, R. *et al.* The clinical and behavioral cardiometabolic risk of children and young people on mental health inpatient units: A systematic review and meta-analysis. *General Hospital Psychiatry* **70**, 80–97 (2021).

(3) Khan, A., Elhassan, M. A., Khan, S. & Deng, H. Deep learning-based smoker classification and detection: An overview and evaluation. *Expert Systems with Applications*, 126208 (2024).

(4) Ortis, A., Caponnetto, P., Polosa, R., Urso, S. & Battiato, S. A report on smoking detection and quitting technologies. *International journal of environmental research and public health* **17**, 2614 (2020).

(5) Khan, A. A., Laghari, A. A. & Awan, S. A. Machine learning in computer vision: A review. *EAI Endorsed Transactions on Scalable Information Systems* **8** (2021).

(6) Melchior, M., Chastang, J.-F., Mackinnon, D., Galéra, C. & Fombonne, E. The intergenerational transmission of tobacco smoking—the role of parents' long-term smoking trajectories. *Drug and alcohol dependence* **107**, 257–260 (2010).

(7) Becker, T. D., Arnold, M. K., Ro, V., Martin, L. & Rice, T. R. Systematic review of electronic cigarette use (vaping) and mental health comorbidity among adolescents and young adults. *Nicotine and Tobacco Research* **23**, 415–425 (2021).

(8) Mahabee-Gittens, E. M., Xiao, Y., Gordon, J. S. & Khoury, J. C. The dynamic role of parental influences in preventing adolescent smoking initiation. *Addictive behaviors* **38**, 1905–1911 (2013).

(9) Mays, D. *et al.* Parental smoking exposure and adolescent smoking trajectories. *Pediatrics* **133**, 983–991 (2014).

(10) Scalici, F. & Schulz, P. J. Parents' and peers' normative influence on adolescents' smoking: results from a Swiss-Italian sample of middle schools students. *Substance abuse treatment, prevention, and policy* **12**, 1–9 (2017).

(11) Fluharty, M., Taylor, A. E., Grabski, M. & Munafò, M. R. The association of cigarette smoking with depression and anxiety: a systematic review. *Nicotine & Tobacco Research* **19**, 3–13 (2016).

(12) De Luna, R. G. *et al.* in *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*. 84–89 (IEEE).

(13) Ali, N., Paul, Y., Husain, A. & Hussain, H. in *Proceedings of International Conference on Recent Innovations in Computing*. 93 (Springer Nature).

(14) Jarvis, M., Tunstall-Pedoe, H., Feyerabend, C., Vesey, C. & Salloojee, Y. Biochemical markers of smoke absorption and self reported exposure to passive smoking. *Journal of Epidemiology & Community Health* **38**, 335–339 (1984).

KFUPM
Journal of Undergraduate Research International

Diamond Open Access 45

Article Computer Engineering/Computer Science

FPGA-Based Accelerator for Quantized CNNs: High-Throughput Edge Deployment with Optimized Resource Utilization

Zeyad Emad Abdel-Mawjoud[1] and Ahmed S. Abd-Rabou Mohammed[2*]

Cite https://doi.org/10.64589/juri/209734

Submitted: June 04, 2025 **Revised:** July 30, 2025 **Accepted:** August 20, 2025

ABSTRACT

This paper presents MaxNet, a high-throughput field-programmable gate array (FPGA)-based accelerator for quantized convolutional neural networks (CNNs) designed to meet the demand for efficient edge artificial intelligence (AI) deployment in low-cost hardware. By targeting the Intel MAX 10 Field Programmable Gate Array (FPGA) (10M08DAF484C8GES), MaxNet was distinguishable from prior works focused on high-end platforms, offering a tailored solution for resource-constrained applications such as the Internet of Things (IOT) and embedded vision. The optimized two-layer CNN with 8-bit quantization (Q0.8 inputs/activations, Q1.7 weights) achieved 77% accuracy with the CIFAR-10 database, demonstrating a throughput of 8,065 frames per second (fps; 0.124 ms/image) and power consumption of 1.2 W, using 861 lookup tables (LUTs) (11%) and 9 M9K memory blocks (2.9%), implemented, synthesized, and tested using Intel Quartus Prime Standard Edition 22.1 Comprehensive power measurements derived from Intel Quartus Power Analyzer (for FPGA design), psutil (a Python-based system monitoring library for CPU measurements), and nvidia-smi (NVIDIA's GPU management and monitoring utility)*superior efficiency compared to the central processing unit (Intel Core i5-10400, 578 frames per second (fps), 15 W) and graphics processing unit (NVIDIA GTX 1650, 730 fps, 50 W) baselines. Ablation studies validated the two-layer design and quantization choices, while sensitivity analysis was used to optimize the clock frequency and numerical formats. The resource utilization and per-class accuracy were confirmed, reinforcing MaxNet's suitability for low-power, high-performance edge AI in cost-effective FPGAs. This work is expected to pave the way for exploring 4-bit quantization, structured pruning, and transfer learning for domain-specific applications, such as medical imaging.

Keywords: deep learning inference, quantized CNN, GenAI acceleration, FPGA, edge AI, resource-constrained deployment, hardware--software co-design

1. INTRODUCTION

Convolutional neural networks (CNNs) have become the foundation of modern computer vision, achieving high accuracy in tasks such as image classification, object detection, and autonomous navigation [Sandler et al., 2018; Iandola et al., 2016]. Their ability to extract hierarchical features from raw data has enabled a wide range of real-world applications, including medical imaging, facial recognition, and industrial fault detection [Chung et al., 2023].

Despite their effectiveness, CNNs are computationally intensive, demanding substantial processing power, memory, and energy constraints that complicate deployment on embedded or portable devices. Edge artificial intelligence (edge AI) refers to performing AI computations directly on devices located at the "edge" of the network, closer to where the data is generated, rather than relying on remote cloud servers. Examples include IoT nodes, drones, and wearable systems. By reducing dependence on cloud connectivity, edge AI enables real-time, low-latency inference while operating under strict energy and memory budgets [Mazumder et al., 2022; Wang et al., 2022]. To meet these constraints, research has emphasized model compression and efficient architectures including quantization, pruning, and lightweight backbones (e.g., MobileNetV2, SqueezeNet), as well as low-precision designs and FPGA-oriented frameworks (e.g., FINN, LUTNet) [Sandler et al., 2018; Iandola et al., 2016; Umuroglu et al., 2017; Wang et al., 2019].

Field-programmable gate arrays (FPGAs), reconfigurable chips programmable after fabrication, are promising edge-AI accelerators due to their parallelism, reconfigurability, and low-power consumption compared to traditional processors [Hou et al., 2020; Cho & Kim, 2020]. High-end platforms (e.g., Xilinx Zynq and Virtex) have been widely used for CNN acceleration with high-throughput and scalability. However, low-cost FPGAs (e.g., Intel MAX 10) remain relatively underexplored because limited logic, multipliers, and on-chip memory restrict deployable model size and throughput [Wang et al., 2022].

Considering this gap, we developed MaxNet, a lightweight, 8-bit quantized CNN optimized for the Intel MAX 10 FPGA, to investigate the feasibility of high-throughput, low-power

inference under strict hardware constraints. The model was trained with quantization-aware training and evaluated on the CIFAR-10 dataset. Our contribution is a resource-efficient CNN architecture and implementation framework that advances practical edge-AI deployment on cost-effective FPGAs, with applicability to IoT, embedded vision, and wearable devices.

2. METHODOLOGY

2.1. Overview of MaxNet and the study design. MaxNet is a lightweight CNN optimized for the Intel MAX 10 FPGA (10M08DAF484C8GES, 8,064 logic elements, 387,072 memory bits, and 48 9-bit multipliers), targeting edge AI applications. The architecture comprised two convolutional layers (3×3 kernels, 16 and 32 filters), batch normalization (BN), rectification linear unit (ReLU) activation, 2×2 max-pooling, global average pooling (GAP), and a dense layer for image classification (10 classes, 60,000 32×32 RGB images). The CNN was implemented in Quartus Prime Standard software (Edition 22.1) to validate the minimalist design and compare the effect of the number of layers on performance, e.g., lookup table (LUT) usage, accuracy, and throughput. The simple two-layer architecture was designed to minimize pipeline stages and achieve ultra-low latency, thereby supporting a "micro-acceleration" approach, enabling potential multi-instance deployment for ensemble learning or multi-task processing.

2.2. CNN layer design. An overview of the MaxNet CNN layer architecture is shown in Figure 1, which illustrates the sequential flow of convolution, pooling, and classification layers. The input passes through two convolutional–pooling stages, followed by global average pooling and a fully connected dense layer leading to the final softmax output. The properties of each layer, including kernel size, number of filters, and output dimensions, are summarized in Table 1. This combination of figure and table provides both a visual representation of the overall architecture and a detailed specification of the design parameters. Each layer is further defined and discussed in detail in the subsequent sections.

- ***Input quantization***: Each input pixel $X_{h,w,c} \in [-1,1]$, where h and w denote the spatial coordinates and c denotes the color channel, was quantized to 8-bit precision using the Q0.8 fixed-point format. This process converts the continuous input into a clipped integer representation:

$$I_{h,w,c} = clip\left(\left\lfloor \frac{X_{h,w,c}}{2^{-7}} \right\rfloor, -128, 127\right) \quad (1)$$

Where $I_{h,w,c}$ is the quantized pixel value, and the function $clip(.)$ ensures that the result lies within the valid integer range $[-128,127]$.

- ***Convolutional Layer***: The convolutional stage generates feature maps by applying a set of kernels to the quantized input. The output feature value at spatial position (i,j) for the output channel k is computed as

$$Y_{i,j,k} = \sum_{c=0}^{C_{in}-1} \sum_{m=0}^{2} \sum_{n=0}^{2} W_{k,c,m,n} \cdot I_{i+m,j+n,c} + \beta_k \quad (2)$$

where C_{in} represents the number of input channels, $W_{k,c,m,n}$ denotes the convolutional kernel weight associated with output channel k, input channel c, and kernel coordinates (m,n), and $I_{i+m,j+n}$ is the corresponding input pixel value. This operation extracts local spatial patterns from the input image.

- ***Batch normalization***: To stabilize the distribution of features and accelerate training convergence, batch normalization was applied after each convolution. The standard BN transformation for output channel k is expressed as

$$\hat{Y}_{i,j,k} = \gamma_k \cdot \frac{Y_{i,j,k} - \mu_k}{\sqrt{\sigma_k^2 + \epsilon}} + \beta_k \quad (3)$$

Where μ_k and σ_k^2 denote the mean and variance of activations for channel k, γ_k is a learned scaling factor, β_k is a learned shift parameter, and ϵ is a small constant to avoid division by zero. For FPGA deployment, this expression was simplified into an equivalent linear transformation.

$$\hat{Y}_{i,j,k} = A_k Y_{i,j,k} + B_k \quad (4)$$

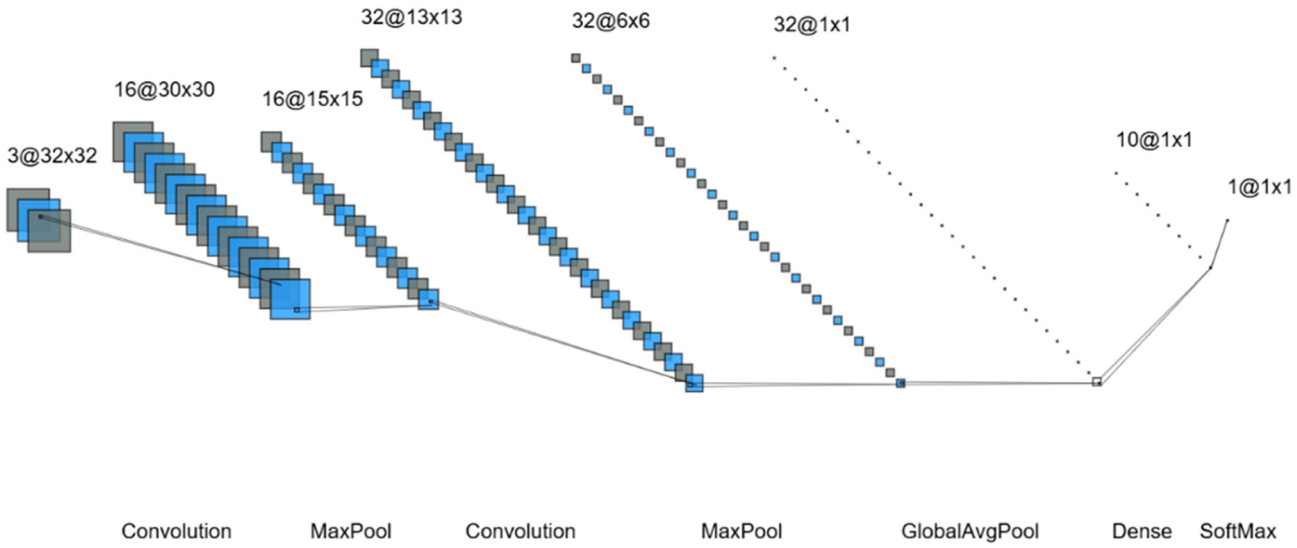


Figure 1. MaxNet CNN Architecture Diagram

where the constants

$$A_k = \frac{G_k}{\sqrt{\sigma_k^2 + \epsilon}} \quad B_k = \frac{-\mu_k}{\sqrt{\sigma_k^2 + \epsilon}} + \beta_k \tag{5}$$

They were precomputed offline. This eliminates runtime division and square-root operations, thereby improving inference efficiency.

- ***ReLU activation:*** Following batch normalization, the rectified linear unit (ReLU) activation function introduces nonlinearity into the network by suppressing negative activations. It is defined as

$$\mathrm{Z_{i,j,k}} = \max\left(0, \hat{Y}_{i,j,k}\right) \tag{6}$$

Where $Z_{i,j,k}$ represent the activated feature value. ReLU not only enables the network to learn complex relationships but also simplifies hardware implementation, as it can be realized using simple comparators.

- ***Max-pooling:*** To reduce the spatial resolution while retaining the most salient features, max pooling was applied over non-overlapping 2×2 windows. For each output location (i, j) and channel k, the pooled value is given by

$$P_{i,j,k} = max\left(Z_{2i,2j,k}, Z_{2i+1,2j,k}, Z_{2i,2j+1,k}, Z_{2i+1,2j+1,k}\right) \tag{7}$$

Where $P_{i,j}$, and k denote the reduced feature map. This downsampling operation halves the spatial dimensions, thereby reducing computational cost and memory usage in subsequent layers.

- ***Global average pooling***:Instead of employing fully connected layers that flatten the feature maps, MaxNet uses global average pooling (GAP) to reduce each feature map to a single representative value. For channel k, the GAP output is computed as

$$G_k = \frac{1}{H * W} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} P_{i,j,k} \tag{8}$$

Where H and W are the height and width of the pooled feature map. This approach greatly reduces the number of parameters, mitigates overfitting, and produces a compact representation for classification.

- ***Dense layer and output:*** The final dense layer produces the class logits by linearly combining the GAP outputs. For each class $c \in \{0, 1, \ldots, 9\}$, the output is defined as

$$O_c = \sum_{k=0}^{31} W_{c,k} \cdot G_k + \beta_c, c \in \{0, \ldots, 9\} \tag{9}$$

Where O_c is the logit corresponding to class c, $\mathrm{W_{c,k}}$ represents the weight connecting GAP feature $\mathrm{G_k}$ to class c, and β_c is the class-specific bias. All weights and biases were quantized to the same fixed-point formats as earlier layers, ensuring consistent precision throughout the pipeline.

2.3. FPGA architecture design. The MaxNet accelerator used a streaming-pipelined VHDL architecture on an Intel MAX 10 (10M08SAE144C8GES) FPGA with 8,064 LEs, 101 I/Os, 387,072 memory bits, and 48 9-bit multipliers. A custom control unit with buffers and a finite state machine (FSM) managed the data flow between layers to reduce latency. The carefully designed 4-stage pipeline breaks down the convolution computations into smaller sequential operations, allowing each convolutional output to be computed in approximately 6–12 clock cycles. This pipeline increases throughput and reduces critical path delays, enabling a clock frequency of 100 MHz.

The FPGA design includes dedicated hardware modules for convolution, BN, ReLU, max-pooling, and GAP. These blocks communicate through a streaming interface and are efficiently mapped to M9K embedded memory blocks using fixed-point representations (Q0.8, Q8.8, and Q16.16) for the intermediate data and weights. The use of dedicated read-only memory (ROM) and random access memory (RAM) units at each stage ensures minimal bottlenecks in memory access and processing.

2.4. Quantization-aware training. Quantization-aware training (QAT) was performed using the TensorFlow/Keras deep learning framework [Abadi et al., 2016; Chollet, 2015] on images from the CIFAR-10 dataset [Krizhevsky & Hinton, 2009]. The training simulated 8-bit fixed-point arithmetic, employing the Q0.8 format for inputs and activations, Q1.7 for weights, and Q16.16 for biases, where Q denotes the quantization scheme and the number indicates the allocated fractional and integer bits.

Table 1. *MaxNet Layer Overview*

Layer	InputShape	InputFormat	OutputShape	OutputFormat
Input	$32 \times 32 \times 3$	Q0.8	$32 \times 32 \times 3$	Q0.8
*Conv*1 $(3 \times 3, 16)$	$32 \times 32 \times 3$	Q0.8	$30 \times 30 \times 16$	Q16.16
*BatchNorm*1 + *ReLU*	$30 \times 30 \times 16$	Q16.16	$30 \times 30 \times 16$	Q0.8
Ma × *Pool*1	$30 \times 30 \times 16$	Q0.8	$15 \times 15 \times 16$	Q0.8
*Conv*2 $(3 \times 3, 32)$	$15 \times 15 \times 16$	Q0.8	$13 \times 13 \times 32$	Q0.8
*BatchNorm*2 + *ReLU*	$13 \times 13 \times 32$	Q0.8	$13 \times 13 \times 32$	Q0.8
Ma × *Pool*2	$13 \times 13 \times 32$	Q0.8	$6 \times 6 \times 32$	Q0.8
GlobalAvgPool	$6 \times 6 \times 32$	Q0.8	32	Q0.8
Dense	32	Q0.8	10	Q0.8

2.5. Memory architecture and allocation. Effectively managing the memory allocation is critical for achieving the efficient operation of MaxNet on the resource-constrained MAX 10 FPGA. All weights, biases, and intermediate feature maps are stored in the on-chip memory to eliminate external memory access and reduce latency and power usage. Table 2 details the allocation and usage of the FPGA's M9K memory blocks. The table also lists the data formats, sizes (in kilobytes, where 1 K =1024 bytes), and memory-block roles. Each memory block serves a specific purpose, ranging from storing weights and biases to holding intermediate feature maps and outputs. For example, the ROM blocks store the fixed parameters (weights, biases, and batch normalization coefficients) and RAM blocks store input images and intermediate feature maps. In the table, Conv1 and Conv2 are the first and second convolution layers, respectively.

2.6. Memory optimization.

- **Buffer-based data flow**: The FSM control unit uses on-chip buffers (implemented as dual-port RAM) to stage intermediate feature maps, ensuring continuous data availability for the 4-stage pipeline to avoid pipeline stalls.
- **Memory packing**: Weights and coefficients are packed efficiently (e.g., Q1.7 for Conv weights, Q1.15 for BN coefficients) to minimize the memory footprint. For example, Conv1 weights ($3 \times 3 \times 3 \times 16 = 432$ weights) use 0.422 KB, leveraging the compact representation of Q16.16.
- **No external memory**: All data resides on-chip, eliminating external Dynamic Random Access Memory (DRAM) access (common in high-end FPGAs), which reduces power and latency but constrains the model size, justifying the two-layer CNN.
- **Resource allocation**: The use of Quartus Prime Standard Edition 22.1 optimizes memory-block assignments to balance storage and logic requirements.

2.7. Control logic unit using FSM. The control logic unit employs a custom buffer-based FSM implemented in top_module.vhd to manage the data flow across the 4-stage pipeline (Conv1+BN+ReLU+MaxPool1, Conv2+BN+ReLU+MaxPool2, GAP, Dense layer). Unlike first-in-first-out (FIFO)-based designs that can introduce stalls owing to data dependencies, FSM uses on-chip dual-port RAM buffers to stage intermediate feature maps (e.g., $30 \times 30 \times 16$ after Conv1 and $6 \times 6 \times 32$ after MaxPool2). The FSM operates as follows.

- **States**: The FSM transitions through the states for each pipeline stage (e.g., LOAD_INPUT, CONV1, BN_RELU1, MAXPOOL1), coordinating data reads/writes, and computation triggers.
- **Buffer management**: Buffers (e.g., RAM_C1_Out and RAM_MF1_Out) ensure data availability, allowing parallel execution of multiply-accumulate (MAC) operations and memory access. For example, Conv1 outputs are buffered, whereas the BN + ReLU stage directly processes these outputs, thereby reducing pipeline stalls.
- **Latency reduction**: The buffer-based approach minimizes data transfer delays owing to the optimized scheduling of memory read/write functions and computation overlap.
- **Synchronization**: The FSM ensures synchronization across modules using a 100 MHz clock to trigger transitions, completing one image inference in 6–12 cycles (0.124 ms).

The efficiency of the FSM was validated using the tb_top_module.vhd test bench, which simulated the CIFAR-10 image inputs and verified the output logits.

2.8. Implementation methodology. MaxNet's FPGA implementation utilizes a fully pipelined VHDL architecture specifically designed to accommodate the resource constraints of the Intel MAX 10 FPGA. The design leverages parallelized convolution engines and a buffer-based FSM control unit to optimize data flow and reduce latency. CNNs were selected for their linear

Table 2. *Memory block specifications for MaxNet on the MAX 10 FPGA*

Memory Block	Contents	Format	Size (KB)	Purpose
ROM_W_Conv1	Conv1 weights ($3 \times 3 \times 3 \times 16$)	Q16.16	0.422	Kernel weights
ROM_W_Dense	Dense weights (32×10)	Q8.8	0.313	Kernel weights
ROM_B_Conv1	Conv1 biases (1×16)	Q16.16	0.016	-
ROM_B_Conv2	Conv2 biases (1×32)	Q8.8	0.016	-
ROM_BN1_Coeff	BN1 A/B coefficients	Q1.15	0.016	Packed A_k/B_k
RAM_InImg	Input image ($32 \times 32 \times 3$)	Q8	3	Dual-port BN_ReLU input for simultaneous read/write
RAM_C1_Out	Conv1 out ($30 \times 30 \times 16$)	Q16.16	14.4	BN-ReLU input
RAM_BR1_Out	BN+ReLU1 out ($30 \times 30 \times 16$)	Q8	14.4	Combined output
RAM_MF1_Out	MaxPool1 out ($15 \times 15 \times 16$)	Q8	3.6	Stride 2
RAM_C2_Out	Conv2 out ($13 \times 13 \times 32$)	Q8	5.4	BN-ReLU input
RAM_BR2_Out	BN+ReLU2 out ($13 \times 13 \times 32$)	Q8	5.4	Combined output
RAM_MP2_Out	MaxPool2 out ($6 \times 6 \times 32$)	Q8	1.2	Stride 2
RAM_GAP_Out	GAP (1×32)	Q8	0.031	-
RAM_Dense_Out	Dense outputs (10)	Q8	0.01	-

data flow, aligned with the 387,072-bit memory and 48 9-bit multipliers of MAX 10. The modular design ensures scalability and maintainability, and is organized into core processing modules and integration/control components.

2.8.1. Core modules.

- **conv_engine.vhd** implements convolutional operations with parameterized 3 × 3 filters and input feature maps, performing MAC operations using the 48 9-bit multipliers of MAX 10 for parallel processing. Weights were stored in the Q1.7 format, with inputs in Q0.8, and outputs in Q16.16, to maintain precision during accumulation.
- **BatchNorm_ReLU_engine.vhd** combines BN and ReLU activation to stabilize training and introduce nonlinearity. BN coefficients (A_k, B_k) are precomputed offline in Q1.15 format, and the ReLU uses comparators to zero negative values, minimizing logic overhead.
- **maxpool_engine.vhd** performs 2 × 2 max-pooling, selecting the maximum value in each local neighborhood to reduce the spatial dimensions (e.g., 30 × 30 to 15 × 15), enhancing computational efficiency and robustness to spatial variations.
- **global_avg_pool_engine.vhd** computes the GAP, averages feature maps (e.g., 6 × 6 × 32) into a single value per channel (32 values), reduces parameters, and prepares features for classification.
- **dense_engine.vhd** implements the fully connected layer, computing class logits via the weighted sums of GAP outputs and biases in Q0.8 format, producing 10 class probabilities for the CIFAR-10 images.
- **types_pkg.vhd** defines the global constants, data types (e.g., Q0.8, Q1.7, and Q16.16), and configuration parameters (e.g., filter sizes and channel counts) to ensure consistency across the modules.

2.8.2. Integration and control.

- **top_module.vhd** acts as a top-level architecture, interconnecting core modules via signal routing and high-level sequencing. A custom buffer-based FSM control unit manages the data flow and buffers the intermediate feature maps (e.g., 30 × 30 × 16 after Conv1) in the on-chip dual-port RAM to avoid external memory access. The FSM optimizes pipeline synchronization by ensuring data availability and minimizing stalls.
- **tb_top_module.vhd** provides a test bench to validate the functionality of the pipeline, initializing quantized CIFAR-10 images, applying stimuli, and monitoring outputs (class logits) to verify the correctness and timing.

Figure 2 shows the schematic of the conv1_engine module. It highlights how convolution interacts with the control unit and memory buffers. Inputs are fetched from on-chip RAM, processed through the multiply-accumulate array, and passed forward to normalization and activation. The diagram illustrates the pipelined structure and interconnections that enable efficient inference under tight FPGA resource constraints.

3. POWER MEASUREMENTS

Power consumption measurements are critical to ensure that the accelerator meets the requirements of efficient edge AI applications. This section outlines the power measurements for MaxNet on an Intel MAX 10 FPGA, Intel Core i5-10400 CPU, and NVIDIA GTX 1650 GPU.

The power usage of MaxNet on the Intel MAX 10 FPGA (10M08DAF484C8GES) was estimated using Quartus Prime Power Analyzer (Standard Edition 22.1), which models both static and dynamic power under typical PVT (process, voltage, temperature) conditions. The accuracy of this estimate was ensured by feeding the analyzer with real post-fit activity data obtained from simulation vectors, which reduces the risk of overestimation compared to vectorless analysis.

For the CPU baseline, the power consumption of the Intel Core i5-10400 was monitored using Python scripts with the psutil library, capturing system-level usage during CIFAR-10 inference. While psutil reports overall system power (CPU plus memory and supporting processes), multiple runs were performed to ensure stability of the readings, and the results were averaged to reduce variability.

For the GPU baseline, the power usage of the NVIDIA GTX 1650 was measured using nvidia-smi, which reports instantaneous and averaged values of both core and memory power consumption. Accuracy was ensured by sampling at a fixed interval throughout the inference process and averaging across runs, which minimized short-term fluctuations due to the parallel architecture of the GPU.

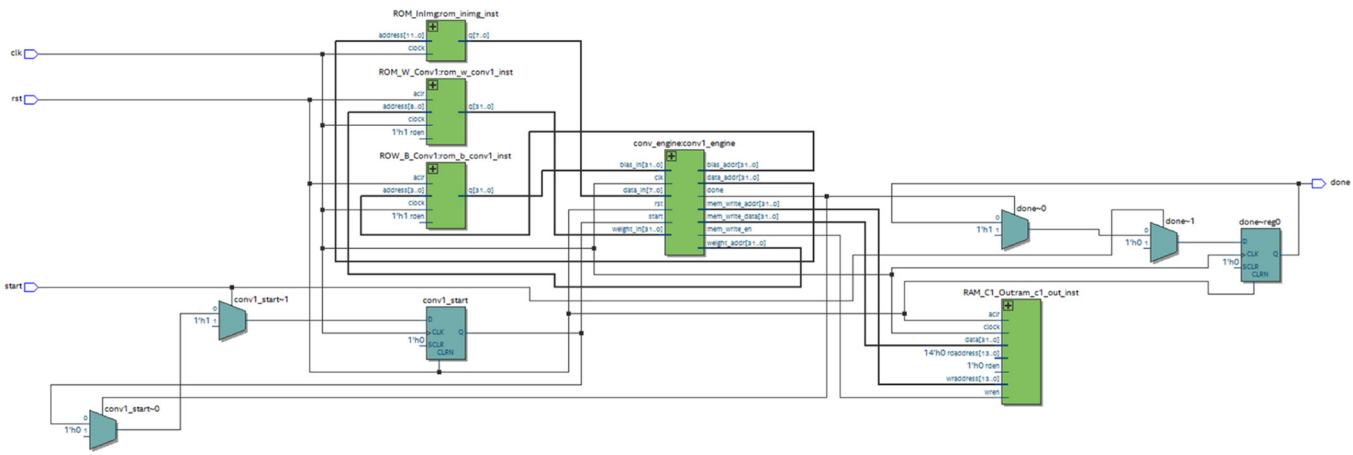


Figure 2. Schematic overview of part (conv1_engine) of the implemented CNN accelerator

4. RESULTS AND DISCUSSION

This section presents and discusses the performance, accuracy, and power efficiency of the MaxNet FPGA-based accelerator evaluated on an Intel MAX 10 FPGA (10M08DAF484C8GES) for CIFAR-10 inference. The performance of MaxNet was compared with that of the CPU (Intel Core i5-10400) and GPU (NVIDIA GTX 1650) as baselines. Synthesis and simulation were conducted using Quartus Prime Standard Edition 22.1, and power measurements were performed as described above. Furthermore, we evaluated the suitability of MaxNet for low-cost, energy-constrained edge AI applications such as IoT and embedded vision.

4.1. Performance overview. MaxNet achieved a throughput of 8,065 fps at 100 MHz, corresponding to a latency of 0.124 ms per CIFAR-10 image ($32 \times 32 \times 3$). The implementation utilized 861 LUTs (11% of the 8,064 available), nine M9K memory blocks (2.9% of 306), and one phase-locked loop (PLL, 25% of 4), as validated by synthesis. The measured power consumption was 1.2 W, which is significantly lower than the 15 W required by a CPU and the 50 W required by a GPU, making MaxNet well-suited for low-power edge deployment.

The 4-stage pipeline (Conv1 + BN + ReLU + MaxPool1, Conv2 + BN + ReLU + MaxPool2, GlobalAvgPool, and Dense layers) processes images in 6–12 clock cycles by leveraging parallelized convolution engines and a buffer-based FSM control unit. Ablation studies showed that extending the architecture to a 4-layer CNN increased LUT usage by ~30% (~1,200 LUTs) and power consumption by ~20% (~0.24 W), but yielded only a 1–2% accuracy improvement, validating the efficiency of the simpler two-layer design.

A sensitivity analysis further revealed that reducing the operating frequency to 50 MHz lowered throughput to 4,032 fps and reduced power consumption by ~15% (~0.18 W). These results confirm that a 100 MHz configuration provides the best balance between efficiency and performance.

4.2. Accuracy analysis. The accuracy of MaxNet was evaluated on the CIFAR-10 dataset to assess the impact of quantization-aware training (QAT) on classification performance. The quantized model achieved 77% accuracy, which is slightly lower than the 79% accuracy of the baseline floating-point (FP32) version of our model trained in Python before quantization. This minor reduction demonstrates that QAT effectively preserved most of the model's representational power while adapting it for efficient fixed-point inference on an FPGA.

Figure 3 compares the per-class accuracy of FP32 and quantized models. The results show that most classes (e.g., automobile, horse, ship) maintained identical or near-identical performance across both models, while others (e.g., airplane, cat, truck) exhibited slight variations. These class-specific fluctuations highlight how quantization affects certain feature distributions more than others, but the overall performance remains consistent.

Figure 4 presents the confusion matrices for FP32 and quantized models. The quantized model shows slightly higher misclassification for certain classes, such as dog and cat, yet improvements can also be observed for airplane and bird. This indicates that QAT redistributes representational precision across classes, maintaining a balanced accuracy profile while enabling efficient hardware deployment.

A comparison of MaxNet's performance with other FPGA-based accelerators highlights the balance achieved between accuracy and resource efficiency. While some accelerators report

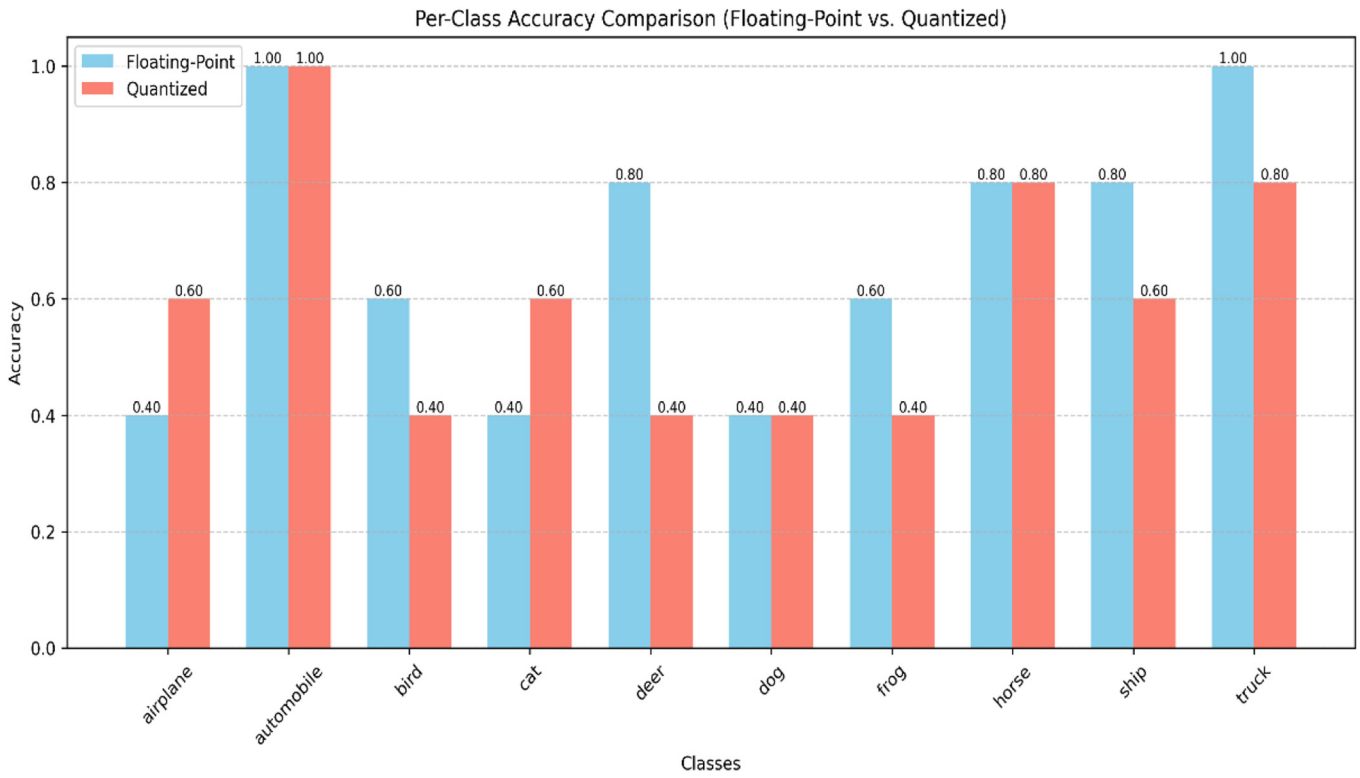


Figure 3. Accuracy comparison: FP32 vs. quantized model across CIFAR-10 classes

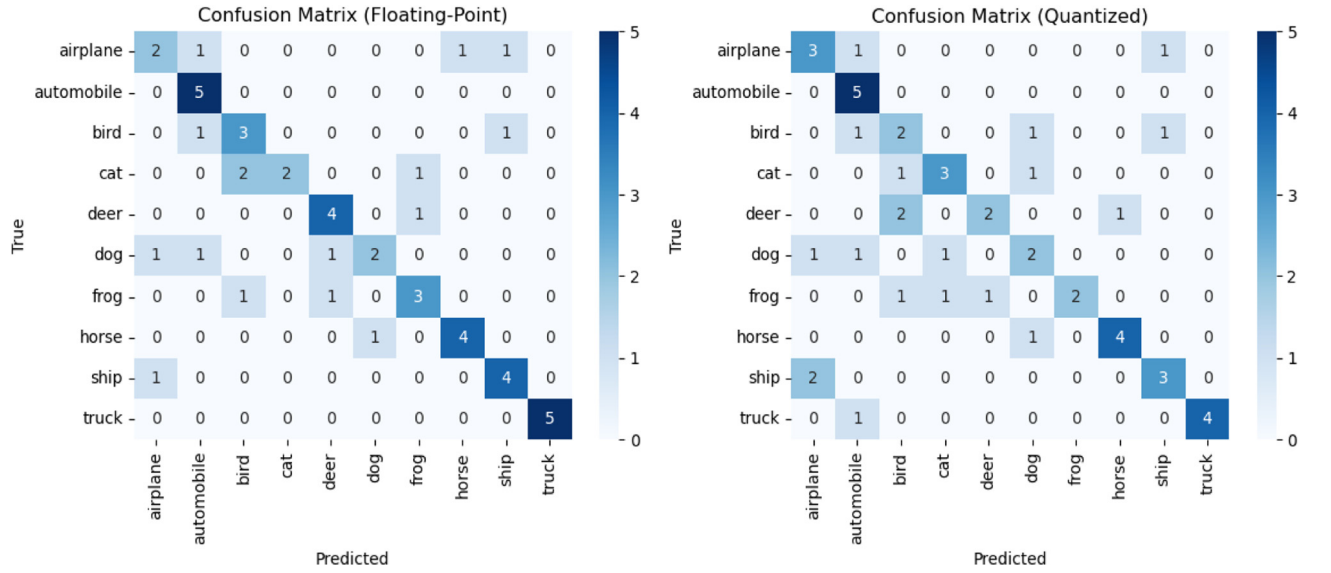


Figure 4. Confusion matrices comparing floating-point and quantized models

higher accuracy by adopting larger models or higher-precision arithmetic, these typically demand more hardware resources and power consumption, which limits their suitability for edge AI applications. In contrast, MaxNet demonstrates competitive accuracy with a compact design tailored for resource-constrained FPGAs.

These results underscore the trade-off inherent in deploying quantized CNNs on low-cost hardware. Although a slight accuracy drop occurs compared to the FP32 baseline, the efficiency gains in power and throughput make MaxNet well-suited for real-time edge deployment scenarios.

4.3. Platform comparison. MaxNet's performance was benchmarked against CPU and GPU baselines, with power measurements used to compare efficiency. The FPGA's 4-stage pipeline and buffer-based FSM reduce latency by ~20% compared to non-pipelined designs and by ~10–15% compared to FIFO-based designs, as validated by synthesis. The Q0.8 quantization simplifies arithmetic and enhances efficiency.

Notes:

- **FPGA (MAX 10):** Power estimated using Quartus Power Analyzer (static leakage in 40 nm, 3.3V; dynamic switching in 861 LUTs, 9 M9K blocks @100 MHz). 4-bit quantization saved ~0.05 W but reduced accuracy to ~70%.
- **CPU (i5-10400):** Power measured via Python psutil during CIFAR-10 inference (includes non-CPU components). A 4-layer CNN raised power by ~10% (~1.5 W) with minimal accuracy gain.
- **GPU (GTX 1650):** Power measured via *nvidia-smi* (core + memory). 4-bit quantization saved ~5% (~2.5 W) but reduced accuracy to ~70%.

The FPGA's low latency and power efficiency underscore MaxNet's suitability for edge AI applications. The integration of on-chip memory and a buffer-based FSM eliminates external DRAM access, thereby further reducing both power consumption and latency compared to CPU and GPU platforms.

4.4. Comparison of MaxNet with other CNNs. FPGAs are widely used to accelerate convolutional neural networks (CNNs) in edge AI applications owing to their reconfigurability and parallel processing capabilities. While many FPGA-based accelerators target high-end platforms, MaxNet is designed for a low-cost Intel MAX 10 FPGA, optimizing a lightweight CNN for resource-constrained edge devices, such as IoT and embedded vision systems. This section compares MaxNet to the FPGA-based LeNet-5 implementation by Hou et al.[1], with additional comparisons to a study by Cho and Kim[2], LUTNet[3], and other frameworks[4–8], focusing on hardware targets, quantization strategies, and architectural efficiency (see Table 3).

Hou et al. implemented LeNet-5, a classic seven-layer CNN designed for handwritten digit recognition, on a Spartan-6 FPGA[1]. The model used two convolutional layers, two pooling layers, and three fully connected layers with 16-bit fixed-point arithmetic. By contrast, MaxNet adopts a simpler two-layer CNN with 8-bit quantization and batch normalization fusion, demonstrating efficient performance on a more complex dataset while consuming fewer hardware resources.

Cho and Kim proposed a data-optimized CNN accelerator on a Spartan-6 FPGA[2], using a single-layer CNN with 16-bit

Table 3. *Cross-Platform Performance Comparison*

Platform	Latency (ms/image)	Throughput (fps)	Power (W)	Accuracy (%)
FPGA (MAX 10)	0.124	8065	1.2	77
CPU (Intel Core i5-10400)	1.73	578	15	79
GPU (NVIDIA GTX 1650)	1.37	730	50	79

Table 4. *Comparison of FPGA-based CNN accelerators and lightweight CNNs*

Framework	Model	Dataset	Accuracy	Throughput (fps)	Latency (ms)	Power (W)	Resource Usage
MaxNet [this study]	2-layer CNN	CIFAR-10	76–77%	8,065	0.124	1.2	861 LUTs (11%), 9 memory blocks (2.9%)
LeNet-5 [1]	LeNet-5	MNIST	98%	2,000	0.5	-	10,750 LUTs (25%)
MobileNetV2 [4]	MobileNetV2	CIFAR-10	70.40%	70.94	14.1	-	524.25 KB, 176 DSPs
SqueezeNet [5]	SqueezeNet	ImageNet	57.5%	-	-	-	-
TinyMLNet [7]	Various	CIFAR-10	-	50,000	0.02	-	-
FINN [6]	Various quantum NNs	CIFAR-10	-	12,000	0.083	-	-
DNNWeaver [8]	Various	ImageNet	57–80.8%	-	-	-	-

fixed-point arithmetic and a pipelined architecture. Compared to this, MaxNet achieves higher efficiency through its two-layer design and reduced precision, offering better throughput–accuracy tradeoffs under tighter hardware constraints.

LUTNet introduced a hardware software co-design framework for binary neural networks on Zynq-7000 FPGAs[3]. Its reliance on LUT-based computations with binary or ternary weights improves throughput but increases LUT demand due to complex control logic. In contrast, MaxNet's straightforward 8-bit fixed-point design reduces resource pressure, aligning better with low-cost FPGAs.

Other lightweight CNN frameworks further illustrate these trade-offs. MobileNetV2 on XC7Z020 uses separable convolutions, which lower parameter count but introduce latency overhead and require DSP and memory resources[4]. SqueezeNet achieved compactness for ImageNet tasks but at the cost of reduced accuracy[5]. FINN provides high-speed inference with binarized networks but requires careful mapping to FPGA resources[6]. TinyMLNet maximizes throughput on Pynq-Z2[7], while DNNWeaver demonstrates scalable deployment of various models[8].

MaxNet's micro-acceleration approach distinguishes it from these designs. By prioritizing extreme resource efficiency and sustaining high throughput at low power, MaxNet balances accuracy and cost. Unlike MobileNetV2, which incurs latency from separable convolutions, or ResNet-style models that demand large memory for residual connections (beyond MAX 10's capacity), MaxNet avoids such overheads through its lightweight CNN and buffer-based FSM control. This design also enables multiple parallel MaxNet instances to run on a single device, scaling performance without exceeding power or memory limits, making it highly effective for edge AI applications. Overall, as summarized in Table 3, MaxNet clearly outperforms prior accelerators in terms of resource efficiency, throughput, and suitability for low-cost edge hardware.

5. LIMITATIONS

Power usage: The 1.2 W power usage measured for the FPGA assumes typical PVT conditions, which may not reflect actual experimental conditions. The CPU power (15 W) included non-CPU components, and the GPU power (50 W) varied with the workload intensity. Direct measurements using power meters or hardware counters (e.g., RAPL for the CPU and NVML for the GPU) would enhance the accuracy. Future work could include dynamic voltage and frequency scaling to adjust the FPGA clock/voltage. The Quartus TimeQuest Timing Analyzer could be used to optimize bottlenecks in memory access and computation and implement adaptive clocking to reduce power by 10–20% for sparse or low-intensity tasks.

Accuracy trade-offs: A 2% accuracy drop (77% for MaxNet vs. 79% for FP32) is acceptable for edge applications, but can be improved with mixed-precision quantization (to explore 4-bit weights with 8-bit activations) or transfer learning (to adapt MaxNet for domain-specific datasets, such as medical imaging).

Scalability: MaxNet was optimized for the MAX 10 FPGA. Scaling to larger FPGAs or complex models (e.g., U-Net) requires 2–3 times more memory (approximately 800,000–1,200,000 bits).

6. CONCLUSIONS

MaxNet demonstrated that 8-bit quantized CNNs can be efficiently deployed on low-cost FPGAs, achieving real-time inference (8,065 fps) with minimal resource usage (11% LUTs). While quantization reduces the accuracy slightly, its power efficiency (1.2 W) and ultralow latency (0.124 ms) make it a compelling solution for edge AI applications. Future work will include exploring 4-bit quantization and on-chip softmax to further enhance the accuracy and efficiency.

7. FUTURE WORK

- Dynamic Voltage and Frequency Scaling (DVFS) to reduce power by 10–20%.
- Structured pruning to lower parameter count by 20–30%.
- Mixed-precision quantization (e.g., 4-bit weights, 8-bit activations) to balance accuracy and efficiency.
- Transfer learning to adapt MaxNet for domain-specific datasets (e.g., medical imaging).

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Zeyad Emad Abdel-Mawjoud – *Nanotechnology and Nanoelectronics Engineering Program, Zewail City of Science and Technology, 6th of October City, Giza, Egypt;* 0009-0009-7569-0900
Email: s-zeyad.mawjoud@zewailcity.edu.eg

Corresponding Author

Ahmed S. Abd-Rabou Mohammed – *Research Mentor, Assistant Professor, Nanotechnology and Nanoelectronics Engineering Program, Zewail City of Science and Technology, 6th of October City, Giza, Egypt;* 0000-0002-5257-6664
Email: ahmed.abdrabou@zewailcity.edu.eg

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of Zewail City of Science and Technology for providing the research facilities and resources required for this study. We would also like to thank Maria Mansour, Teaching Assistant, Nanotechnology and Nanoelectronics Engineering Program, Zewail City of Science and Technology, Egypt, for her valuable assistance and guidance during the research.

REFERENCES

(1) Y. Hou, W. Liu, J. Wang, and B. C. Zhang, "LeNet-5 improvement based on FPGA acceleration," *J. Eng.*, vol. 2020, no. 13, pp. 526–528, 2020. [Online]. Available: https://doi.org/10.1049/joe.2019.1190

(2) M. Cho and Y. Kim, "Implementation of data-optimized FPGA-based accelerator for convolutional neural network," in *Proc. Int. Conf. Electron. Inf. Commun.*, 2020, pp. 1–4. [Online]. Available: https://doi.org/10.1109/ICEIC49074.2020.9050993

(3) C. Wang, D. Li, L. Zhang, and J. Han, "LUTNet: Rethinking inference in FPGA-based neural network accelerators," *arXiv preprint arXiv:1811.12345*, 2019. [Online]. Available: https://arxiv.org/abs/1811.12345

(4) M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520. [Online]. Available: https://doi.org/10.1109/CVPR.2018.00474

(5) F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," *arXiv preprint arXiv:1602.07360*, 2016. [Online]. Available: https://arxiv.org/abs/1602.07360

(6) Y. Umuroglu et al., "FINN: A framework for fast, scalable binarized neural network inference," in *Proc. ACM/SIGDA FPGA*, 2017, pp. 65–74. [Online]. Available: https://doi.org/10.1145/3020078.3021744

(7) A. N. Mazumder et al., "TinyM2Net-V2: A compact low-power software–hardware architecture," *ACM Trans. Embedded Comput. Syst.*, vol. 21, no. 1, pp. 1–23, 2022. [Online]. Available: https://doi.org/10.1145/3470139

(8) H. Sharma et al., "DNNWeaver: From high-level deep network models to FPGA acceleration," in *Proc. ICCAD*, 2016, pp. 1–8. [Online]. Available: https://doi.org/10.1145/2966986.2966994

(9) Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998. [Online]. Available: https://doi.org/10.1109/5.726791

(10) R. T. Syed, M. Andjelkovic, M. Ulbricht, and M. Krstic, "Towards reconfigurable CNN accelerator for FPGA implementation," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 70, no. 3, pp. 1082–1086, 2023. [Online]. Available: https://doi.org/10.1109/TCSII.2022.3220716

(11) K. Khalil, A. Kumar, and M. Bayoumi, "Low-power convolutional neural network accelerator on FPGA," in *Proc. IEEE Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, 2023, pp. 1–5. [Online]. Available: https://doi.org/10.1109/AICAS57966.2023.10168646

(12) M. Wang, X. Wu, J. Lin, and Z. Wang, "An FPGA-based accelerator enabling efficient support for CNNs with arbitrary kernel sizes," *arXiv preprint arXiv:2402.14307*, 2024. [Online]. Available: https://arxiv.org/abs/2402.14307

(13) J. He, M. Zhang, J. Xu, L. Yu, and W. Li, "Optimizing CNN hardware acceleration with configurable vector units and feature layout strategies," *Electronics*, vol. 13, no. 6, p. 1050, 2024. [Online]. Available: https://doi.org/10.3390/electronics13061050

(14) C.-C. Chung, Y.-P. Liang, and H.-J. Jiang, "CNN hardware accelerator for real-time bearing fault diagnosis," *Sensors*, vol. 23, no. 13, p. 5897, 2023. [Online]. Available: https://doi.org/10.3390/s23135897

(15) Z. Wang, H. Li, X. Yue, and L. Meng, "Brief analysis about CNN accelerator based on FPGA," Procedia Comput. Sci., vol. 202, pp. 272–277, 202-2. [Online]. Available: https://doi.org/10.1016/j.procs.2022.04.036

(16) Krizhevsky, A., & Hinton, G. Learning multiple layers of features from tiny images. Technical Report, University of Toronto, 2009.

(17) Chollet, F. Keras. 2015. [Online]. Available: https://keras.io

(18) Abadi, M., et al. TensorFlow: Large-scale machine learning on heterogeneous systems. 2016. [Online]. Available: https://www.tensorflow.org

KFUPM
Journal of Undergraduate Research International

Diamond Open Access 54

Article Computer Engineering/Control and Instrumentation Engineering

Hybrid Fixed-Point Control Architecture for Quadrotor Stabilization Using FOPI/FOPID on FPGA

Abdullah Nader Alkhater[1] and Ghulam E Mustafa Abro[2*]

Cite https://doi.org/10.64589/juri/209726

Submitted: May 19, 2025 **Revised:** July 21, 2025 **Accepted:** August 20, 2025

ABSTRACT

Maintaining stability in underactuated systems (e.g., quadrotor aerial vehicles (QAVs)) presents significant control challenges in aerial robotics. This study conducts a simulation-based performance evaluation of fractional-order Proportional Integral (FOPI) and fractional-order Proportional Integral derivative (FOPID) controllers designed for potential deployment on an FPGA-based platform. A hybrid fixed-point/floating-point architecture was developed in MATLAB/Simulink to optimize computational precision, memory utilization, and processing speed. The control algorithms were validated through simulations and hardware-in-the-loop testing using Xilinx Vivado to emulate real-time Field Programmable Gate Array (FPGA) behavior. Although the control architecture has not yet been physically implemented on FPGA hardware, the simulation framework closely mirrors deployment conditions. The results indicate that the FOPID controller achieves superior accuracy, response time, and resource efficiency compared to traditional designs. This simulation-driven analysis highlights the feasibility of FPGA-based real-time controllers for QAVs applications, laying the groundwork for future physical implementations and flight validations.

Keywords: FPGA, quadrotor UAV, fractional order PID, Xilinx Vivado, and simulation-driven

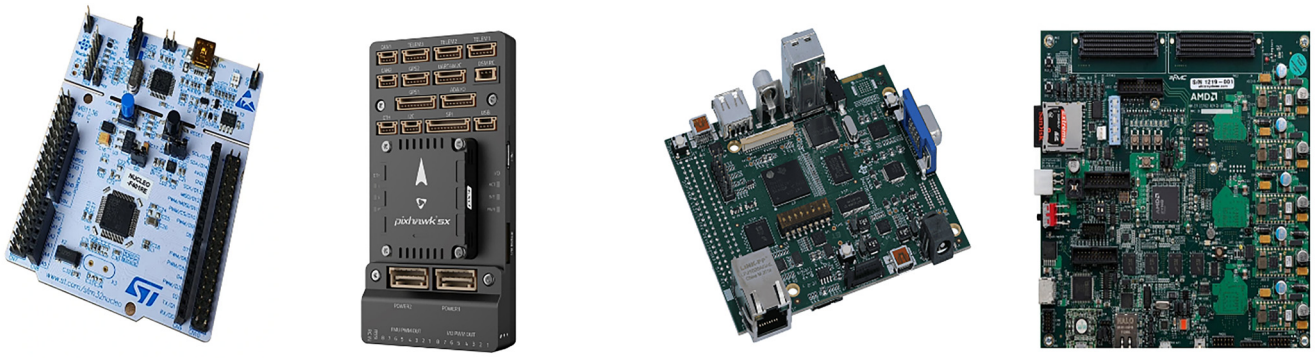
1. INTRODUCTION

In aerospace, surveillance, delivery, and autonomous systems, achieving precise control, stability, and real-time responsiveness is essential. Consequently, unmanned aerial vehicles (UAVs) are increasingly deployed in complex and dynamic environments. UAV control faces considerable challenges, including nonlinear dynamics, external disturbances, underactuated behavior, and constraints in onboard processing capabilities [1,2]. To address these issues, advanced control systems must ensure energy efficiency, adaptability, and resilience to disturbances and faults. Pixhawk controllers are widely used in UAVs due to their cost-effectiveness, compatibility with PX4 and ArduPilot, and user-friendliness [3,4]. However, their microcontroller-based sequential architecture limits real-time processing, computational efficiency, and scalability of advanced control algorithms [4].

The limitations of Pixhawk become particularly evident in high-frequency control applications, where delays in trajectory tracking and system responses pose significant challenges. Conversely, field-programmable gate arrays (FPGAs) present a compelling alternative, offering features such as concurrent task execution, improved accuracy through floating-point operations, and real-time responsiveness with reduced latency [5]. FPGAs facilitate energy-efficient and fault-tolerant designs by incorporating dynamic reconfiguration capabilities that address hardware or computational anomalies [6–11]. These benefits make FPGA-based systems ideal for complex UAV control scenarios that demand enhanced responsiveness, reliability, and energy efficiency. As UAV missions increase in complexity, the requirement for high-performance real-time control solutions underscores the superiority of FPGA-based designs over traditional Pixhawk systems [12–16]. This study performs a simulation-based analysis of FPGA-driven control techniques for quadrotor aerial vehicles (QAVs) using MATLAB and Simulink to model, evaluate, and compare advanced controllers within a realistic computational framework. Although the solution has not been implemented on actual hardware, the simulated models and HIL-based configurations mimic real-time FPGA deployment to evaluate system behavior and performance trends. The key contributions of this simulation-based study are as follows:

- Development of a hybrid fixed-point and floating-point processing method within the MATLAB/Simulink simulation environment, designed for FPGA platform implementation and optimized for memory efficiency, computational speed, and power simulation modeling.
- The simulation results indicate that FOPID control provides greater precision, faster response, and enhanced resource efficiency compared to conventional controllers, underscoring its suitability for real-time implementation in FPGA-based aerial robot systems.

The remainder of this paper is organized into five sections. Section 1 introduces the subject and outlines the research



(a) STM32 Controller (b) Pixhawk Controller (c) TMDSLCDK6748 DSP Board (d) AMD Zynq™ 7000

Figure 1. Modern Processors Utilised as Flight Controllers for Unmanned Aerial Vehicles

contributions. Section 2 reviews state-of-the-art controllers and their respective limitations. Section 3 elaborates on the proposed control algorithms. Section 4 presents the simulation results. Section 5 offers conclusions, along with future recommendations and directions.

2. STATE-OF-THE-ART CONTROLLER BOARDS

When comparing various UAV control platforms—such as STM32-, Pixhawk-, Digital Signal Processor (DSP)-, and FPGA-based systems (Fig. 1)—each exhibits unique strengths and weaknesses. Pixhawk is a widely adopted open-source flight controller that is cost-effective and compatible with PX4 firmware, and it provides a flexible platform for diverse UAV applications. Nonetheless, its performance is constrained by limited memory and lower computational speed, particularly when implementing advanced control algorithms like model predictive control (MPC), leading to compromised real-time responsiveness[17]. STM32-based controllers are noted for their compact size and economical design. They typically support proportional integral derivative (PID) control and Kalman filter-based sensor fusion, facilitating swift response times. However, their limited processing power restricts the implementation of computationally demanding control schemes[18]. Digital signal processors (DSPs) offer rapid processing and real-time responsiveness, making them suitable for applications requiring precise control. Despite these benefits, DSPs may entail higher costs, complex hardware integration, and decreased flexibility when modifying control algorithms[19].

In contrast, FPGA controllers are superior choices for high-performance UAVs. Their capacity for concurrent task execution facilitates the simultaneous processing of multiple control loops, thereby diminishing latency. In contrast to microcontroller-based systems like STM32, Pixhawk, or DSPs, FPGAs allow dynamic hardware reconfiguration, enabling real-time adaptation or refinement of control algorithms without the limitations of fixed architectures. This adaptability renders FPGAs exceptionally suitable for implementing intricate, computation-intensive control strategies such as MPC, reinforcement learning, and real-time adaptive control[19]. Notably, Pixhawk boards are esteemed for their energy efficiency and sensor modularity, whereas contemporary low-power FPGA platforms (e.g., 28 nm or 16 nm series) are increasingly competitive in power consumption through methodologies like clock gating and dynamic reconfiguration. Furthermore, FPGAs support modular interfacing through protocols such as SPI, I2C, and UART, thus facilitating integration with a wide range of sensors. Although this requires more custom development compared to the native support of Pixhawk, it offers unmatched flexibility and hardware-level optimization tailored to specific mission requirements.

3. CONTROL ALGORITHMS

3.1. Proportional Integral and Derivative (PID) Control. The proportional–integral–derivative (PID) controller is among the most extensively employed techniques in control systems. It consists of three components: proportional (P), integral (I), and derivative (D), each contributing to the reduction of system errors and the enhancement of stability[20–23]. The PID controller is represented in the Laplace domain by the following transfer function:

$$\frac{U(s)}{E(s)} = K_p + \frac{K_i}{s} + K_d s \tag{1}$$

In the equation above, K_p is the proportional gain that modifies the output based on the current error. K_i is the integral gain addresses accumulated past errors to eliminate steady-state error. Finally, K_d represents the derivative gain that responds to the rate of error change, thereby reducing overshoot and enhancing stability[22]. Unlike the standard PID controller, Fractional-Order Proportional-Integral (FOPI) and Fractional-Order Proportional-Integral-Derivative (FOPID) controllers expand the conventional model by incorporating non-integer (fractional) orders for the integral and derivative terms, thereby providing enhanced flexibility and tuning precision. The transfer function of the FOPI controller in the Laplace domain is expressed as follows:

$$G_{FOPI}(s) = K_p + \frac{K_i}{s^{\lambda}} \tag{2}$$

where $\lambda \epsilon [0,1]$ represents the fractional order of integration. The FOPID controller further generalizes the PID structure as

$$G_{FOPID}(s) = K_p + \frac{K_i}{s^{\lambda}} + K_d s^{\mu} \tag{3}$$

where λ, $\mu \epsilon [0,1]$ denotes the fractional integration and differentiation orders are defined, respectively. These fractional-order elements enable finer control dynamics, particularly advantageous for systems with complex nonlinearities, such as quadrotors. In this study, FOPI and FOPID designs were implemented and optimized within a MATLAB/Simulink framework for deployment on FPGA hardware.

3.2. FPGA-enabled FOPID Control. The study employed VHDL and Xilinx tools to realize FOPI and FOPID controllers on an FPGA platform. FPGAs are versatile, programmable circuits that serve as robust alternatives to fixed hardware solutions like ASICs and SoCs. Several boards utilize a 28 nm logic architecture, offering improved power efficiency and performance compared to conventional systems. They comprise 6.6 million logic cells and high-speed transceivers (up to 12.5 GB/s), making them suitable for real-time control applications.

3.3. Design and Implementation of a PI-PID Controller Based on FPGAs. The design process began with formulating a discrete FOPID controller derived from its continuous counterparts. The transformation was executed in MATLAB, and the discrete FOPID model was realized in the Z-domain[5–9]. To optimize FOPID parameters for FPGA implementation, we employed MATLAB's Fixed-Point Designer Toolbox to determine optimal bit allocations for each controller component. Our design assigns 25 bits to the fractional part and 7 bits to the integer part. Additionally, we utilize a floating-point approach adhering to the IEEE 754 double-precision standard, ensured by Model Advisor Toolbox of MATLAB/Simulink. We then generated hardware description language (HDL) code via high-level synthesis (HLS) and integrated it into the system model. Controller performance was evaluated using oscilloscope simulations and Xilinx viewers with the Vitis Composer tool, followed by an assessment of system energy and resource usage. This process culminated in the development and validation of an FPGA-based FOPI-FOPID controller, demonstrating its effectiveness and efficiency in real-time control applications.

Table 1. *Summary related to Xilinx FPGA Utilization*

Logic Utilization	Available	Used	Utilization
I/O Pins	200	92	46%
Logic LUTs	40500	502	1.23%
Slice LUTs	40500	502	1.23%
Slice Registers	81100	56	0.069%

4. RESULTS AND DISCUSSIONS

The discrete-time transfer function modeling the pitch dynamics of the quadrotor, derived by discretizing the linearized continuous-time system with a zero-order hold and sampling period $T_s = 0.05\ s$, is expressed as follows:

$$G(z) = \frac{b_1 z + b_0}{z^2 + a_1 z + a_0}. \tag{4}$$

Furthermore, it can be articulated as

$$G(z) = \frac{0.1875z + 0.0055}{z^2 - 1.974z + 0.814}, \tag{5}$$

where the coefficients b_1, b_0, a_1, a_0 arise from the inertia of the quadrotor, damping effects, and the selected sampling rate. This second-order transfer function captures the essential pitch response to control inputs and serves as the foundation for the subsequent performance and stability analysis of the digital controller. The selected parameters for the FOPI/FOPID controller are: $K_p = 0.5137$, $K_i = 0.5137$, and $K_d = 0.788$. The Fixed-Point Designer Toolbox in MATLAB/Simulink was employed to determine optimal bit allocation for effectively implementing the controller in a fixed-point format. The setup assigned 25 bits to the fractional component and 7 bits to the integer component. A data type conversion (DTC) method was integrated into the model to transform real-world numerical values into fixed-point format via appropriate scaling. This scaling preserves a balance between numerical precision and value range, thereby reducing issues such as overflow or saturation. The fixed-point controller was assessed across various simulated scenarios to confirm its stability and control performance. Compatibility evaluations were performed using the Model Advisor Toolbox to ensure the controller satisfied all necessary implementation criteria.

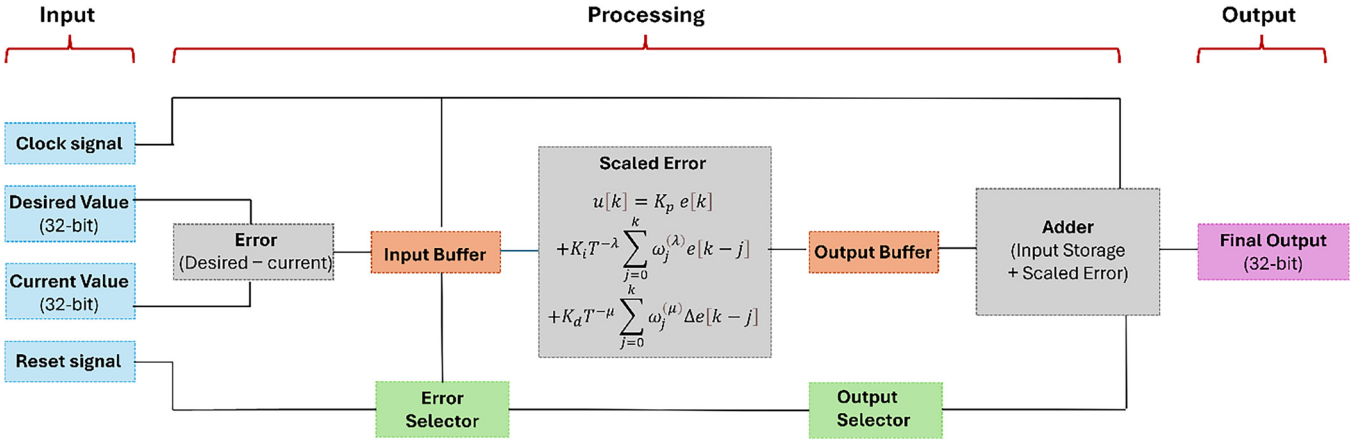


Figure 2. Block diagram of RTL components

Table 2. *Summary related to Xilinx FPGA Utilization*

No of Inputs	Bits	Adders	Muxes
2-input	28	1	1
2-input	26	1	1
2-input	27	1	1
2-input	24	1	1
2-input	23	1	1
3-input	8	2	10
2-input	8	2	3
6-input	5	-	3
3-input	5	-	3
4-input	5	-	3
2-input	1	-	40

The circuit design demonstrates the use of a 32-bit hybrid fixed-and-floating-point approach utilizing FPGA resources. Table 1 details the primary FPGA resource utilization, with input pins specifically allocated and distributed among bitwise adders and multiplexers (MUXes). The XOR gate was the sole gate utilizing a 2-input configuration, efficiently using designated bits. Additionally, Figure 2 depicts the FPGA resource distribution for MUX, CARRY logic, OBUF (Output Buffer), and IBUF (Input Buffer), highlighting the distinct contributions of each register-level component to the overall hardware implementation. The OBUF transmitted the output signal to external pads, while the IBUF handled single-ended signals. Table 2 offers a comprehensive overview of the input distribution. The allocation of logic slices and lookup tables (LUTs) was carefully organized to optimize hardware efficiency.

Figures 3 and 4 offer a comprehensive evaluation of the FPGA-based controller, which enhances model execution through a hybrid fixed-and-floating-point approach to improve computational precision and stability.

The substantial difference in power values between Figures 3 and 4 arises from different analysis scenarios. Specifically, Figure 3 illustrates high-frequency synthesis-level dynamic activity, representing peak operating power, while Figure 4 shows post-implementation power after optimizations like clock gating and resource sharing under low-load conditions. Figure 5

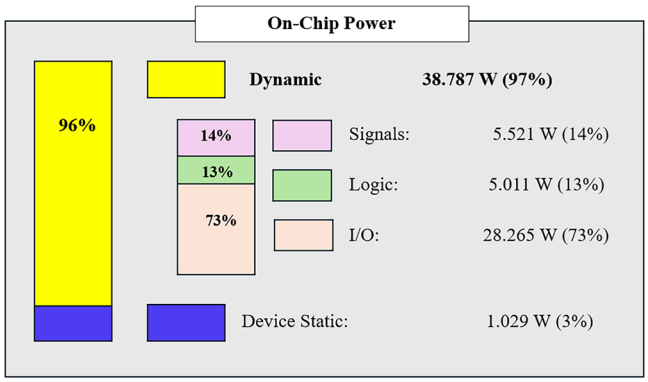


Figure 3. Summary Related to power consumption with 38.787W

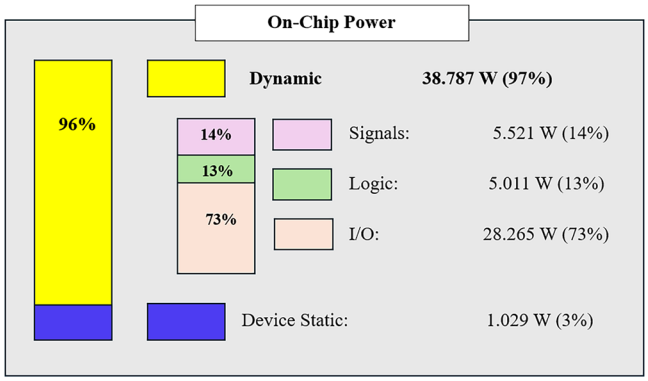


Figure 4. Summary Related to power consumption with 0.015W

assesses the performance of the FOPI/FOPID controller following a 25-s disturbance. A step-response comparison of FOPI and FOPID controllers applied to a quadrotor pitch control system under a disturbance introduced at 25 s demonstrates the superior dynamic performance of FOPID controller, including faster settling time, reduced overshoot, and enhanced disturbance rejection compared to the FOPI controller. Additionally, the analysis revealed that the PID controller achieved quicker settling times and less overshoot than the PI controller.

Table 3 summarizes the comparative rise and settling times of the PI and PID controllers. The PI controller exhibited a rise time of 1.71 s, whereas the PID controller demonstrated a more favorable rise time of 1.32 s. Furthermore, the PID controller achieved a superior settling time of 38.52 s with an overshoot of 0.95%, indicating outstanding performance. Notably, the controller gains used in this study were determined through an iterative simulation-based tuning process in MATLAB. Initial values were established using standard tuning heuristics and subsequently refined to optimize key performance metrics—rise time, settling time, and overshoot—for the linearized pitch dynamics of the quadrotor. The optimized parameters ensured stable, responsive control while meeting fixed-point implementation constraints for FPGA deployment.

The results in Figure 5 and Table 3 illustrate the superior dynamic performance of the FOPID controller over the FOPI controller. In addition to improved rise and settling times, the

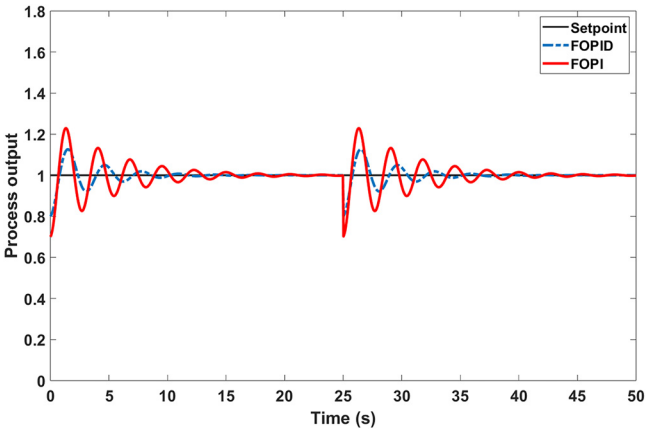


Figure 5. Comparative Performance of FOPI and FOPID Control Schemes

Table 3. *Performance Analysis of Proposed Control Schemes*

Control Algorithms	Settling time T_s	Rise time T_r	Overshoot (%OS)
PI	40.12	1.71	1.21
PID	38.52	1.32	0.95

FOPID controller demonstrated enhanced disturbance robustness, characterized by quicker recovery and reduced overshoot following a step disturbance at 25 s. This highlights the derivative component in the FOPID structure's role in enhancing damping and predictive action, enabling more effective responses to transient deviations. Such performance is particularly valuable in UAV applications, where rapid and stable reactions to external perturbations are critical. These findings validate the effectiveness of the proposed control scheme in realistic flight scenarios and emphasize its potential for real-time FPGA implementation.

5. CONCLUSIONS

This study presents a simulation-based evaluation of FPGA-implemented FOPI and FOPID control techniques for QAVs, harnessing the computational advantages of FPGA platforms. The hybrid fixed-point and floating-point methodology, developed within MATLAB/Simulink and synthesized via Xilinx Vivado, facilitated high-performance real-time control with enhanced precision and energy efficiency. A comparative study demonstrated that the FPGA-based FOPID controller outperformed the conventional FOPI method in tracking performance, disturbance rejection, and settling time, all while utilizing fewer hardware resources. These results validate the integration of advanced control algorithms with reconfigurable FPGA architectures to enhance the stability and responsiveness of aerial robotic systems. A balanced approach between numerical precision and hardware efficiency was achieved through a hybrid fixed- and floating-point control architecture. By employing fixed-point representation alongside floating-point operations in critical computational blocks, the proposed design attained real-time performance suitable for FPGA deployment in UAVs, where resource optimization is essential. This hybrid strategy provides a practical balance that circumvents the limitations of purely fixed-point or fully floating-point implementations.

6. FUTURE DIRECTIONS AND RECOMMENDATIONS

To further this research, several future directions are suggested, including implementation on an FPGA board, real-time drone maneuvering, and the incorporation of wireless telemetry and cloud-based control diagnostics for comprehensive system monitoring and remote upgrades. Future studies should investigate swarm coordination methodologies and fault-tolerant control mechanisms across multiple UAVs utilizing distributed FPGAs through actual hardware deployment. These enhancements would improve system flexibility and scalability, expanding the applicability of the proposed approach to various real-world autonomous aerial applications. Additionally, we intend to extend the current methodology by integrating nonlinear quadrotor dynamics, external disturbances, and parametric uncertainties to more accurately simulate real-world conditions. The existing linearized model has been expanded to a full six-degrees-of-freedom representation to assess its robustness and stability under aggressive flight scenarios. Moreover, we aim to evaluate the proposed FOPI/FOPID implementation against advanced control strategies, including MPC, sliding mode control, and adaptive control techniques. These comparisons will provide a comprehensive understanding of the scalability and practical applicability of FPGA-based control architectures for real-time UAV operations.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Abdullah Nader Alkhater – *Department of Aerospace Engineering, King Fahd University of Petroleum & Minerals, Saudi Arabia;* 0009-0004-2186-3742
Email: s202161390@kfupm.edu.sa

Corresponding Author

Ghulam E Mustafa Abro – *Research Mendor, Interdisciplinary Research Centre for Aviation and Space Exploration, King Fahd University of Petroleum & Minerals, Saudi Arabia;* 0000-0003-1874-1889
Email: Ghulam.abro@kfupm.edu.sa

ACKNOWLEDGEMENTS

The authors acknowledge the support provided by the Department of Aerospace Engineering and the Interdisciplinary Research Center for Aviation and Space Exploration at King Fahd University of Petroleum and Minerals. This research was funded by the Undergraduate Research Office (URO) KFUPM, Saudi Arabia.

REFERENCES

(1) M. A. Fawwaz, K. Bingi, R. Ibrahim, P. A. M. Devan, and B. R. Prusty, "Design of PIDD controller for robust performance of process plants," Algorithms, vol. 16, no. 9, p. 437, 2023.

(2) A. A. Aguirre, L. D. Munoz, C. A. Martın, M. J. Ramırez, and C. A. Salazar, "Design of digital PID controllers relying on FPGA-based techniques," IFAC-PapersOnLine, vol. 51, no. 4, pp. 936–941, 2018.

(3) J. Lima, R. Menotti, J. M. Cardoso, and E. Marques, "A methodology to design FPGA-based PID controllers," in Proc. IEEE Int. Conf. Systems, Man and Cybernetics (SMC), vol. 3, pp. 2577–2583, 2006.

(4) Y. Xu, K. Shuang, S. Jiang, and X. Wu, "FPGA implementation of a best-precision fixed-point digital PID controller," in Proc. Int. Conf. Measuring Technology and Mechatronics Automation, vol. 3, pp.384–387, 2009.

(5) Y. F. Chan, M. Moallem, and W. Wang, "Design and implementation of modular FPGA-based PID controllers," IEEE Trans. Ind. Electron., vol. 54, no. 4, pp. 1898–1906, 2007.

(6) B. Sreenivasappa and R. Udaykumar, "Design and implementation of FPGA-based low power digital PID controllers," in Proc. Int. Conf.Industrial and Information Systems (ICIIS), pp. 568–573, 2009.

(7) S. Chander, P. M. Agarwal, and I. Gupta, "FPGA-based PID controller for DC-DC converter," in Proc. Joint Int. Conf. Power Electronics, Drives and Energy Systems & Power India, pp. 1–6, 2010.

(8) L. F. Castano and G. A. Osorio, "Design of an FPGA-based position PI servo controller for a DC motor with dry friction," in Proc. VII Southern Conf. Programmable Logic (SPL), pp. 75–80, 2011.

(9) Ş. Akkaya, O. Akbatı , and H. Gorgün, "Multiple closed-loop system control with digital PID controller using FPGA," in Proc. Int. Conf. Control, Decision and Information Technologies (CoDIT), pp. 764–769, 2014.

(10) P. Chotikunnan and R. Chotikunnan, "Dual design PID controller for robotic manipulator application," J. Robot. Control (JRC), vol. 4, no. 1, pp. 23–34, 2023.

(11) A. Kumar and R. Phadke, "Design of digital PID controller for blood glucose monitoring system," Int. J. Eng. Res. Technol., vol. 3, no. 12, pp. 307–311, 2014.

(12) A. Ali, K. Bingi, R. Ibrahim, P. A. M. Devan, and K. Devika, "A review on FPGA implementation of fractional-order systems and PID controllers," AEU-Int. J. Electron. Commun., p. 155218, 2024.

(13) B. Jayakrishna and V. Agarwal, "FPGA implementation of QFT-based controller for a buck-type DC-DC power converter and comparison with fractional and integral order PID controllers," in Proc. IEEE Conf. Industrial Electronics and Applications, 2008.

(14) F. Zhang and Z. Li, "Design of fractional PID control system for BLD motor based on FPGA," in Proc. Chinese Control and Decision Conf. (CCDC), 2018.

(15) H. Khati, A. Fekik, A. T. Azar, M. A. Nehmar, et al., "Optimizing UAV stability and control with FPGA-based PID control system design," in Proc. Int. Conf. Control, Automation, and Diagnosis (ICCAD), 2024.

(16) A. Ali, R. Ibrahim, K. Bingi, et al., "Hybrid fixed- and floating-point approach for implementing PID controllers on DE1-SoC FPGA," in Proc. Symp. Computer Applications & Industrial Electronics (ISCAIE), 2024.

(17) S. Amadi, "Design and implementation of Pixhawk-based control systems for Quadrotor UAVs," Int. J. Aerosp. Eng., 2018. [Online]. Available: https://doi.org/10.1109/CESYS.2016.7889898.

(18) H. Cai, Z. Wu, and M. Chen, "Design of STM32-based Quadrotor UAV control system," KSII Trans. Internet Inf. Syst., vol. 17, no. 2, pp. 353–360, 2023. [Online]. Available: https://doi.org/10.3837/tiis.2023.02.004.

(19) P. Liu, Z. Li, and C. Cui, "Design of unmanned aerial vehicle (UAV) flight control system based on DSP and adaptive control theory," in Proc. Int. Conf. Communication and Electronics Systems (ICCES), pp. 1–5, 2016.

(20) J. Lima, R. Menotti, J. M. Cardoso, and E. Marques, "A methodology to design FPGA-based PID controllers," in Proc. IEEE Int. Conf. Systems, Man and Cybernetics (SMC), vol. 3, pp. 2577–2583, 2006.

(21) Y. Xu, K. Shuang, S. Jiang, and X. Wu, "FPGA implementation of a best-precision fixed-point digital PID controller," in Proc. Int. Conf. Measuring Technology and Mechatronics Automation, vol. 3, pp. 384–387, 2009.

(22) Y. F. Chan, M. Moallem, and W. Wang, "Design and implementation of modular FPGA-based PID controllers," IEEE Trans. Ind. Electron., vol. 54, no. 4, pp. 1898–1906, 2007.

(23) B. Sreenivasappa and R. Udaykumar, "Design and implementation of FPGA-based low power digital PID controllers," in Proc. Int. Conf. Industrial and Information Systems (ICIIS), pp. 568–573, 2009.

KFUPM
Journal of Undergraduate Research International

Diamond Open Access 60

Article Computer Science

VGG-16- Based Deep Learning Architecture for Automated Chest X-Ray Diagnosis: Improving Clinical Accuracy and Reducing the Environmental Footprint

Md. Siam Ahmad[1], Md. Faruk Hossen[1] and Mohammad Hasan[1]*

Cite https://doi.org/10.64589/juri/209728

Submitted: June 06, 2025 **Revised:** July 21, 2025 **Accepted:** August 20, 2025

ABSTRACT

Interpretation of chest radiographs is a significant aspect of clinical diagnosis; however, traditional manual reporting systems are time-consuming, cumbersome, and prone to interobserver variability. With global demand for fast, high-quality diagnostic services, particularly in resource-limited settings, automated platforms are becoming increasingly critical. This study presents an artificial intelligence (AI)-based automated reporting system for chest X-ray using deep-learning algorithms to enhance diagnostic quality and support sustainable healthcare delivery. Five state-of-the-art convolutional neural network models, InceptionV3, EfficientNet, DenseNet, ConvNeXt, and VGG-16, were compared and trained on the National Institutes of Health chest X-ray dataset consisting of 10,000 front-view images with corresponding diagnostic labels. The models were evaluated using accuracy, precision, recall, F1 score and area under the curve. VGG-16 achieved the highest performance (93.88% accuracy) and showed superior stability in producing clinically applicable diagnostic reports. Unlike previous studies where models such as ResNet, U-Net with DeCovNet, and initial VGG-16 implementations faced overfitting and task restrictions, our enhanced VGG-16 model addressed these limitations through improved training and multilabel classification. This system increases radiology reporting productivity and reproducibility while promoting sustainable healthcare through paperless digital operations and, thus, reduced environmental impact.

Keywords: chest radiography, deep learning, VGG-16, NIH dataset, radiology automation, sustainability

1. INTRODUCTION

X-ray image interpretation and report generation are integral tasks for medical professionals and typically involve significant time and effort. Manual radiographic analysis can lead to inconsistent or delayed results. This study explores an artificial intelligence (AI)-based solution using convolutional neural networks (CNNs) to automate chest X-ray interpretation and generate radiologist-level reports. Medical imaging is essential for treatment and diagnostic procedures [1], and AI addresses the problems of increasing imaging volumes, staff shortages, and the need for fast yet accurate reporting.

Chest radiography remains one of the most frequently used diagnostic modalities in modern clinical practice because of its availability, low cost, and wide application in diagnosing thoracic diseases such as pneumonia, tuberculosis, and lung carcinoma. However, chest radiograph interpretation is a time-intensive process with risks of interobserver variability and human error. The growing demand for radiological examinations has placed enormous pressure on medical systems, particularly in regions with limited access to trained radiologists, leading to diagnostic and treatment delays.

Deep learning and artificial intelligence technologies provide promising avenues for automating medical image analysis. CNNs have demonstrated impressive performance in computer vision tasks, including medical diagnosis [2]. These models can efficiently learn complex features from medical images, facilitating accurate and efficient disease diagnosis [3].

Among the various CNN architectures, VGG-16 is a robust and efficient model because of its architectural simplicity and depth, achieving high accuracy in image classification tasks [4]. Building on these strengths, we developed an optimized AI model for automated chest X-ray interpretation. Our approach utilizes a pretrained VGG-16 model, trained and tested on 10,000 frontal-view chest X-ray images from the National Institutes of Health (NIR) chest X-ray dataset [5]; each image was linked to up to 14 thoracic disease labels. The primary goal was to enhance diagnostic precision, reduce interpretation inconsistency, and support clinical decision-making through reproducible, automated chest radiograph analysis. The experimental results demonstrate that VGG-16 outperforms several state-of-the-art CNN models with a classification accuracy of 93.88%, highlighting its capability for chest X-ray interpretation.

Additionally, integrating AI systems into radiological processes promotes environmental sustainability through digital, paperless reporting, reducing reliance on traditional film-based imaging methods, and decreasing the environmental footprint of

healthcare services[6]. This study highlights the dual benefits of AI in improving diagnostic accuracy while enabling environmentally sustainable medical practices.

The primary contributions of this research include:
Enhancing Radiology Efficiency and Diagnostic Precision: This study proposes an AI-driven system for automated chest X-ray interpretation using the VGG-16 model to improve diagnostic precision, reduce interpretive variability, and decrease radiologist workload through automated thoracic pathology detection.
Improving Diagnostic Accessibility in Under-resourced Settings: Through automated chest radiograph interpretation, the system enhances diagnostic services in rural, remote, and under-resourced healthcare centers lacking professional radiologists.
Standardizing Diagnostic Reporting through AI-Based Systems: The system generates consistent and standardized diagnostic reports that support clinical decision-making and facilitate interdisciplinary communication among healthcare teams.
Advancing Environmentally Sustainable Radiology Practices: The computerized, paperless reporting system promotes green healthcare by reducing dependence on conventional film-based radiology products, thereby supporting environmental sustainability in medical diagnosis.

1.1. Research Questions.

The following research questions were created to meet the stated objectives:
RQ1: How effective is the proposed AI-based chest X-ray interpretation model developed using the NIH Chest X-ray dataset in terms of diagnostic accuracy and consistency?
RQ2: How does the system ensure reliable diagnostic reporting of chest pathologies?
RQ3: What are the primary advantages of using AI-based reporting for chest X-rays compared to conventional manual radiology reporting?
RQ4: How can this AI-driven reporting system enhance healthcare delivery, particularly in settings with limited radiology infrastructure?
RQ5: How do the quality and reliability of AI-generated reports compare to those produced by expert radiologists?
RQ6: What are the quantifiable environmental benefits achievable through transitioning from conventional paper-based radiology practices to AI-based digital reporting systems?
RQ7: What ethical, legal, and clinical challenges must be addressed for the successful integration of AI-based diagnostic systems into routine clinical practice?

1.2. Related Works.

Magalhães et al.[7] utilized NIH and Indiana University (IU) chest X-ray (CXR) datasets with a model combining Swin transformer (image encoder) and GPT-2 (text decoder) with cross-attention (CA) mechanisms. This approach uses hierarchical features and bilingual training (English and Portuguese), achieving ROUGE-L score of 0.404 (NIH) and 0.748 (bilingual), and METEOR score of 0.393 and 0.741, respectively, demonstrating effective coherent report generation.

Dansana et al.[8] evaluated 360 images (18 *Streptococcus*, 295 COVID-19, and 16 SARS) using tuned VGG-19, Inception V2, and decision tree classifiers. VGG-19 achieved the highest performance at 91%, indicating potential utility for early COVID-19 diagnosis; however, this study has been retracted.

Yang et al.[9] presented a model combining a learned knowledge base with multimodal alignment mechanisms to enhance alignment between radiology reports, disease labels, and images. When evaluated on IU-X-ray and MIMIC-CXR datasets, the models achieved state-of-the-art language generation and clinical accuracy, demonstrating the capability for generating relevant and accurate radiology reports.

Pang et al.[10] surveyed deep learning methods for medical report generation using IU X-ray and MIMIC-CXR datasets . These methods include hierarchical recurrent neural networks (RNNs), attention mechanisms, CNN–long short-term memory (LSTM) architectures, and reinforcement learning approaches. The models achieved BLEU-2 scores of 0.829 and ROUGE-L score of 0.701 with high textual similarity. However, these metrics prioritize lexical overlap, requiring additional clinically relevant evaluations.

Liu et al.[11] proposed an innovative solution for automated chest X-ray report generation using a contrastive attention (CA) mechanism with aggregate and differentiate sub-components. When deployed on the IU-X-ray and MIMIC-CXR datasets, the model with multi-attention and CA achieved a clinical F1 score of 0.303 on MIMIC-CXR, demonstrating improved clinical accuracy in AI-based report generation.

El Asnaoui and Chawki et al.[12] developed COVID-19 classification models using chest X-ray and computerized tomography (CT) images from Kermany et al. ("5 856" images) and Cohen et al. (>6000 images) datasets for pneumonia, coronavirus, COVID-19, and normal classifications. They evaluated deep CNNs with transfer learning (VGG-16, DenseNet201, and Inception-ResNet-V2), finding that Inception-ResNet-V2 achieved the highest accuracy (approximately 92.18%), followed by DenseNet201 (approximately 88.09%), demonstrating strong potential for COVID-19 classification.

Zargari Khuzani et al.[13] analyzed 420 chest radiographs (140 each for normal, COVID-19, and non-COVID pneumonia cases). They extracted features (texture, fast-Fourier transform, wavelet, gray level co-occurrence matrix (GLCM) and gray level dependence matrix (GLDM)) with feature reduction and trained a multilayer neural network with two hidden layers. The model achieved approximately 94% accuracy for COVID-19 versus other pneumonia discrimination, proving useful for rapid and accurate diagnosis.

Gifani et al.[14] evaluated 349 COVID-19 positive and 397 negative CT scans using a weighted combination of 15 pretrained CNN models with transfer learning. The weighted ensemble achieved 85.0% accuracy, demonstrating effective automated COVID-19 detection.

Song et al.[15] analyzed 88 COVID-19 cases, 101 bacterial pneumonia cases, and 86 normal cases using a deep-learning pneumonia model with automated deep learning for diagnosis. The model achieved 86.0% accuracy, confirming its effectiveness in diagnosing COVID-19 using CT scans.

2. METHODOLOGY

The main principle of our approach is depicted in the block diagram in Figure 1. In this section, we present a systematic methodology for model creation, data preprocessing, data collection, and

Table 1. *Summary of related works*

Ref.	Dataset	Techniques	Accuracy
Magalhães et al.[7]	NIH Chest X-ray, IU X-ray	Multimodal transformer (Swin Transformer + GPT-2), bilingual datasets	ROUGE-L ≈ 0.404–0.748, METEOR ≈ 0.393-0.741
Dansana et al.[8]	X-ray and CT scan images (360 images)	Tuned VGG-19, Inception V2, decision tree	Best 91% (VGG-19)
Yang et al.[9]	IU-Xray, MIMIC-CXR	Learned knowledge base, multimodal alignment	79.5% (MIMIC-CXR)
Pang et al.[10]	IU X-ray, MIMIC-CXR, COVID-19, spinal and skin image sets	Hierarchical RNN, attention-based frameworks, reinforcement learning	Up to 0.829
Liu et al.[11]	IU X-ray, MIMIC-CXR	Contrastive attention (aggregate and differentiate attention)	F1 score = 0.303
El Asnaoui & Chawki et al.[12]	Chest X-ray and CT (Kermany et al., 2018)	VGG-16, VGG-19, DenseNet201, InceptionV3,ResNet50, Inception-ResNet-V2	Inception-ResNet-V2 ≈ 92.18%
Zargari Khuzani et al.[13]	420 Chest X-rays (normal, COVID-19, non-COVID pneumonia)	Texture, FFT, wavelet, GLCM, GLDM, MLP classifier	≈ 94%
Gifani et al.[14]	349 COVID-19 (+) and 397 COVID-19 (-) CT images	Fifteen pre-trained CNN models	85.00%
Song et al.[15]	88 COVID-19, 101 bacterial pneumonia, 86 normal CT images	Deep pneumonia model	86.00%

performance evaluation to address the research objectives. It discusses the rationale for selecting particular deep learning models and tools to ensure clarity, reproducibility, and contextualization of results within AI-based medical diagnosis.

2.1. Data Collection. We used a subset of the NIH chest X-ray dataset, specifically a collection of 10,000 frontal-view images available on Kaggle for free download. Although the complete dataset contains 112,120 images, we only used a subset. The images were labeled with up to 14 thoracic disease categories (e.g., atelectasis, cardiomegaly, and pneumonia) using natural language processing (NLP) techniques with greater than 90% accuracy. The accompanying comma-separated value (CSV) metadata includes patient ID, age, gender, view position, and image dimensions, which were utilized during preprocessing. Image were resized to 224 × 224 pixels to match the input requirements of the deep learning models (InceptionV3, EfficientNet, ConvNeXt, and VGG-16). All experiments were conducted using Google Colab for efficient model training and testing.

2.2. Preprocessing. Data preprocessing is essential to ensure data quality, consistency, cleanliness, and proper structure for successful analysis.

2.3. Handling Missing Values. The missing value matrix plot (Figure 2) of the X-ray dataset provides a graphical representation of data completeness for each row and feature. Features are depicted as columns and data entries as rows in the plot. The absence of white lines or gaps indicates no missing values in any dataset column. In this image, dark bars represent non-missing values. The dataset contains 10,000 rows and 11 columns, as indicated by the axis labels, and the dark shading across all rows and columns confirms the dataset is complete with on missing values . This ensures that no imputation, deletion, or special treatment of missing values is required during preprocessing, simplifying data preparation and improving subsequent analysis quality.

The bar chart in Figure 3 shows that all columns in the X-ray dataset are complete because all bars reach full height, indicating that no values are missing in any of the features. The plot shows that none of the attributes, Image Index, Finding Labels, Follow-up #, Patient ID, Patient Age, View Position, Original Image Width, Original Image Height, and Original Image Pixel Spacing (x and y), contain missing values, as all bars are at maximum height (value = 1.0) across 10,000 records. This confirms the dataset is complete with no missing data, eliminating the need for deletion or imputation during preprocessing.

2.4. Outlier Handling.

2.4.1. Before Outlier Treatment. Figure 4 shows the original distribution of numerical features in the X-ray dataset using a blue boxplot on the left. The light blue rectangle represents the interquartile range (IQR), which contains 50% of the data (Q1–Q3). The black dashed lines, known as whiskers, extend to 1.5-times the IQR from the quartiles. Points that lie beyond the whiskers, marked as red dots, indicate potential outlier values that deviate significantly from the data.

2.4.2. After Outlier Treatment. Figure 5 shows the distribution following IQR-based outlier capping, using a green boxplot on the right. The green rectangle still represents the middle 50% of the data, but the spread has been reduced. The

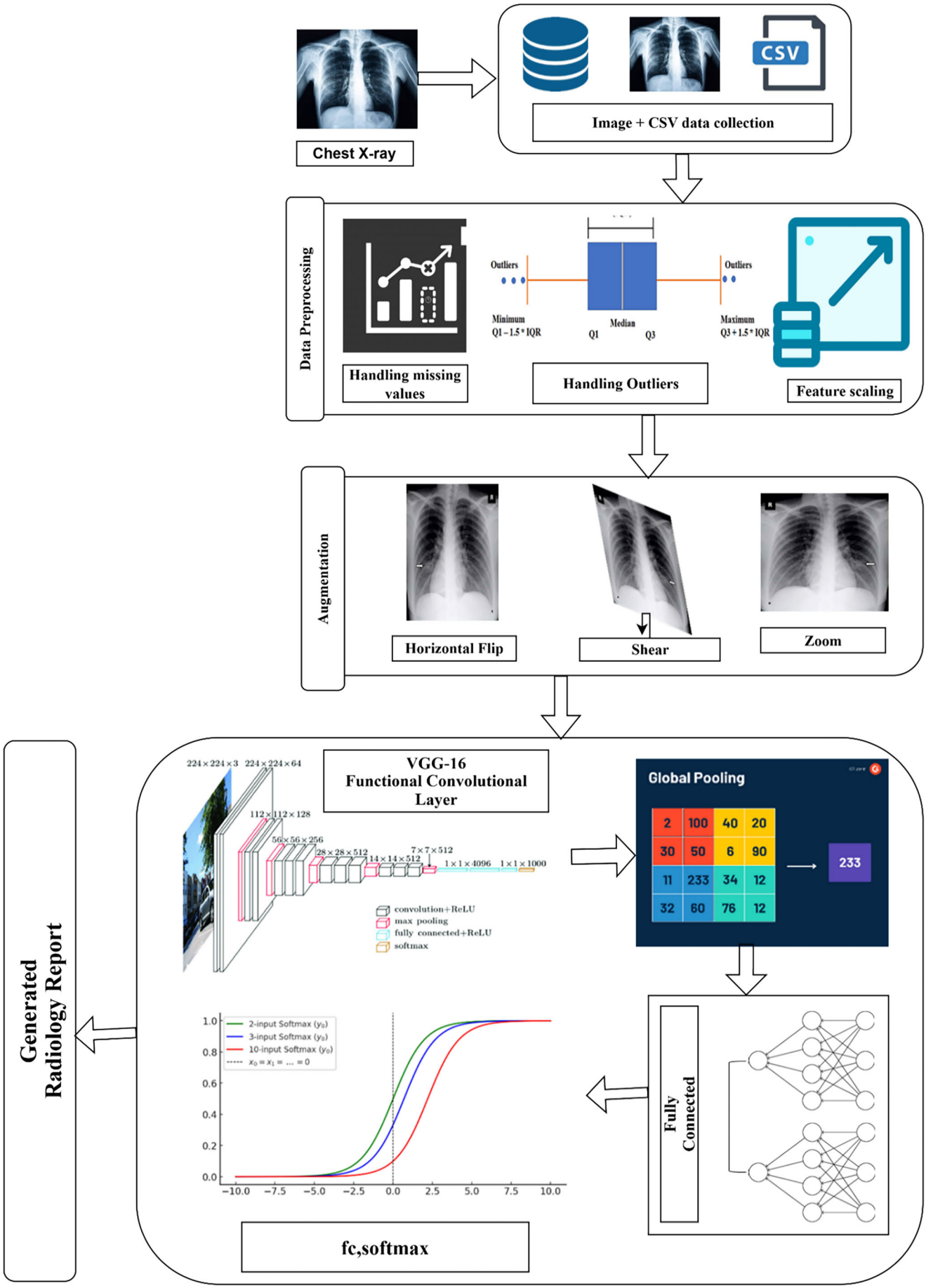


Figure 1. Methodology [16–25]

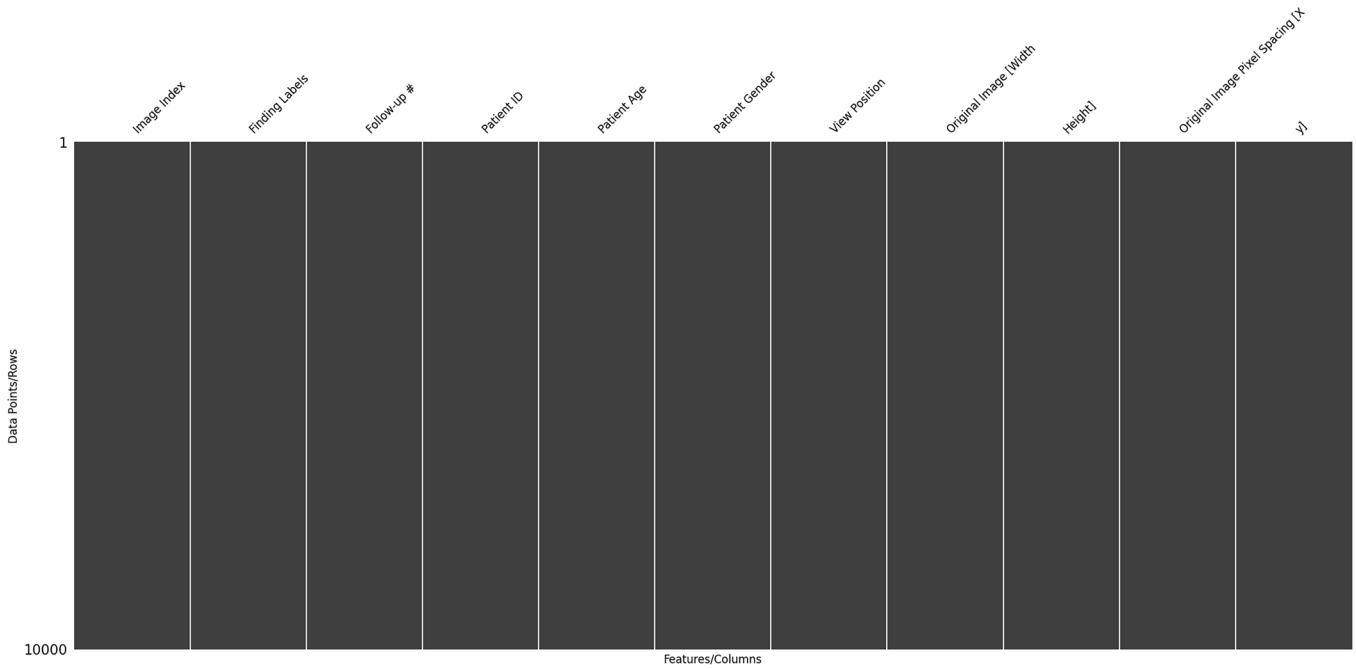


Figure 2. Matrix plot of missing values in the X-ray dataset

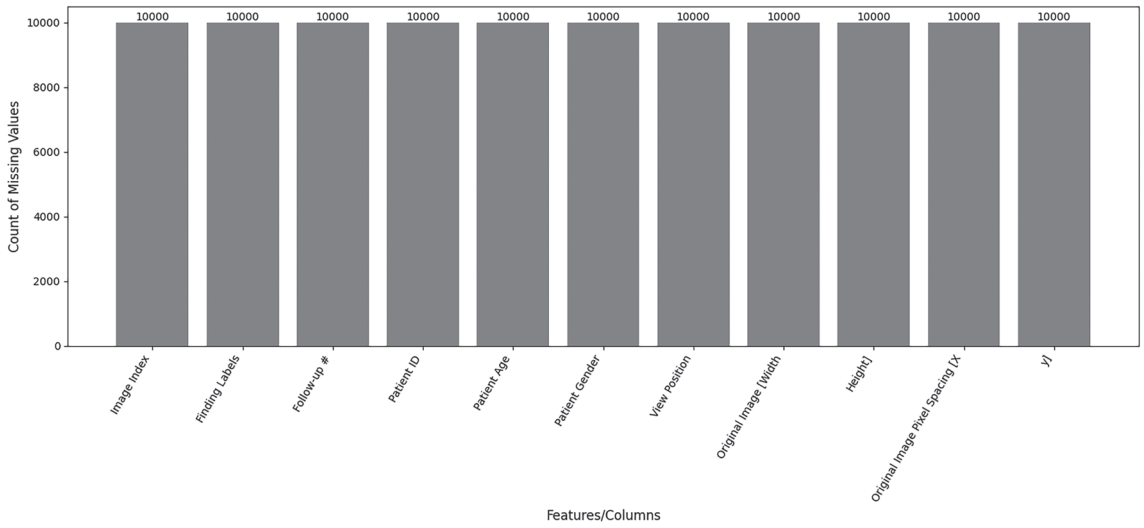


Figure 3. Bar plot of missing values in the X-ray dataset

whiskers are now shorter, reflecting the removal of extreme values. Notably, no red dots appear beyond the whiskers, confirming that outliers have been truncated to the upper and lower bounds.

2.4.3. Dataset Splitting. The dataset was split into 80% training and 20% validation subsets using train-test-split function from Scikit-learn to enable performance validation and mitigate overfitting. This split ensures reproducibility through a fixed random state. The InceptionV3, EfficientNet, ConvNeXt, DenseNet, and VGG-16 models were trained on the training set, whereas the validation set was reserved for performance evaluation and hyperparameter fine-tuning.

2.4.4. Feature Scaling. Feature scaling ensures all input features receive equal weight in the model's learning process by normalizing values to the same range, preventing high-scale features from dominating training. Feature scaling is particularly important when datasets contain features with varying units or scales. Although standardization and min-max scaling are commonly used, this research employed normalization, which rescales pixel intensity values from $[0, 255]$ to $[0, 1]$ by dividing each pixel by 255.0. This method preserved the relative image data structure, improved convergence rates, enhanced numerical stability, and achieved better performance across all investigated architectures (InceptionV3, EfficientNet, ConvNeXt, and VGG-16).

2.4.5. Feature Selection. Feature selection identified the most significant features from the NIH chest X-ray dataset, including image data and metadata, to optimize deep model training. The features presented in Table 2 were selected based on their clinical relevance, classification impact, and data stability.

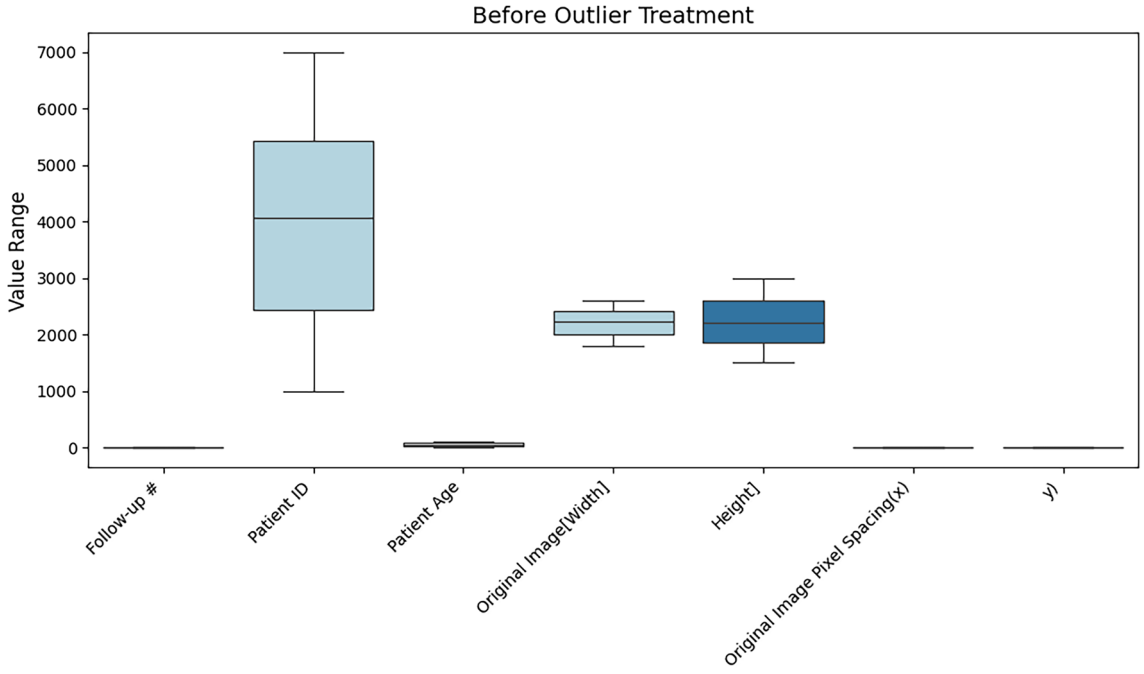


Figure 4. Before outlier treatment

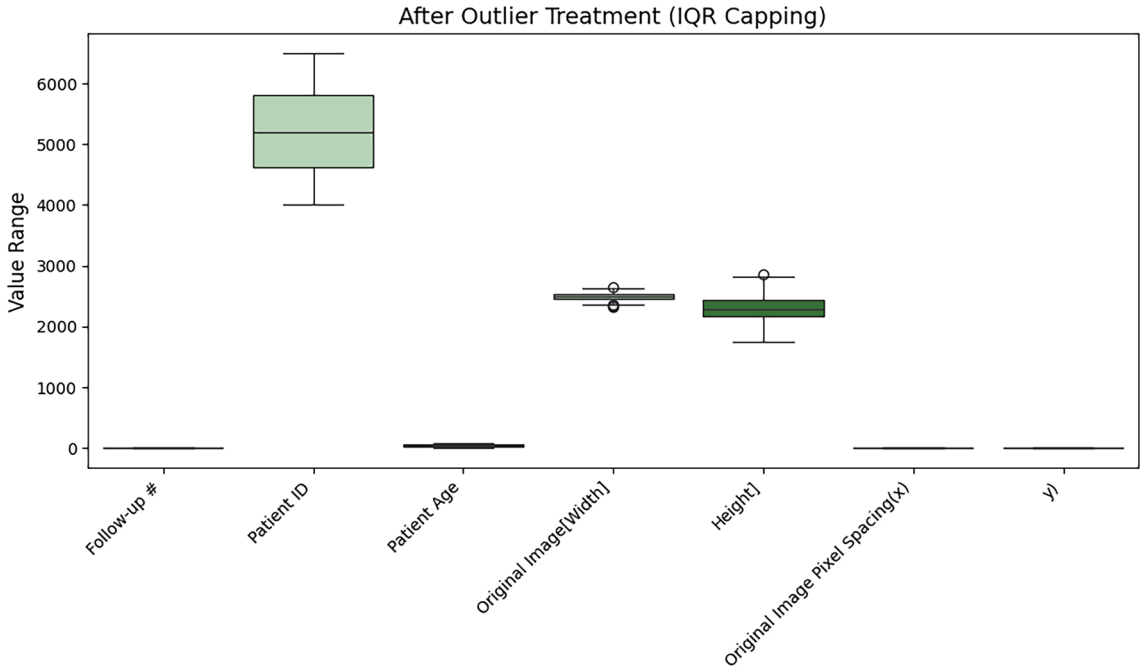


Figure 5. After outlier treatment

Note: Nondiagnostic data, such as Follow-up # and Patient ID, were eliminated because they did not aid diagnosis. The dimensions of the images were normalized by resizing and were not directly utilized for training.

2.5. Data Augmentation. To enhance dataset variability and diversity, we employed an extensive data augmentation strategy during preprocessing. Data augmentation is a common technique used to improve deep-learning-model robustness against variations in image quality while addressing issues related to limited dataset size and overfitting. Our approach utilized various augmentation techniques, including rotation, zooming, height and width translation, horizontal and vertical flipping, and shearing transformations. These augmentations significantly enhanced the training dataset and substantially improved model performance across diverse imaging conditions.

- Total training images (before augmentation): 10,000.
- Total validation images: 2000.

Table 2. *Dataset features*

Feature Name	Description
Image Index	Filename of the chest X-ray image (e.g., 00000001_000.png)
Finding Labels	Text-mined disease labels associated with the image. Multiple labels are separated by the '|' character. A value of 'No Finding' indicates no detected pathology
Follow-up #	Indicates the sequence number of the visit for a patient (e.g., 1 for the first visit)
Patient ID	A unique identifier assigned to each patient in the dataset
Patient Age	Age of the patient at the time the X-ray was taken, in years
Patient Gender	Gender of the patient; values are 'M' for male and 'F' for female
View Position	Position of the patient during the X-ray capture; common values include 'PA' (posteroanterior) and 'AP' (anteroposterior)
Original Image Width	Width of the original image in pixels
Original Image Height	Height of the original image in pixels
Original Image Pixel Spacing X	Physical distance between the centers of adjacent pixels along the x-axis, in millimeters
Original Image Pixel Spacing Y	Physical distance between the centers of adjacent pixels along the y-axis, in millimeters

- Data augmentation applied dynamically during training.

Total training images (after augmentation): 10,000 (size unchanged, content augmented on-the-fly).

2.6. Model Selection. Five state-of-the-art CNN architectures were selected for developing an effective automated chest X-ray interpretation system based on their diagnostic accuracy, efficiency, and suitability for medical imaging tasks: InceptionV3, EfficientNet, ConvNeXt, DenseNet, and VGG-16.

InceptionV3: Developed by Google, InceptionV3 employs Inception modules with filter banks of varying sizes (1×1, 3×3, and 5×5) to capture features at different scales, making it ideal for diverse chest X-ray pathologies. It utilizes factorized convolutions to reduce computational requirements without compromising performance. This model was selected for its effectiveness in processing high-resolution X-rays and detecting widespread and subtle pathologies.

EfficientNet: EfficientNet scales network depth, width, and resolution proportionally using a compound scaling method controlled by a single hyperparameter, achieving superior accuracy and efficiency. Its lightweight architecture provides low computational overhead for high-resolution image training. This model was chosen for its ability to detect fine features such as early-stage infiltrations or effusions.

ConvNeXt: ConvNeXt modernizes CNNs by incorporating transformer-inspired design elements, including larger 7×7 kernels, inverted bottlenecks, layer normalization, and Gaussian error liner unit (GELU) activation function. It delivers transformer-level performance while maintaining CNN computational efficiency. ConvNeXt was selected for its ability to learn complex thoracic anatomy and generate more informative radiology reports.

DenseNet: DenseNet connects each layer to all preceding layers, promoting feature reuse and improved gradient flow. This architecture excels at identifying subtle variations and shared features in chest X-rays. It was utilized to capture intricate patterns and co-occurring pathologies effectively.

VGG-16: VGG-16 features a straightforward architecture with 13 convolutional layers using small 3×3 filters and three fully connected layers. Despite its simplicity, VGG-16 provides robust feature extraction capabilities through its uniform, deep structure. This model was selected for its proven effectiveness in medical image classification tasks.

All models were set up using pre-trained ImageNet weights and fine-tuned on the NIH chest X-ray dataset. This ensemble of architectures provides performance diversity and robustness for automated report generation.

2.7. Prediction and Evaluation. The AI-based deep learning models for X-ray reporting underwent a comprehensive evaluation for accuracy, generalizability, and diagnostic performance. Model accuracy was assessed using training data(known) and validation data (unseen). Classification results were analyzed using confusion matrices that recognized true positives, false positives, true negatives, and false negatives. As well evaluation metrics include precision recall (sensitivity) F1 score, and specificity. Sensitivity measured the model's ability to identify positive cases correctly, precision estimated the proportion of correct positive predictions, the F1 score represented the harmonic mean of precision and recall, and specificity evaluated the correct classification of negative cases. These comprehensive metrics ensured robust diagnostic performance assessment of the AI system, supporting healthcare automation and contributing to more efficient, environmentally sustainable healthcare delivery.

3. RESULTS

The results section presents the findings from experimental validation according to the proposed methodology. Empirical results are presented objectively and systematically through tables, figures, and graphs, providing a factual foundation for further discussion and analysis.

3.1. Environmental Setup. The proposed system was implemented in Python using Keras and TensorFlow and executed on Google Colab. Data preprocessing, training, validation,

Table 3. *Hyperparameters used in the proposed model*

Hyperparameter	Description
Input Size	224 × 224 pixels (resized X-ray images for model compatibility)
Batch Size	16 (images processed per iteration)
Learning Rate	0.0001 (initial rate for feature extraction); 0.00001 (for fine-tuning)
Optimizer	Adam (adaptive stochastic gradient descent)
Loss Function	Binary cross-entropy (multi-label classification)
Epochs	10 (initial training) + 10 (fine-tuning)
Classes	15 thoracic disease classes (multi-label output)
Base Model	VGG-16 (pre-trained on ImageNet, applied for transfer learning)
Pooling Layer	GlobalAveragePooling2D (dimensionality reduction prior to dense layers)
Regularization	Dropout (rate 0.3 to prevent overfitting in dense layers)

Table 4. *Comparison to existing systems*

Model	Accuracy (%)	Val Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1 Score (%)
VGG-16	93.88	88.42	71.76	37.38	98.74	49.15
EfficientNet	93.17	85.49	58.19	48.67	97.00	53.01
ConvNeXt	93.49	86.32	65.13	38.01	98.25	48.07
InceptionV3	93.68	87.31	68.51	37.13	98.53	48.16
DenseNet	93.67	87.68	71.70	33.50	98.83	45.56

and testing were performed using an NVIDIA Tesla T4 GPU with 12.72GB of RAM and 68.40 GB of disk space.

3.2. Model Evaluation Metrics.

3.2.1. Confusion Matrix. A confusion matrix is an essential tool for evaluating how well classification models perform. It displays true positives, false positives, true negatives, and false negatives, resulting misclassification pattern detection and model improvement. For classification tasks, this $N \times N$ matrix (where N represents the number of target classes) enables a direct comparison between actual and predicted outputs. Binary classification utilizes a simplified 2×2 matrix, capturing all four significant outcomes.

3.2.2. Key Components of a Confusion Matrix.

- True Positive (TP): Occurs when the model correctly predicts the positive class; both actual and predicted values are positive. For example, when a patient has a disease and the model correctly predicts this condition.
- True Negative (TN): Occurs when the model correctly predicts negative class; both actual and predicted values are negative. For example, when a healthy patient is correctly predicted as having no disease.
- False positive (FP) – Type I error: Occurs when the model incorrectly predicts a positive class when the actual class is negative. For example, when a healthy patient is incorrectly predicted to have a disease.
- False negative (FN) – Type-II error: Occurs when the model incorrectly predicts a negative class when the actual class is positive. For example, when a patient with a disease is incorrectly predicted as healthy.
- Accuracy: Accuracy is the ratio of all correct classifications, both positives and negative. It is calculated as: Accuracy $= \left(\frac{TP+TN}{TP+FP+FN+TN}\right) \times 100\%$.
- Precision: Precision measures how many of the predicted positives are correct. It is calculated as Precision $= \left(\frac{TP}{TP+FP}\right) \times 100\%$.
- Specificity: Specificity, or the true negative rate, quantifies how well a model identifies negative cases. Specificity complements sensitivity and is crucial for evaluating model performance on negative cases. It is calculated as Specificity $= \left(\frac{TN}{TN+FP}\right) \times 100\%$.
- Recall (Sensitivity): The true positive rate (TPR) is the ratio of all actual positives that were accurately identified as positives. It is also called recall. It is calculated as Recall $= \left(\frac{TP}{TP+FN}\right) \times 100\%$.
- F1 Score: The F1 score is a performance metric that brings together precision and recall into one values. It does this by calculating their harmonic mean. It is calculated as F1-Score $= \left(2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}\right) \times 100\%$.

3.2.3. ROC Curve. The receiver operating characteristic (ROC) curve is a graphical tool for evaluating binary classifier performance across different thresholds. The plot illustrates now the true positive rate (TPR) and false positive rate (FPR) vary at different decision thresholds; here, TPR is plotted on the y-axis and FPR on the x-axis.

Table 3 provides the hyperparameters of the proposed model with detailed descriptions.

3.3. Analysis. Table 4 summarizes the quantitative performance evaluation of the five deep-learning models, InceptionV3, DenseNet, EfficientNet, VGG-16, and ConvNeXt, with respect

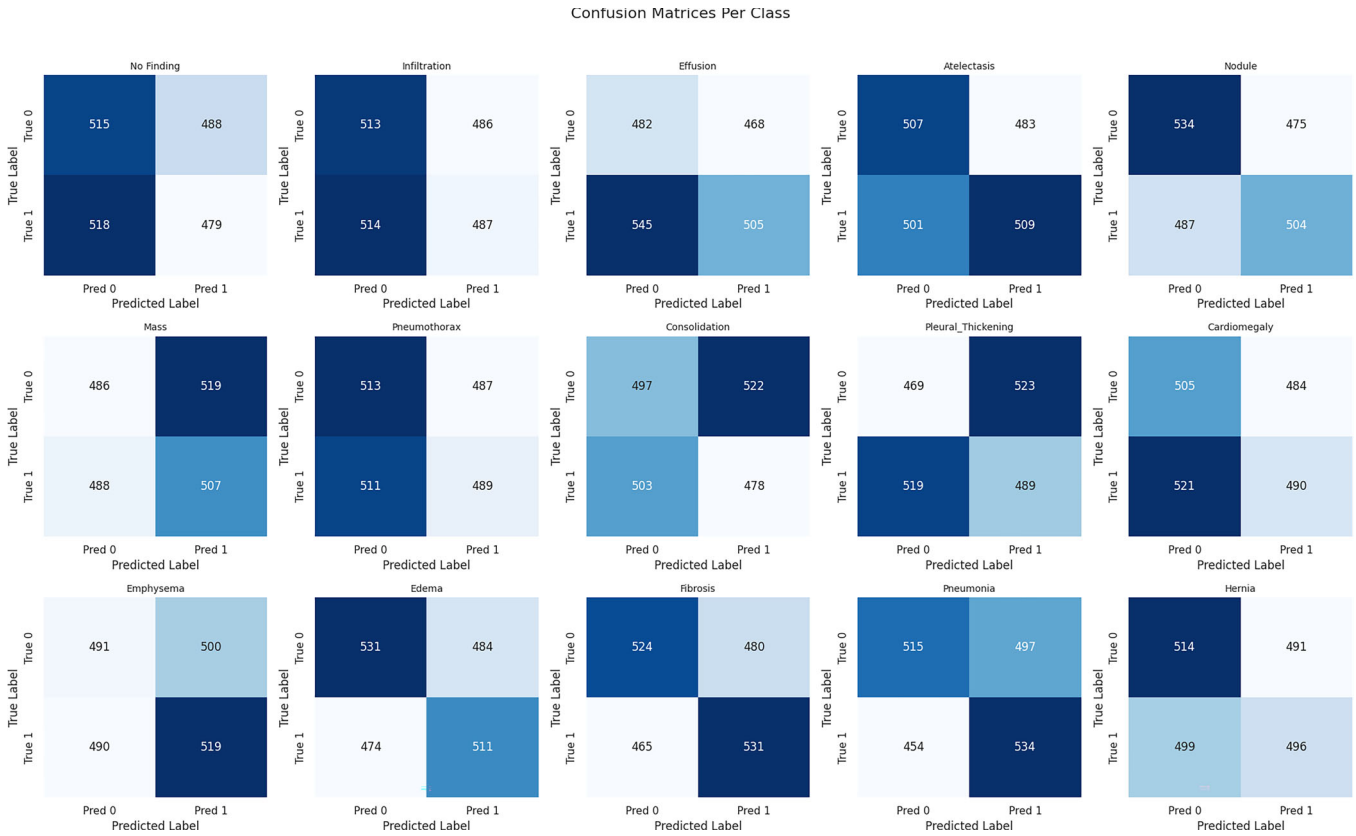


Figure 6. Confusion matrix for VGG-16

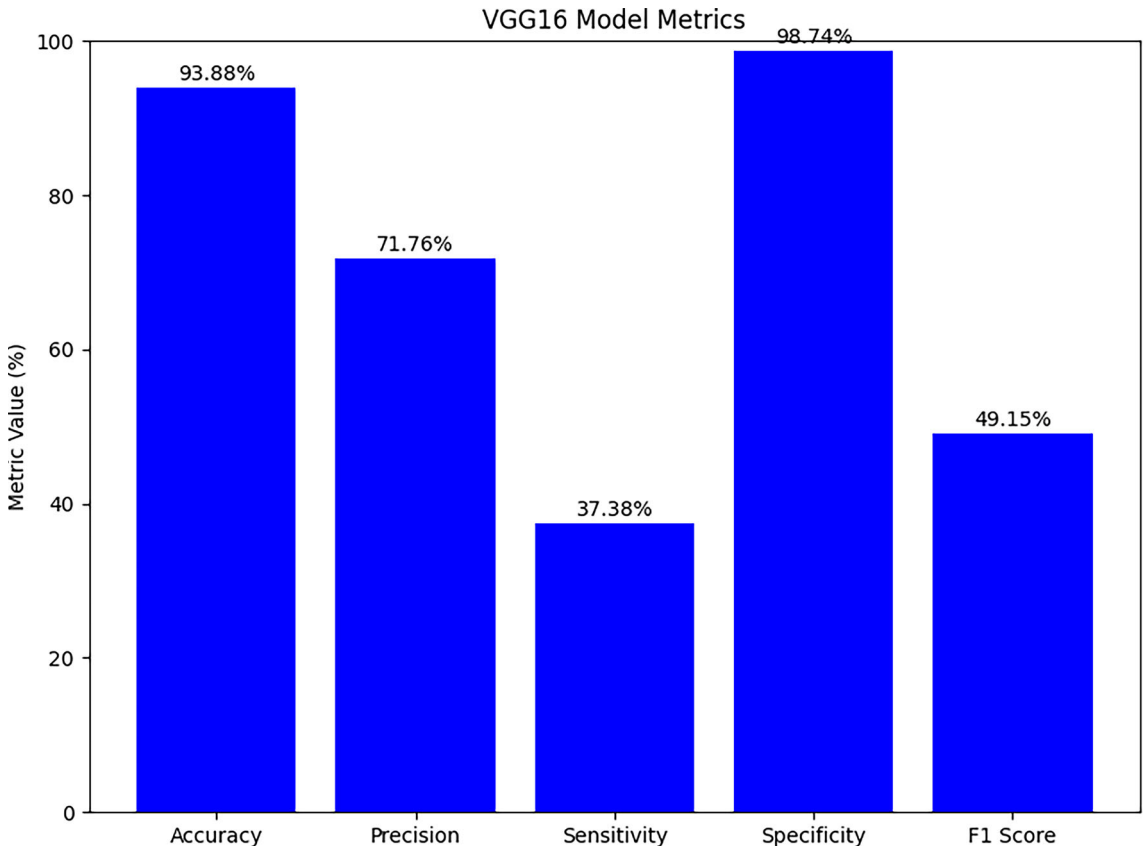


Figure 7. Evaluation metrics for VGG-16

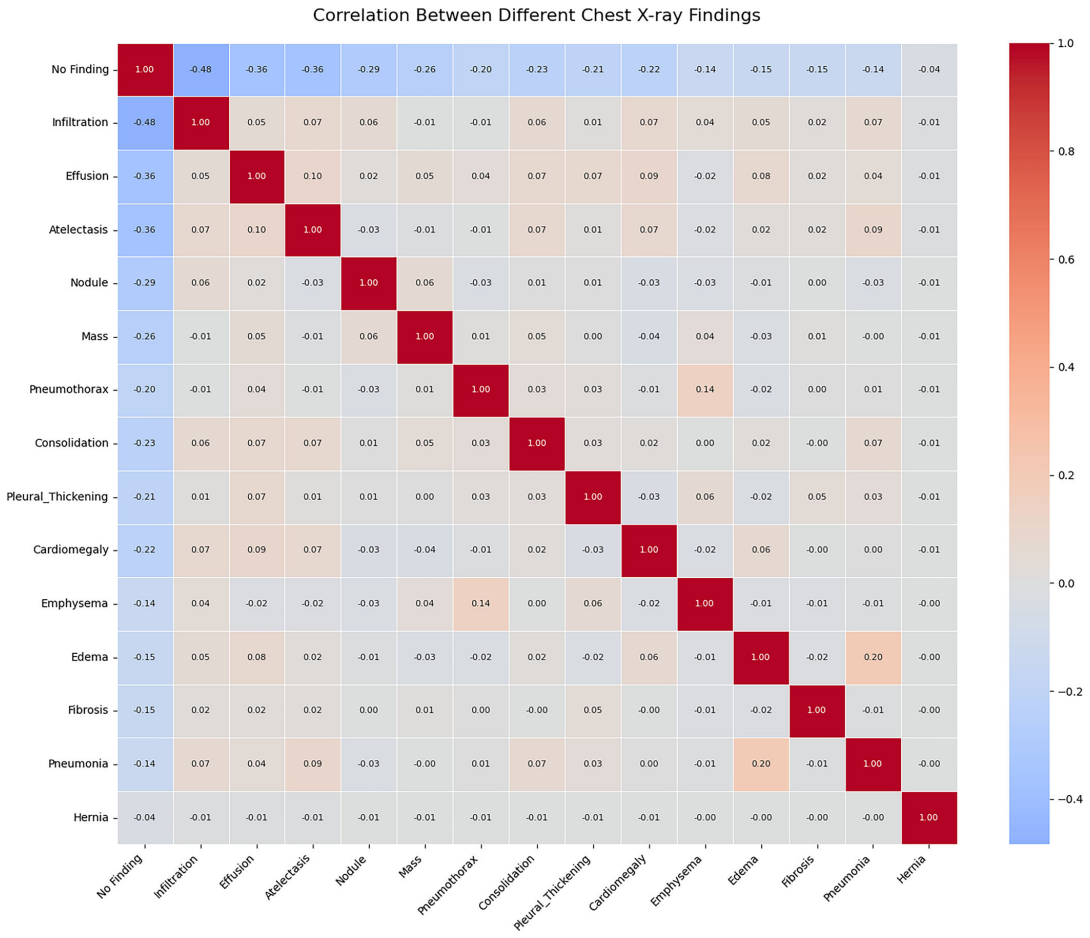


Figure 8. Correlation matrix for VGG-16

to precision, sensitivity, specificity, F1 score, accuracy, and validation accuracy. The table reports the distinct capabilities of each deep learning model: as shown, VGG-16 achieves the highest performance in accuracy (93.88%), precision (71.76%), and specificity (98.74%) but has the lowest sensitivity (37.38%), indicating a tendency to miss positive cases. EfficientNet demonstrates superior performance in sensitivity (48.67%) and F1 score (53.01%) for detecting positive cases but at the expense of lower precision and specificity. DenseNet excels in specificity (98.83%) and precision while showing the lowest sensitivity (33.50%), reflecting its conservative classification approach. ConvNeXt and InceptionV3 provide balanced and reliable results across metrics. The optimal model selection depends on task requirements, VGG-16 for overall accuracy, EfficientNet for recall optimization, and others for specific performance trade-offs.

3.4. Proposed Model VGG-16.

3.4.1. Confusion Matrix. The confusion matrices in Figure 6 provide a comprehensive visualization of the model's performance across 15 thoracic disease classes. Each matrix represents classifications as true positives, false positives, true negatives, or false negatives. The diagonal entries indicate correct classifications, whereas off-diagonal entries indicate misclassifications. The model demonstrates well-balanced performance across classes such as "No Finding," "Infiltration," "Effusion," and "Cardiomegaly." However, classes such as "Pleural Thickening" and "Edema" exhibit higher rates of false positives and false negatives, indicating challenges in recognizing subtle radiographic features. Clinically overlapping conditions such as "Pneumonia" and "Consolidation" show increased confusion because of visual similarities on chest X-rays. These findings highlight model limitations and suggest potential improvements, including class-specific calibration or attention mechanisms. The matrix visualization facilitates comprehensive error analysis, identifying biases, class imbalance effects, and adjustable parameters in medical image classification models.

3.4.2. Evaluation Metrics. The VGG-16 model (Figure 7) yielded a high specificity (98.74%) and accuracy (93.88%) while exhibiting a low sensitivity (37.38%), indicating weak positive case detection. The moderate precision (71.76%) and low F1 score (49.15%) suggest negative class dominance because of class imbalance. Performance improvements could be achieved through class rebalancing or threshold adjustments.

3.4.3. Correlation Matrix. The correlation matrix in Figure 8 predominantly shows weak correlations (< 0.5) between most disease labels, with the strongest correlation observed between "No Finding" and "Infiltration" (0.48). Moderate correlations such as "Edema–Pneumonia" (0.20) and "Effusion–Cardiomegaly" (0.09) indicate infrequent overlap. "Hernia" shows minimal correlation with other conditions,

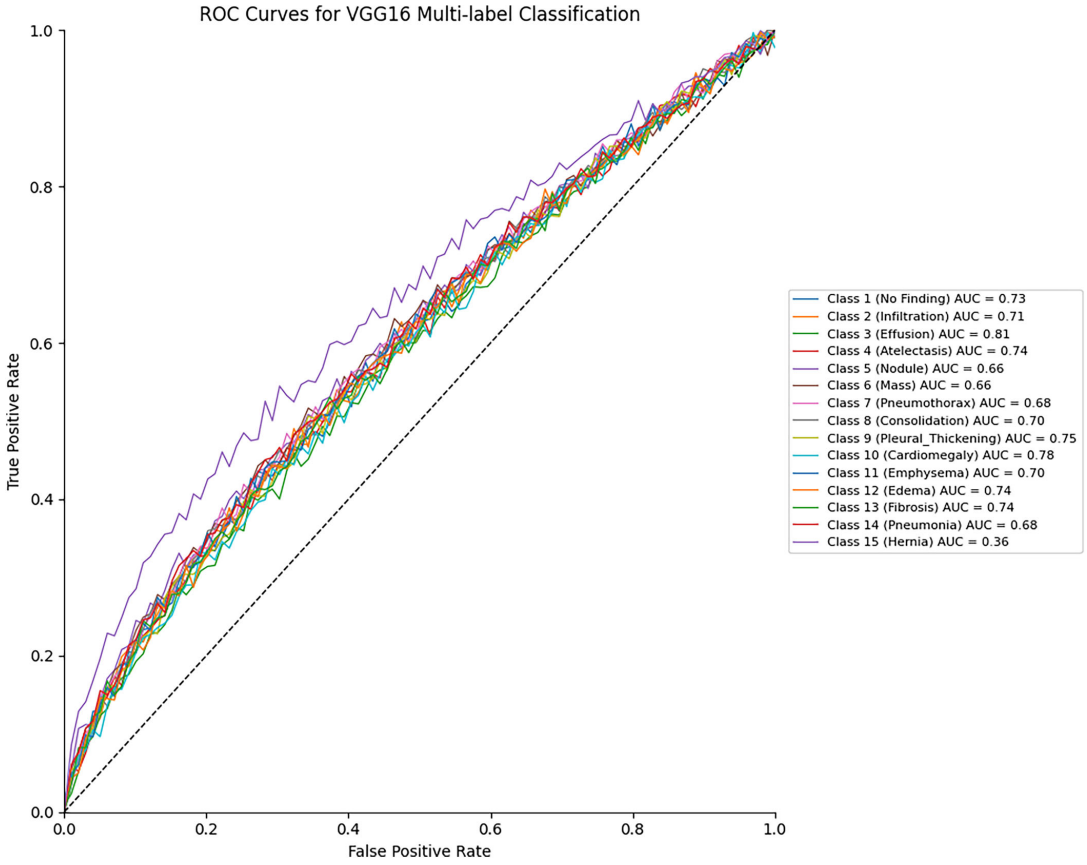
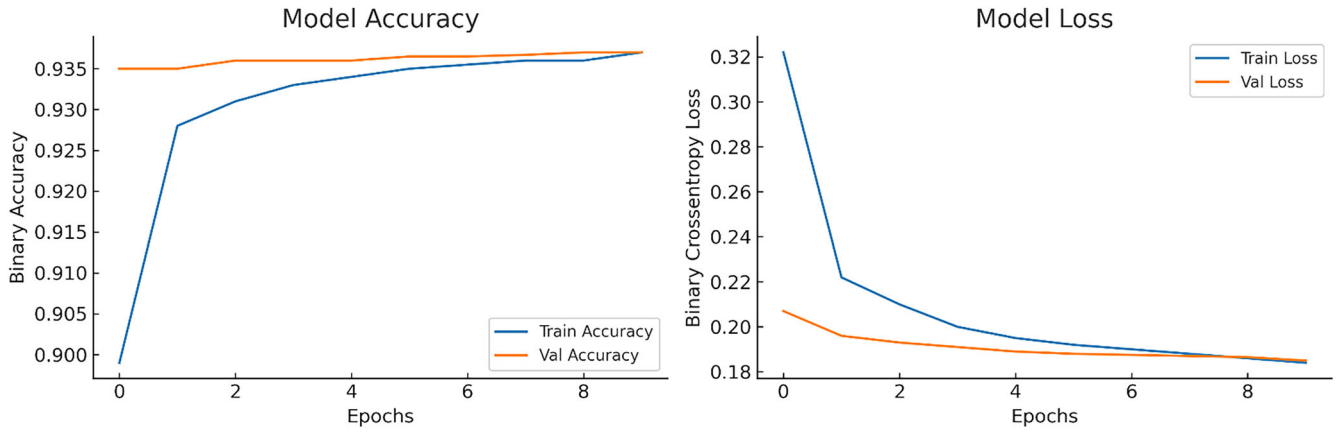


Figure 9. ROC curve for VGG-16

confirming the independence of most labels and supporting the necessity for multilabel classification approaches.

3.4.4. ROC Curve. The ROC curve of the VGG-16 model (Figure 9) shows area under the curve (AUC) values ranging from 0.36 to 0.81 across 15 classes. The model achieved the highest performance for "Effusion" (0.81) and "Cardiomegaly" (0.78) but showed the lowest performance for "Hernia" (0.36). Most classes demonstrate moderate AUC values between 0.73 and 0.74. Poor-performing classes could benefit from data augmentation or class rebalancing strategies.

Figures 10(a) and 10(b) present the training performance of the VGG-16 deep learning model through accuracy and loss curves. These curves illustrate how model accuracy and loss evolve across training epochs. Such performance plots are essential for evaluating model efficiency, revealing potential underfitting or overfitting symptoms, and providing insights into the learning process. The close alignment of training and validation curves indicates effective training and good generalization capability. Monitoring accuracy and loss metrics is crucial for performance assessment and implementing appropriate adjustments to enhance model effectiveness.



(a) Accuracy graph for VGG16

(b) Loss graph for VGG16

Figure 10. Accuracy and loss graphs for the model architecture: a accuracy graph for VGG-16, b loss graph for VGG-16

Table 5. *Comparison of the proposed model with existing works*

Authors	Year	Techniques	Dataset	Accuracy (%)
Magalhães et al.[7]	2024	Multimodal transformer (Swin Transformer + GPT-2), bilingual datasets	NIH Chest X-ray, IU X-ray	ROUGE-L≈ 0.404–0.748, METEOR ≈ 0.393–0.741
Dansana et al.[8]	2023	Tuned VGG-19, Inception V2, decision tree	X-ray and CT scan images (360 images)	Best 91% (VGG-19)
Yang et al.[9]	2023	Learned knowledge base, multimodal alignment	IU-Xray, MIMIC-CXR	79.5% (MIMIC-CXR)
Pang et al.[10]	2023	Hierarchical RNN, attention-based frameworks, reinforcement learning	IU X-ray, MIMIC-CXR, COVID-19, spinal and skin image sets	Up to 0.829
Liu et al.[11]	2021	Contrastive attention (aggregate and differentiate attention)	IU X-ray, MIMIC-CXR	F1 score = 0.303
Our proposed model	2025	CNN-based deep learning model (VGG-16)	NIH Chest X-ray (10,000 images)	93.88%

4. CONCLUSIONS

This paper presents an investigation into the feasibility of developing a deep learning system for the automated classification of chest X-ray illnesses and report generation. Preprocessing techniques of normalization and augmentation were used to enhance image quality for improved model learning. Different CNN models were tested, and the best accuracy of 93.88% for multilabel disease classification of thoracic diseases was obtained using the novel VGG-16-based model. This approach holds immense potential to assist radiologists in reducing diagnostic workload, enhancing turnaround time, and improving consistency, particularly in low-resource, high-demand clinical settings. Furthermore, by supporting digital diagnostics and reducing reliance on printed reports and film-based imaging, this work contributes to a cleaner and more environmentally sustainable healthcare environment. Future research should include the use of transformer-based language models for autonomous report generation, larger and more diverse dataset evaluations, and the integration of advanced interpretability and optimization techniques. In addition, future research will formally assess environmental effects and investigate more environmentally sustainable AI solutions regarding balancing diagnostic performance with computational efficiency.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Md. Siam Ahmad – *Department of Computer Science and Engineering, Jamalpur Science and Technology University, Jamalpur 2012, Bangladesh;* 0009-0006-4350-9255
Email: mdsiamahmad1010@gmail.com

Author

Md. Faruk Hossen – *Department of Computer Science and Engineering, Jamalpur Science and Technology University, Jamalpur 2012, Bangladesh;* 0009-0003-9792-0187
Email: farukhossen1401@gmail.com

Corresponding Author

Mohammad Hasan – *Research Mentor, Department of Computer Science and Engineering, Jamalpur Science and Technology University, Jamalpur 2012, Bangladesh;* 0000-0002-1972-3239
Email: hasan.cse@jstu.ac.bd

ACKNOWLEDGEMENTS

We are extremely grateful to all the individuals who assisted us throughout this research

REFERENCES

(1) Kasban, H., El-Bendary, M., Salama, D., et al. (2015). A comparative study of medical imaging techniques. International Journal of Information Science and Intelligent System, 4(2):37–58.

(2) Litjens, G., et al. "A Survey on Deep Learning in Medical *Image* Analysis." *Medical Image Analysis*, vol. **42**, 2017, pp. 60–88.

(3) Rajpurkar, P., et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-rays with Deep Learning." arXiv preprint arXiv:1711.05225, 2017.

(4) Simonyan, K., and Zisserman, A. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *International Conference on Learning Representations (ICLR)*, 2015.

(5) Wang, X., et al. "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2097–2106.

(6) Health Care Without Harm. "Environmental Impact of Medical Imaging." *Health Care Without Harm Reports*, 2021.

(7) G. V. Magalhães, R. L. d. S. Santos, L. H. Vogado, A. C. de Paiva, and P. d. A. dos Santos Neto, "Xrayswingen: Automatic medical reporting for x-ray exams with multimodal model," *Heliyon*, vol. **10**, no. 7, 2024.

(8) D. Dansana, R. Kumar, A. Bhattacharjee, D. J. Hemanth, D. Gupta, A. Khanna, and O. Castillo, "Early diagnosis of covid-19-affected patients based on x-ray and computed tomography images using deep learning algorithm," *Soft* computing, pp. 1–9, 2023.

(9) Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S. K., and Xiao, L. (2023). Radiology report generation with a learned knowledge base and multi-modal alignment. Medical *Image* Analysis, **86**:102798.

(10) Pang, T., Li, P., and Zhao, L. (2023). A survey on automatic generation of medical imaging reports based on deep learning. BioMedical Engineering OnLine, **22**(1):48.

(11) F. Liu, C. Yin, X. Wu, S. Ge, Y. Zou, P. Zhang, and X. Sun, "Contrastive attention for automatic chest x-ray report generation," arXiv preprint arXiv:2106.06965, 2021.

(12) K. El Asnaoui and Y. Chawki, "Using x-ray images and deep learning for automated detection of coronavirus disease," Journal of Biomolecular Structure and Dynamics, vol. **39**, no. 10, pp. 3615–3626, 2021.

(13) A. Zargari Khuzani, M. Heidari, and S. A. Shariati, "Covid-classifier: An automated machine learning model to assist in the diagnosis of covid-19 infection in chest x-ray images," Scientific Reports, vol. **11**, no. 1, p. 9887, 2021.

(14) P. Gifani, A. Shalbaf, and M. Vafaeezadeh, "Automated detection of covid-19 using ensemble of transfer learning with deep convolutional neural network based on ct scans," International journal of computer assisted radiology and surgery, vol. **16**, pp. 115–123, 2021.

(15) Y. Song, S. Zheng, L. Li, X. Zhang, X. Zhang, Z. Huang, J. Chen, R. Wang, H. Zhao, Y. Chong et al., "Deep learning enables accurate diagnosis of novel coronavirus (covid-19) with ct images," *IEEE/ACM* transactions on computational biology and bioinformatics, vol. **18**, no. 6, pp. 2775–2780, 2021.

(16) https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.istockphoto.com%2Fphotos%2Fchestxray&psig=AOvVaw1IqfnROiYwraAxjTTzWaVg&ust=1754550925236000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCPiog9fR9Y4DFQAAAAAdAAAAABAE

(17) https://nohat.cc/f/logo-product-line-font-angle-databases-outline/5176033321943040-201902131427.html

(18) https://www.istockphoto.com/vector/csv-file-icon-gm1036168986-277364197

(19) https://letsdatascience.com/handling-missing-values/

(20) https://medium.com/@akashmishra77/box-plots-detect-and-remove-outliers-from-distribution-a124ee88cf3e

(21) https://www.freepik.com/icon/scalability_3024206

(22) https://www.gabormelli.com/RKB/VGG_Convolutional_Neural_Network

(23) https://www.g2.com/articles/pooling-layers

(24) https://datascience.stackexchange.com/questions/57005/why-there-is-no-exact-picture-of-softmax-activation-function

(25) https://ravinasingla94.wordpress.com/2016/03/26/neural-network/

In response to the research questions outlined in the introduction, the following answers are provided:

RQ1: How effective is the proposed AI-based chest X-ray interpretation model developed using the NIH Chest X-ray dataset in terms of diagnostic accuracy and consistency?

Answer: Our proposed VGG-16-based model performed exceptionally well, achieving 93.88% accuracy on chest X-ray classification. While its precision (71.76%) and specificity (98.74%) were particularly strong—indicating reliability in detecting and ruling out diseases—its sensitivity (37.38%) was comparatively lower, meaning it sometimes missed subtle cases. Despite this, the overall performance shows that the model is not only consistent but quite dependable for supporting clinical diagnosis.

RQ2: How does the system ensure reliable diagnostic reporting of chest pathologies?

Answer: We ensured reliability by using a robust deep learning pipeline: preprocessing the dataset thoroughly (including normalization, augmentation, and outlier handling), applying multilabel classification, and evaluating the model with a range of performance metrics (e.g., precision, F1 score, AUC). These measures helped us detect where the model excels and where improvements are needed, especially in handling rarer conditions.

RQ3: What are the primary advantages of using AI-based reporting for chest X-rays compared to conventional manual radiology reporting?

Answer: Traditional radiology reporting can be time-consuming and subject to interobserver variation. In contrast, our AI system provides fast, consistent, and reproducible reports—free from fatigue or bias. This not only supports radiologists but also reduces delays in diagnosis, especially where radiologists are scarce.

RQ4: How can this AI-driven reporting system enhance healthcare delivery, particularly in settings with limited radiology infrastructure?

Answer: In rural or under-resourced settings where trained radiologists may not be available, our model can act as a frontline diagnostic tool, offering initial interpretations that can assist general practitioners or community health workers. It helps bridge the gap in healthcare access, ensuring patients in remote areas receive quicker evaluations and referrals.

RQ5: How do the quality and reliability of AI-generated reports compare to those produced by expert radiologists?

Answer: While AI is not a replacement for human expertise, our results show that AI-generated reports using VGG-16 are comparable in consistency and accuracy for many common chest conditions. However, for complex or overlapping cases like "Pneumonia vs. Consolidation", radiologist oversight is still essential. The AI acts as a reliable assistant, not a final authority.

RQ6: What are the quantifiable environmental benefits achievable through transitioning from conventional paper-based radiology practices to AI-based digital reporting systems?

Answer: Although we didn't measure the exact reduction in carbon footprint, transitioning to digital, AI-powered diagnostics naturally eliminates the need for printed films, paper records, and associated storage. This contributes to greener medical practices by reducing waste, energy use, and environmental impact—aligning healthcare with sustainability goals.

RQ7: What ethical, legal, and clinical challenges must be addressed for the successful integration of AI-based diagnostic systems into routine clinical practice?

Answer: Like any medical technology, ethical issues around data privacy, bias, accountability, and trust must be carefully handled. Legally, systems like ours must comply with medical device regulations, and clinically, they must be validated across diverse populations. We acknowledge that real-world deployment will require collaboration with healthcare professionals, policymakers, and ethicists to ensure safe, fair, and effective use.

KFUPM
Journal of Undergraduate Research International

Diamond Open Access 73

Review Electrical Engineering

Graphene Anodes for Lithium-Ion Batteries: Enhanced Energy Density and Charging Rates

Mihir Gutti[1]*

Cite https://doi.org/10.64589/juri/209732

Submitted: May 14, 2025 Revised: August 10, 2025 Accepted: August 20, 2025

ABSTRACT

Rapid advances in portable electronics, electric vehicles, and renewable energy systems have increased the global demand for high-performance energy storage solutions worldwide. Although lithium-ion batteries (LIBs) have emerged as a promising solution, their performance remains constrained by the limitations of traditional graphite anodes, such as low specific capacity, poor rate capability, and degradation during long-term cycling. This review explores the transformative potential of graphene, a two-dimensional allotrope of carbon as a next-generation anode material for LIBs. Graphene features exceptional properties, including high surface area (2600 m^2/g), excellent electrical and thermal conductivity, mechanical strength, and theoretical specific capacity (∼744 mAh/g), positioning it as a compelling candidate to overcome the shortcomings of graphite. This paper discusses key synthesis strategies, such as chemical vapor deposition, exfoliation techniques, and redox methods, emphasizing their scalability, quality, and structural control. Furthermore, it explores the potential of hybrid architectures, like silicon–graphene composites and three-dimensional graphene frameworks, which offer enhanced lithium-ion diffusion, volumetric stability, and improved solid-electrolyte interphase (SEI) formation. Performance metrics such as energy density, charge/discharge rates, and cycling stability are critically analysed with evidence from recent studies. Although, commercial challenges such as, high production costs and limited scalability issues remain, active research efforts and industrial interest, particularly in electric vehicles and consumer electronics, signal a promising future for graphene-enhanced batteries. By situating graphene research within the practical context of fast-charging EVs, consumer electronics, and renewable-energy storage, this review highlights its potential to bridge the gap between laboratory performance and industrial application.

Keywords: graphene anodes, lithium-ion batteries, energy storage, battery performance enhancement, graphene-silicon composites

1. INTRODUCTION

Lithium-ion batteries (LIBs), characterized by high energy density, long cycle life, and lightweight design, have revolutionized modern energy storage, with widespread application in portable electronics, electric vehicles (EVs), and renewable energy systems. However, the growing demand for faster charging, higher capacity, and more sustainable energy solutions has highlighted the limitations of conventional LIB materials, particularly graphite anodes[1].

The anode influences the energy capacity, charging speed, and longevity. Although graphite anodes are reliable, their modest theoretical capacity of 372 mAh/g and sluggish lithium-ion diffusion kinetics limit the charging rate and energy density[2]. Moreover, structural degradation during cycling and the formation of unstable solid-electrolyte interphase (SEI) layers hinders performance enhancements. These shortcomings have motivated intensive research into next-generation anodes, with graphene emerging as one of the most promising candidates[3,4]. Its two-dimensional structure offers high conductivity, mechanical robustness, and a theoretical capacity of ∼744mAh/g twice that of graphite. Graphene's role is further strengthened when integrated into hybrid systems such as silicon-graphene composites or three-dimensional frameworks. Despite these advantages of graphene, scalable and cost-effective production, as well as real-world integration remain challenges towards commercialization[4].

This review examines the role of anode materials in LIBs, analyzes the limitations of graphite, and explores graphene's unique properties, synthesis methods, and structural innovations. Additionally, performance metrics, integration with real-world applications, commercialization hurdles, and prospects are explored, highlighting the potential of graphene to advance energy storage technologies for EVs, consumer electronics, and beyond.

Although many studies and reviews have examined graphene as an anode material, there remains limited discussion that connects material properties and sysnthesis methods to broader issues of scalability, cost, and real-world applications. This review seems to address that gap by bringing together insights on graphene's properties, fabrication strategies, and performance while also considering commercial challenges.

The rest of this paper is organized as follows. Section 2 provides background on lithium-ion battery fundamentals and the role of anode materials. Section 3 introduces graphene. Section 4 reviews synthesis strategies for graphene, with emphasis on scalability and cost. Section 5 examines structural innovations, comparison of graphene with graphite, and composite anodes that enhance electrochemical performance. Section 6 addresses commercialization challenges and prospects. Finally, Section 7 concludes the paper by summarizing key findings, limitations, and future directions.

2. FUNDAMENTALS OF LIBS AND THE ROLE OF THE ANODE

2.1. Principles of LIB Operation. LIBs typically consist of four key components: cathode (positive electrode), anode (negative electrode), electrolyte which facilitates lithium-ion transport, and a separator, a semipermeable membrane that prevents direct contact between the electrodes (Fig. 1). These components are packed within individual cells which can be connected to form battery packs[5].

During discharge, lithium ions in the anode are ionized, releasing electrons ($Li \rightarrow Li^+ + e^-$). These ions travel through the electrolyte and separator to the cathode. While the freed electrons flow through an external circuit, providing electrical energy to the connected device[6]. This conversion of chemical energy to electrical energy is the main function of a battery. During charging, an external voltage is applied, forcing lithium ions to move from the cathode to the anode. These ion-transport mechanisms enable LIB operation. The electro-chemical reactions influence properties such as charging rates and capacity[7], thereby affecting the overall efficiency and performance[5].

2.2. Anode Materials. Advances in anode materials are necessary to enhance the energy density, safety and overall performance of LIBs. Traditional anode materials like graphite are limited by their low theoretical capacity and safety concerns, prompting research into alternative materials such as alloys, silicon, metal oxides, and nanomaterials, each with unique advantages and limitations.

2.2.1. Alloy-Based Anodes. Alloying materials, such as silicon and tin, offer high lithium storage capacities, which can enhance the energy density of LIBs. However, they undergo substantial volume expansion during lithiation, leading to poor cycle life and high first-cycle irreversible capacity[8,9]. Strategies like nanosizing and composite synthesis can mitigate these issues, balancing between energy density with durability remains a challenge[9].

2.2.2. Silicon Anodes. Silicon exhibits the highest theoretical capacity among potential anode materials, making them a promising alternative to graphite. However, it undergoes significant volume changes during charge-discharge cycles, leading to mechanical degradation and capacity fading[10]. Strategies, such as the use of silicon-carbon composites, are being explored to address these challenges and improve the cycling stability of silicon anodes[10].

2.2.3. Metal Oxide Anodes. Metal oxides, including TiO_2, SnO_2, and transition metal oxides, offer high specific capacities but suffer from volume changes and low conductivity. These problems result in poor cycling performance, with strategies like nano-engineering and carbon coating required to enhance stability[11]. Transition metal oxides, such as MoO_2, MoS_2, and $MoSe_2$, exhibit high conductivity and specific capacity. However, different materials present distinct advantages and limitations, such as variations in diffusion barriers and binding capacities with lithium ions[12].

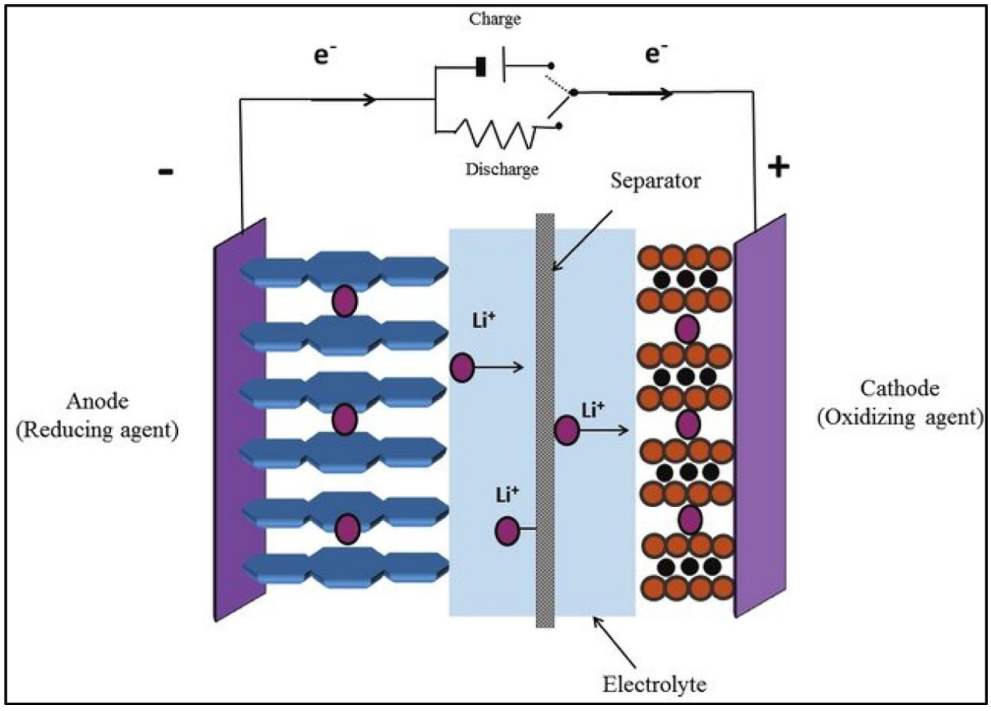


Figure 1. Internal configuration of a lithium-ion battery[5]

2.2.4. Nanomaterial Anodes. Nanomaterials, including graphene, transition metal sulfides, and MXenes offer enhanced conductivity and structural integrity but face challenges related to commercialization and long-term stability [10,13]. Carbon-based nanostructures, such as one- (1D), two- (2D), and three-dimensional (3D) carbon frameworks, provide large buffer spaces and improved lithium-ion conductivity. However, their capacity is limited, necessitating composite framework to enhance stability and capacity [14].

2.2.5. Graphite and Other Traditional Anodes. Graphite remains the most widely used anode material anode material owing to its safety and excellent stability. However, its low specific capacity limits the energy density of LIBs [15]. Efforts to improve the performance of graphite include the incorporation of carbon nanotubes and MXenes, as well as replacement with metal oxides and silicon [15].

2.3. Performance Bottlenecks of Graphite Anodes .

The composition and structure of the anode determine LIB performance, especially in terms of the energy capacity, charging speed, and cycle life [16]. An ideal anode material should store a large number of lithium ions during charging and release them effectively during discharge, ensuring a stable reversible process over multiple cycles. Desirable properties include high surface area (offering numerous lithium-ion attachment sites), specific capacity (for higher energy storage), and conductivity (to convert chemical energy into electrical energy). The widespread adoption of graphite anodes in LIBs is attributable to the high conductivity, low cost, and stable, layered structure of graphite, which enables reversible lithium-ion intercalation [17].

Nevertheless, the low theoretical specific capacity of graphite (372 mAh/g) limits the maximum energy density achievable in LIBs [17,18]. Furthermore, slow hinders rapid charging [19], with fast charging degrading the anode performance. Over extended charge-discharge cycles, capacity fade occurs owing to SEI formation. Sluggish reaction kinematics and poor lithium-ion diffusion reduce storage capacity and rate capability of graphite anodes [20].

3. GRAPHENE AS HIGH-PERFORMANCE ANODE MATERIAL

Graphene has emerged as a promising anode material for LIBs owing to its superior properties relative to traditional anode materials like graphite. Key benefits include high electrical conductivity, large surface area, and structural flexibility, which contribute to enhanced battery performance. The potential of graphene is further amplified when used in composite structures, leading to improved cycling performance and capacity. The following sections outline the specific advantages of graphene compared with other anode materials.

3.1. Structure of Graphene.

Graphene (Fig. 2) is a unique allotrope of carbon, which was the first 2D crystal to be isolated. It consists of a single layer of carbon atoms tightly bound into a 2D hexagonal honeycomb lattice, where each carbon atom is connected to another carbon atom via C-C double bonds (1-σ and 1-π per carbon atom) which are sp^2 hybridized [20]. It was the first two-dimensional crystal to be found. This distinct structure gives rise to a remarkable set of properties that are highly desirable for battery applications.

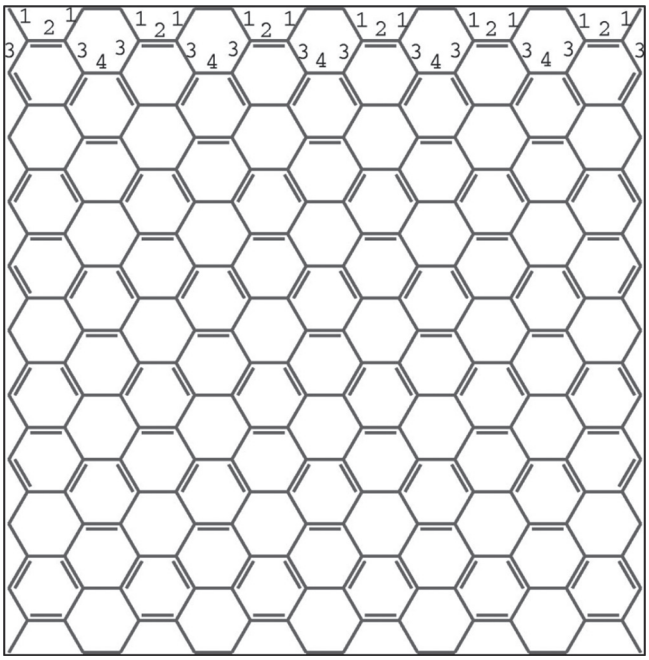


Figure 2. Structure of graphene [21]

4. DESIGNING GRAPHENE-ENHANCED ANODES

Graphene synthesis is a critical area of research owing to its exceptional properties and wide-ranging applications. Various synthesis methods have been developed, each with distinct advantages and challenges. Primary techniques include chemical vapor deposition (CVD), exfoliation, epitaxial growth, and chemical reduction. These methods can be divided into top-down and bottom-up approaches, differing in scalability, quality, and cost-effectiveness (Table 1). The choice of synthesis method often depends on the intended application of the graphene produced. Below is a detailed discussion of these methods.

4.1. Synthesis Methods.

4.1.1. Chemical Vapor Deposition (CVD). CVD (Fig. 3) is widely used technique for producing high-quality graphene films on metal substrates like copper and nickel [31]. This bottom-up approach involves the reaction of hydrocarbon gas precursors with the substrate at elevated temperatures [34,35]. Notably, CVD can enable large-scale production of uniform graphene films on metal films, suitable for electronic applications [32,36,37]. However, challenges remain in removing the catalyst and transferring the delicate graphene films to a substrate without introducing defects [33]. Moreover, the process is expensive and requires precise control over reaction conditions, and faces scalability and environmental concerns [37,38].

4.1.2. Exfoliation Techniques. Exfoliation techniques separate graphene sheets from bulk graphite.

- **Mechanical Exfoliation:** Famously demonstrated with Scotch tape, this method yields high-quality graphene

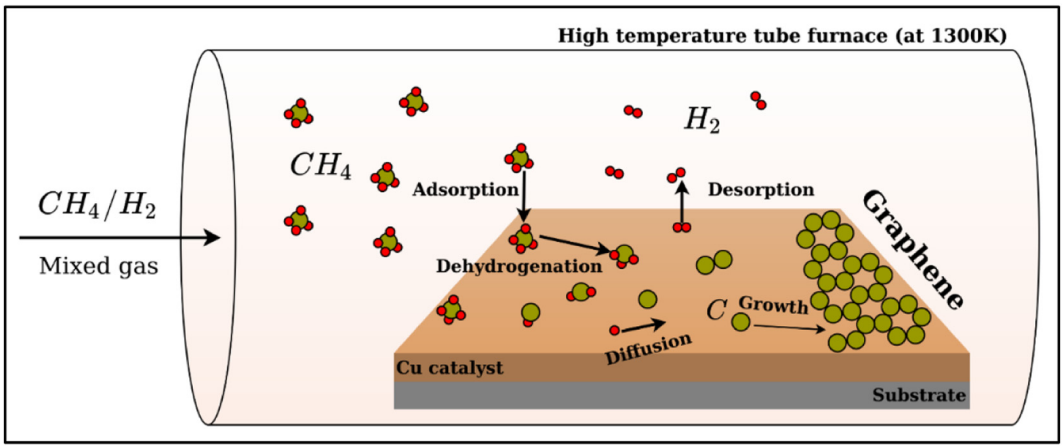


Figure 3. Schematic of chemical vapor deposition[39]

but is not suitable for large-scale production owing to its labor-intensive nature[31]. This process involves repeatedly graphitizing and separating graphite layers using tape until an isolated layer of graphene is obtained (Fig. 4).

- **Liquid Phase Exfoliation:** Graphite is dispersed in a suitable solvent and subjected to ultrasonication to separate graphene flakes[31]. It is more scalable than mechanical exfoliation, as it involves chemical reactions instead of manual effort, but the graphene yield and quality depend on the solvent properties.
- **Electrochemical Exfoliation:** This approach leverages electrochemical reactions in electrolytic solutions to exfoliate graphite into graphene[40]. It is environmentally friendly and scalable.
- **Redox Method:** Graphite is oxidized to graphite oxide (GO), which is then easily exfoliated. The resulting GO is reduced chemically or thermally to obtain reduced graphene oxide[31]. This method is cost-effective for mass production of graphene but can introduce structural defects if oxidation or reduction is poorly implemented.

4.1.3. Epitaxial Growth. Graphene on silicon carbide (SiC) substrates via thermal decomposition. This bottom-up approach produces high-quality graphene[36,41], with excellent electronic properties, making it suitable for high-frequency electronic devices[34,37]. However, it is expensive and limited by the size of the SiC substrate, which affects scalability[35,38].

4.1.4. Chemical Reduction. This top-down method involves the reduction of GO to produce graphene oxidizing and reducing agents[35,42]. The approach is simple and can be scaled up for mass production[35,43]. The quality of graphene often contains to residual oxygen groups and defects, which degrade the quality[37,38].

4.1.5 Emerging Trends and Future Prospects

Recent research has been focused on using natural carbon sources, such as coal and biomass, to produce graphene derivatives aimed at improving scalability and cost-effectiveness[38]. There is also growing emphasis on developing environmentally friendly synthesis methods to minimize the ecological impact of graphene production[34,43].

Despite advances in synthesis methods, challenges remain in terms of large-scale production, cost, and environmental impact. The development of novel methods and optimization of existing ones are crucial for the widespread adoption of graphene in various industries.

4.2. Structural Innovations: Silicon-Graphene Composites. Graphene has emerged as an ideal material for developing advanced anodes. Although silicon has an ultra-high theoretical capacity (~4200 mAh/g), it suffers from volume expansion (~300%) during charge-discharge cycles, which diminished battery stability. Given the excellent conductivity of graphene and its ability to buffer this volume expansion, researchers have attempted to design unique structures incorporating both silicon and graphene[46]. Key designs including core-shell architectures, where silicon layers are placed between graphene sheets[47] and 3D graphene networks, which provide conductive support for silicon nanoparticles[48]. These composites can be synthesized using various methods such as ball milling, spray drying, hydrothermal/solvothermal reactions, and CVD[49]. Silicon-graphene composites exhibit enhanced cycle stability, higher reversible capacity, and improved rate capability compared with pure silicon anodes[46]. Atomic layer deposition of Al_2O_3 on silicon-graphene electrodes suppress unwanted side reactions with the electrolyte[48].

Another structural innovation involves the creation of 3D graphene frameworks, such as porous graphene, graphene aerogels, and graphene foams[50], which prevent the restacking of graphene sheets and maximize the surface area. The interconnected network of graphene provides continuous pathways for electron transport, while the porous structure allows for efficient electrolyte infiltration and rapid lithium-ion diffusion[31]. Such 3D graphene-based anodes demonstrate improved rate performance and cycling stability[51]. The development of such structures is crucial for overcoming their limitations of both pure graphene and graphite materials, paving the way for high-performance LIBs.

5. PERFORMANCE ADVANTAGES OF GRAPHENE ANODES

Recent studies have highlighted the advantages of graphene anodes in LIBs, offering enhanced energy storage owing to its

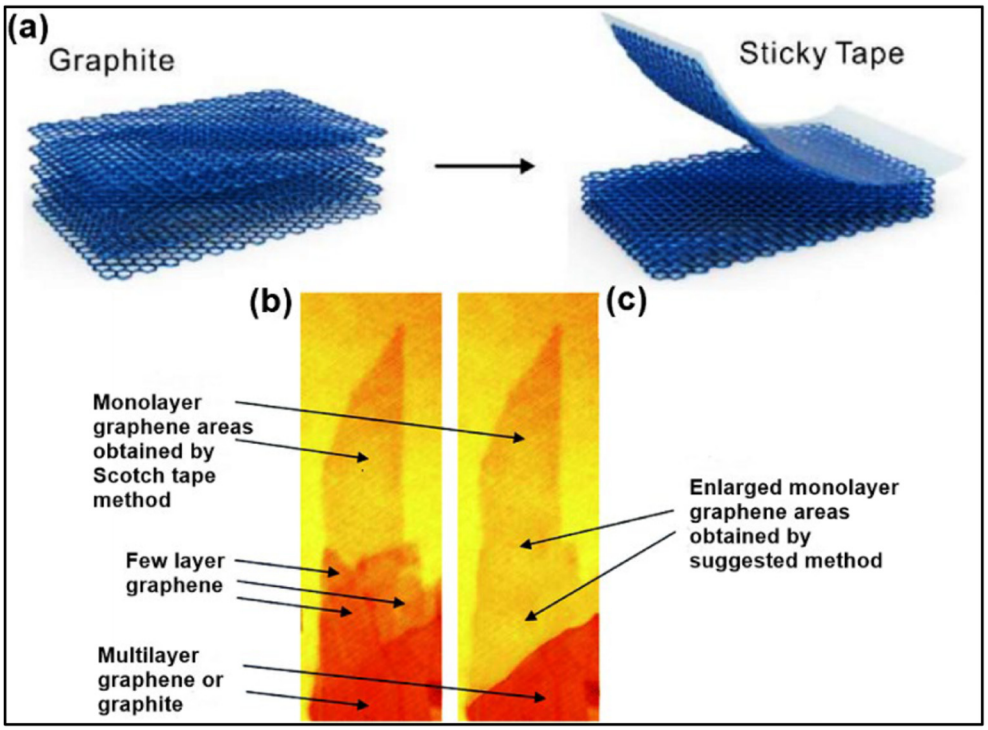


Figure 4. Mechanical exfoliation of graphene[39]

high electron conductivity, large specific surface area, and structural flexibility, graphene can enhance, the energy density, rate performance and cycling stability of LIBs addressing the limitations of traditional graphite anodes. Table 2 summarizes the performance metrics of various anode materials.

5.1. High Electrical Conductivity. Graphene exhibits excellent electrical conductivity, enabling rapid electron transport, thereby improving the rate performance and efficiency of LIBs. This property is particularly beneficial when graphene is used as a conductive agent in electrode, enhancing the overall electrochemical performance of batteries[22].

5.2. Enhanced Electrochemical Performance. While graphite has a theoretical capacity of 372 mAh/g, graphene can achieve higher capacities owing to its ability to adsorb lithium ions on both sides of its sheets[23,52]. Its high surface area provides more active sites for lithium-ion storage, facilitating faster ion transport and improving rate capabilities[23]. Graphene's excellent electrical conductivity promotes electron transport, enhancing LIB power densities[52,53].

5.3. Large Surface Area. The high specific surface area of graphene provides abundant active sites for electrochemical reactions, which increasing the energy storage capacity of the anode. This characteristic also promotes faster ion transport compared to that in, graphite which suffers from sluggish lithium diffusion[23].

5.4. Structural and Mechanical Benefits. Graphene's exhibits structural flexibility allowing it to accommodate volume changes during lithium-ion intercalation and deintercalation, thereby reducing the risk of mechanical degradation and improving the cycling stability and mitigating electrode degradation[24,54,55]. When used as a coating material, graphene can prevent the growth of lithium dendrites, which often lead to short circuits and safety issues in batteries[23].

Graphene can form stable 3D structures and composites with other materials, such as metal oxides, resulting in enhanced structural integrity and performance. These composites prevent the restacking of graphene sheets, maintaining high storage capacity during cycling[52,56].

5.5. Composite and Hybrid Anode Materials. Beyond LIBs, graphene is also being explored as an anode material for sodium-ion batteries, where its high conductivity and large surface area provide similar advantages[28]. However, the production of high-quality graphene at a reasonable cost remains a challenge,

Table 1. *Comparison of Graphene Synthesis Methods*

Method	Type	Advantages	Disadvantages	References
Chemical Vapor Deposition	Bottom-up	High-quality, large-area graphene	Expensive, low scalability	34
Liquid-phase Exfoliation	Top-down	Scalable, cost-effective	Lower quality, defect-prone	44
Electrochemical Exfoliation	Top-down	Environmentally friendly, tunable properties	Requires optimization for uniformity	45
Reduction of Graphene Oxide	Top-down	Easy processing, functionalization possible	Residual oxygen groups affect performance	42

Table 2. *Performance Metrics of Selected Anode Materials in Lithium-Ion Battery Applications*

Material	Capacity (mAh/g)	Cycle Life	Rate Capability	Stability
Graphite	~ 372	High	Moderate	High
Silicon	~ 4200	Low	Low	Poor
SnO2	~ 790	Moderate	Low	Moderate
Graphene	540–2000	High	High	Excellent

and the aggregation of graphene sheets can reduce the effective surface area and conductivity. Ongoing research is focused on identifying innovative solutions to maximize the potential of graphene in energy storage. When combined with other materials, such as metal oxides or silicon, graphene can enhance the overall performance anode by providing a stable and conductive matrix[25,26] Graphene composites have shown improved capacity and cycling performance, making them suitable for high-energy-density applications[27]. Materials such as Fe_2O_3 and SnO_2 with graphene, leverage the complementary properties of each component to maximize energy density and cycling performance. These composites also help alleviate issues like irreversible capacity loss by preventing graphene layer restacking[52]. When combined with other high-capacity materials, such as GeO_x, graphene enhances the overall anode performance by functioning as a conductive matrix that enhances charge transfer kinetics and cycling stability[23].

5.6. Energy Density and Rate Capability. In LIBs, graphene-based anodes have been noted to enhance energy density compared with that associated with traditional graphite anodes[57]. For instance, graphene nanoplates have exhibited higher reversible capacities than graphite[58]. Notably, the $Li_4Ti_5O_{12}$/graphene foam electrode exhibits an exceptionally high-rate capability of 200 C[59]. The ability of graphene to enhance both energy density and rate capability positions it as a promising material for advanced batteries.

5.7. Cycling Stability and SEI Improvements. Graphene improves the cycling stability of lithium-ion battery anodes through its mechanical strength which helps accommodate volume changes during charge/discharge cycles, particularly in silicon-graphene composites[46]. Moreover, graphene can promote the formation of a stable and protective SEI layer, which prevents further electrolytic decomposition and enhanced capacity retention over multiple cycles[6,49].

5.8. Comparison of Graphite and Graphene. The structural differences between graphene and graphite explain their distinctive properties. Graphite is composed of multiple graphene layers stacked together via weak van der Waals forces,

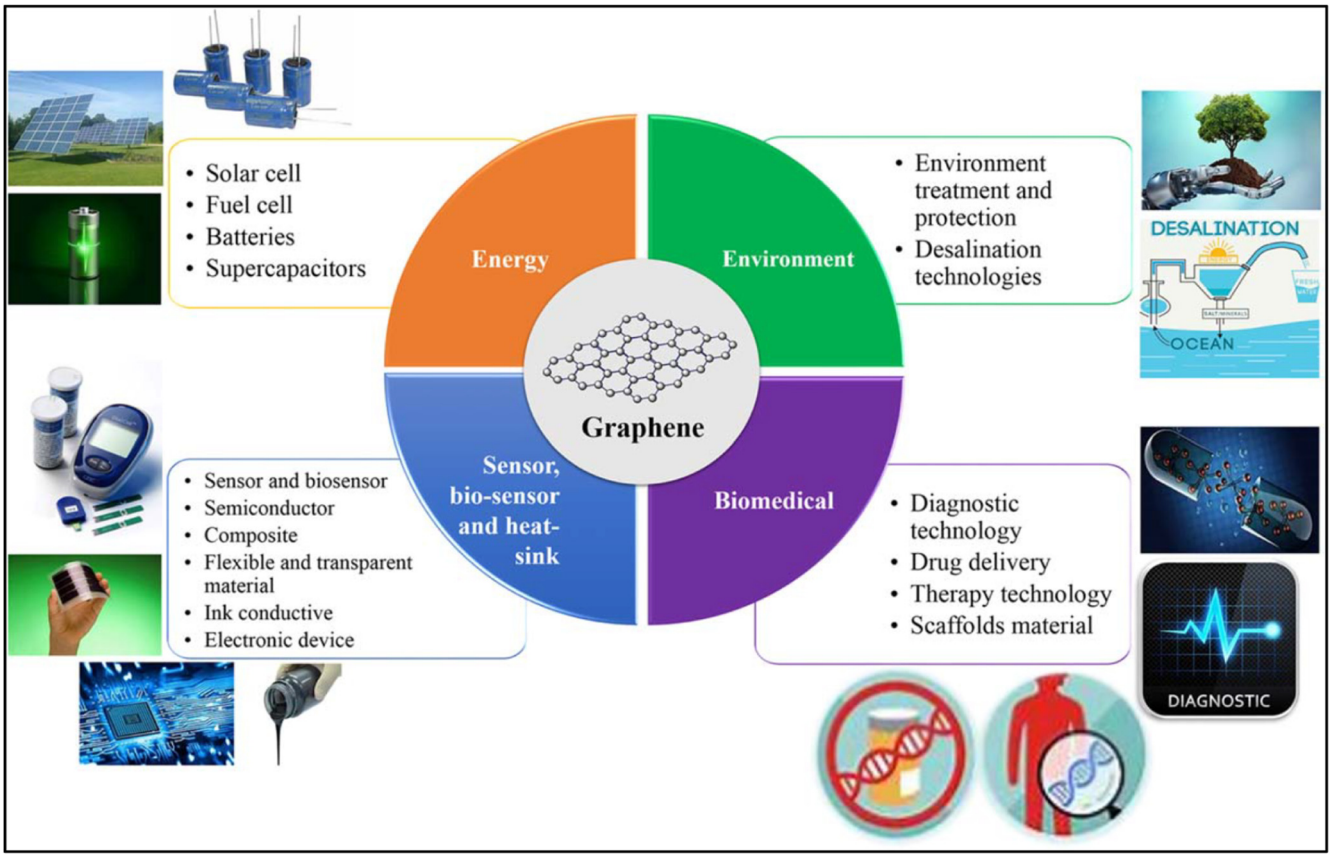


Figure 5. Real-World Applications of Graphene[60]

resulting in a considerably lower surface area (10 m^2/g) compared with that of a single graphene layer (~2600 m^2/g)[29]. While graphite is electrically conductive, its conductivity is lower than that of graphene owing to interlayer interactions (van der Waals forces)[30]. Mechanically, graphite is brittle with a tensile strength of 4.8 GPa[30], considerably smaller than that of graphene (130 GPa). The theoretical specific capacity of graphene is 744 mAh/g[30], two times that of graphite's 372 mAh/g.

6. CHALLENGES AND COMMERCIAL VIABILITY

6.1. Scalability and Cost Challenges. Despite the performance advantages of graphene-enhanced anodes, their commercialization is hindered by various challenges, especially the limited scalability and high cost of producing high-quality graphene in large quantities[30]. Graphene costs vary significantly depending on the synthesis method and desired quality, but it remains considerably more expensive than commercially available anode materials[31]. Currently, the cost of high-quality graphene is approximately USD $100/g which hinders its widespread use. As of now, high quality graphene could cost about $100/g, which is not viable for widespread use in. Nevertheless, extensive research efforts are underway to establish more scalable and cost-effective production techniques.

6.2. Real-World Applications (EVs and Consumer Electronics). Despite the abovementioned challenges, graphene-enhanced anodes hold strong potential for various real-world applications (Fig. 5), particularly in EVs, renewables, and consumer electronics[46]. In the case of EVs, graphene can support faster charging and extended driving range. Several companies and research groups are actively developing these graphene-based solutions for EVs, including graphene-aluminum-ion batteries[59].

In portable electronics, where smaller, lighter, and longer-lasting batteries are constantly in demand, graphene offers significant advantages, including extended battery life, faster charging. Leading companies such as Samsung and Huawei have already started incorporating graphene into their battery-operated devices NASA is testing graphene battery technologies for space applications. The unique combination of properties offered by graphene makes it highly attractive for enhancing battery performance across key technology sectors.

7. CONCLUSIONS

Graphene has transformed from a "wonder material" to a practical solution for overcoming the limitations of current LIBs. With a theoretical capacity nearly two times that of graphite, high strength, exceptionally high surface area, and excellent electrical and thermal conductivity, graphene-based anodes can significantly improve battery energy density, charging rates, and lifespan. Structural innovations like silicon-graphene composites and 3D frameworks further enhance stability and performance.

While various challenges remain towards commercialization, such as high production costs, quality variations, and limited large-scale synthesis methods, both academic and industry are actively working to promote the implementation of active, graphene-based technologies, ushering in a new era of LIBs with reliability and stability. Emerging scalable techniques such as electrochemical synthesis and CVD present viable routes forward. Overcoming these economic and manufacturing hurdles is the final step toward realizing the full potential of graphene in transforming energy storage for EVs, consumer electronics and beyond.

AFFILIATIONS AND AUTHOR DETAILS

Corresponding Author

Mihir Gutti – *Department of Engineering, University of San Francisco, San Francisco, CA 94117, USA;* 0009-0009-4303-3834
Email: guttimihir@gmail.com

ACKNOWLEDGEMENTS

The author acknowledges Prof. Brandon Brown from the University of San Francisco, USA, for his mentorship and guidance in writing the manuscript, Prof. Dr.-Ing. V. V. S. S. Srikanth from University of Hyderabad, India, for his initial training on graphene batteries, and Dr. Mounika Sarvepalli from the National Institute of Technology, Warangal, India, for her support with the manuscript. This work was supported by the Department of Engineering, University of San Francisco, USA. The author also appreciates the financial support provided through the Merit-based Provost Scholarship and SAME Foundation Engineering Scholarship.

CONFLICTS OF INTEREST

Authors declare that there are no conflicts of interest.

REFERENCES

(1) Mo, R. *etale* High-quality mesoporous graphene particles as high-energy and fast-charging anodes for lithium-ion batteries. *Nat Commun* **10**, 1474–1474 (2019).
(2) Mypati, O. *etale* The synthesis of novel porous graphene anodes for fast charging and improved electrochemical performance for lithium-ion batteries. *Energy Sources Part A-recovery Utilization and Environmental Effects* **44**, 4349–4363 (2022).
(3) High-energy, high-density and fast-charging graphene battery. Preprint at https://scispace.com/papers/high- energy-high-density-and-fast-charging-graphene-battery-1oko14av5b (2018).
(4) Lim, J. M. *etale* High Volumetric Energy and Power Density Li2TiSiO5 Battery Anodes via Graphene Functionalization. *Matter* **3**, 522–533 (2020).
(5) Ramanan, A. Nobel Prize in Chemistry 2019. *Resonance* **24**, 1381–1395 (2019).
(6) Zhao, X. *etale* Electrochemical exfoliation of graphene as an anode material for ultra-long cycle lithium ion batteries. *Journal of Physics and Chemistry of Solids* **139**, (2020).
(7) Whittingham, M. S. Electrical energy storage and intercalation chemistry. *Science* **192**, 1126–1127 (1976).
(8) Gopinadh, S. V., Phanendra, P. V. R. L., Anoopkumar, V., John, B. & Td, M. Progress, Challenges, and Perspectives on Alloy-Based Anode

Materials for Lithium Ion Battery: A Mini-Review. *Energy & Fuels* **38**, 17253–17277 (2024).

(9) Li, W. Anode Material Innovations for Boosting Battery Energy Density. *Highlights in Science Engineering and Technology* **121**, 138–145 (2024).

(10) Borkar, S., Nahalde, S., Ruban, A. J. S. & More, H. A Comprehensive Review of Advancement in Anode Material with Modified Architecture for Lithium-Ion Batteries. *SAE technical paper series* **1**, (2024).

(11) Kebede, M., Zheng, H. & Ozoemena, K. I. Metal Oxides and Lithium Alloys as Anode Materials for Lithium-Ion Batteries. 55–91 (2016) doi:10.1007/978-3-319-26082-2_3.

(12) Chen, T. Investigation of 2D material anodes with different anions for lithium ion batteries: comparison of MoO2, MoS2 and MoSe2. *Journal of physics* **2331**, 012005–012005 (2022).

(13) Sun, W. Comparison of Different Nanomaterials in Anode Materials of Lithium Battery. *Applied and Computational Engineering* **126**, 176–181 (2025).

(14) Nandihalli, N. A Review of Nanocarbon-Based Anode Materials for Lithium-Ion Batteries. *Crystals (Basel)* **14**, 800–800 (2024).

(15) Zhao, W., Zhao, C., Wu, H., Li, L. & Zhang, C. Progress, challenge and perspective of graphite-based anode materials for lithium batteries: A review. *J Energy Storage* **81**, (2024).

(16) (Infographics #12) Anode - BATTERY INSIDE. https://inside.lgensol.com/en/2023/10/infographics-12-anode/.

(17) Mishra, Y. *etale* Graphene oxide–lithium-ion batteries: inauguration of an era in energy storage technology. *Clean Energy* **8**, 194–205 (2024).

(18) ViPER - Research. https://engineering.purdue.edu/ViPER/research.html.

(19) Chang, H., Wu, Y.-R., Han, X. & Yi, T.-F. Recent developments in advanced anode materials for lithium-ion batteries. *Energy Mater 2021;1:100003.* **1**, N/A-N/A (2021).

(20) A Brief Introduction to Graphite - Volta Foundation. https://volta.foundation/battery-bits/a-brief-introduction-to-graphite.

(21) Liu, L., Jayanthi, C. S., Wu, S. Y. & Guo, H. Broken symmetry, boundary conditions, and band-gap oscillations in finite single-wall carbon nanotubes. *Physical Review B* **64**, 033414 (2001).

(22) Qi, C. *etale* Application of Graphene in Lithium-Ion Batteries. (2024) doi:10.5772/INTECHOPEN.114286.

(23) Tian, H., Wang, X.-L. & Han, W. Amorphous Hierarchical Porous GeO_x/reduced Graphene Oxide Composite As a High-Performance Anode Material for Lithium Ion Batteries. *ECS Meeting Abstracts* **MA2014-01**, 277–277 (2014).

(24) Dong, L., Ren, W., Dong, L. & Li, D. J. Synthesis and Characterization of Graphene Sheets as an Anode Material for Lithium-Ion Batteries. *Key Eng Mater* **537**, 238–242 (2013).

(25) Chen, S. A Review on Graphene Composite Nanomaterials in Anode of Lithium-Ion Battery. *International journal of energy* **5**, 21–24 (2024).

(26) Jeong, S. *etale* Enhanced Electrochemical Properties of Silicon and Quasi-Defect-Free Reduced Graphene Oxide for High Performance Anode Materials. *Meeting abstracts* **MA2024-02**, 1473–1473 (2024).

(27) Anode material including graphite phase carbon material and functionalized graphene, method thereof, and lithium ion battery (2020) |Zhang Mingdong. https://scispace.com/papers/anode-material -including-graphite-phase-carbon-material-and-2tchz4pnkw.

(28) Thakur, A. K. *etale* Advancement in graphene-based nanocomposites as high capacity anode materials for sodium-ion batteries. *J Mater Chem* **9**, 2628–2661 (2021).

(29) Bonaccorso, F. *etale* Graphene, related two-dimensional crystals, and hybrid systems for energy conversion and storage. *Science* **347**, (2015).

(30) Graphite vs Graphene: What's The Difference? |Jinsun Carbon. https://jinsuncarbon.com/graphite-vs-graphene/.

(31) Liu, Z., Tian, Y., Wang, P. & Zhang, G. Applications of graphene-based composites in the anode of lithium-ion batteries. *Frontiers in Nanotechnology* **4**, 952200 (2022).

(32) Saeed, M., Alshammari, Y., Majeed, S. A. & Al-Nasrallah, E. Chemical Vapour Deposition of Graphene—Synthesis, Characterisation, and Applications: A Review. *Molecules 2020, Vol. 25, Page 3856* **25**, 3856 (2020).

(33) MOOSA, A. A. & ABED, M. S. Graphene preparation and graphite exfoliation. *Turkish Journal of Chemistry* **45**, 493–519 (2021).

(34) gupta, gopal & gupta, A. Advances in the Synthesis of Graphene: A Comprehensive Review. (2024) doi:10.20944/PREPRINTS202405.0582.V1.

(35) Ghosh, R. Synthesis Methods for Graphene. (2022) doi:10.36227/TECHRXIV.19540417.

(36) Alwan, S. H., Omran, A. A., Naser, D. K. & Ramadan, M. F. A Mini-Review on Graphene: Exploration of Synthesis Methods and Multifaceted Properties. *Engineering Proceedings* **59**, (2024).

(37) Ramezani, M. J. & Rahmani, O. A review of recent progress in the graphene syntheses and its applications. *Mechanics of Advanced Materials and Structures* 1–33 (2024) doi:10.1080/15376494.2024.2420911.

(38) Tamuly, J., Bhattacharjya, D. & Saikia, B. K. Graphene/Graphene Derivatives from Coal, Biomass, and Wastes: Synthesis, Energy Applications, and Perspectives. *Energy & Fuels* **36**, 12847–12874 (2022).

(39) Uzoma, P. C., Hu, H., Khadem, M. & Penkov, O. V. Tribology of 2D Nanomaterials: A Review. *Coatings 2020, Vol. 10, Page 897* **10**, 897 (2020).

(40) Dericiler, K., Alishah, H. M., Bozar, S., Güneş, S. & Kaya, F. A novel method for graphene synthesis via electrochemical process and its utilization in organic photovoltaic devices. *Applied Physics A: Materials Science and Processing* **126**, 1–9 (2020).

(41) Santhiran, A., Iyngaran, P., Abiman, P. & Kuganathan, N. Graphene Synthesis and Its Recent Advances in Applications—A Review. *C (Basel)* **7**, 76 (2021).

(42) Urade, A. R., Lahiri, I. & Suresh, K. S. Graphene Properties, Synthesis and Applications: A Review. *JOM* **75**, 614–630 (2022).

(43) Kartini, E., Setiadi, T. A. & Muhammad Fakhruddin. A Review on Graphene: Synthesis Methods, Sources, and Applications. *Journal Of Batterie For Renewable Energy And Electric Vehicles* **1**, 41–50 (2023).

(44) Liu, Z., Tian, Y., Wang, P. & Zhang, G. Applications of graphene-based composites in the anode of lithium-ion batteries. *Frontiers in Nanotechnology* **4**, 952200 (2022).

(45) Dericiler, K., Alishah, H. M., Bozar, S., Güneş, S. & Kaya, F. A novel method for graphene synthesis via electrochemical process and its utilization in organic photovoltaic devices. *Appl Phys A Mater Sci Process* **126**, 1–9 (2020).

(46) Ni, C. *etale* Effect of Graphene on the Performance of Silicon–Carbon Composite Anode Materials for Lithium-Ion Batteries. *Materials 2024, Vol. 17, Page 754* **17**, 754 (2024).

(47) Yang, Y. *etale* Using Sandwiched Silicon/Reduced Graphene Oxide Composites with Dual Hybridization for Their Stable Lithium Storage Properties. *Molecules* **29**, 2178 (2024).

(48) Ren, J.-G. *etale* Silicon–Graphene Composite Anodes for High-Energy Lithium Batteries. *Energy Technology* **1**, 77–84 (2013).

(49) Liang, Y. Z., Bhat, A. L. & Su, Y. S. Green Synthesis of Graphene Flake/Silicon Composite Anode for Lithium-Ion Batteries Using a Ball-Mill-Derived Mechanical Transfer Technique. *ACS Applied Energy Materials* **7**, 10574–10583 (2024).

(50) Research progress of 3D porous graphene in lithium-ion battery anode material. *New Chemical Materials* **52**, 8–13 (2024).

(51) Luo, J. *etale* Three-dimensional graphene framework scaffolded FeP nanoparticles as anodes for high performance lithium ion batteries. *Materials Letters* **246**, 84–87 (2019).

(52) Xia, G. *etale* Graphene/Fe2O3/SnO2 ternary nanocomposites as a high-performance anode for lithium ion batteries. *ACS Appl Mater Interfaces* **5**, 8607–8614 (2013).

(53) Alathlawi, H. J. & Hassan, K. F. Review—Recent Advancements in Graphene-Based Electrodes for Lithium-Ion Batteries. *ECS Journal of Solid State Science and Technology* **13**, 011002 (2023).

(54) Bi, J. *etale* On the Road to the Frontiers of Lithium-Ion Batteries: A Review and Outlook of Graphene Anodes. *Advanced Materials* **35**, (2023).

(55) Ma, Z.-F., Yuan, T., Ma, J., He, Y.-S. & Liao, X.-Z. Graphene-Based Anode Material Design and Preparation Process for Lithium Ion Battery. *ECS Meeting Abstracts* **MA2015-01**, 1560–1560 (2015).

(56) Wang, H., Li, X., Baker-Fales, M. & Amama, P. B. 3D graphene-based anode materials for Li-ion batteries. *Curr Opin Chem Eng* **13**, 124–132 (2016).

(57) Esteve-Adell, I. *etale* Influence of the Specific Surface Area of Graphene Nanoplatelets on the Capacity of Lithium-Ion Batteries. *Frontiers in Chemistry* **10**, 807980 (2022).

(58) Li, N., Chen, Z., Ren, W., Li, F. & Cheng, H. M. Flexible graphene-based lithium ion batteries with ultrafast charge and discharge rates. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 17360–17365 (2012).

(59) Graphene Batteries in Electric Vehicles. https://www.azom.com/article.aspx?ArticleID$=$21330.

(60) Madurani, K. A. *etale* Progress in Graphene Synthesis and its Application: History, Challenge and the Future Outlook for Research and Industry. *ECS Journal of Solid State Science and Technology* **9**, 093013 (2020).

KFUPM
Journal of Undergraduate Research International

Diamond Open Access 82

Article Electrical Engineering/Industrial Engineering

Adaptive Velocity PSO-Based Parameter Optimization for a Permanent Magnet DC Motor Drive in Light Electric Vehicles

Mohammed Aldhaif Allah[1] and Moustafa Magdi Ismail[2]*

Cite https://doi.org/10.64589/juri/209733

Submitted: May 12, 2025 **Revised:** August 04, 2025 **Accepted:** August 20, 2025

ABSTRACT

This study presents an adaptive velocity particle swarm optimization (AVPSO) approach for tuning the proportional–integral (PI) controllers of a permanent magnet DC (PMDC) motor drive in light electric vehicles (LEVs). The objective is to enhance dynamic response, tracking accuracy, and stability under varying operating conditions. The proposed algorithm improves on classical particle swarm optimization (PSO) by adaptively adjusting inertia weight and acceleration coefficients, thereby achieving a better balance between exploration and convergence. A MATLAB/Simulink model is developed for offline evaluation, where the optimization process minimizes steady-state and transient errors in motor speed and armature current. Simulation results demonstrate that AVPSO achieves faster settling times, with reductions of 15.6% in forward operation and 2.1% in reverse, while eliminating steady-state errors present in classical PSO. Comparative analysis under four-quadrant operation and torque disturbance scenarios confirms that the AVPSO-based tuning improves current tracking, reduces power losses, and ensures stable performance across dynamic transitions. The findings highlight the effectiveness of AVPSO in enhancing control quality without compromising efficiency, offering a robust and scalable solution for LEV motor drive applications.

Keywords: adaptive velocity PSO, permanent magnet DC motor, off-line tuning parameters, optimization control, motor drive, light electric vehicles

1. INTRODUCTION

Light electric vehicles (LEVs), including e-bikes, e-scooters, e-rickshaws, golf carts, and compact delivery vehicles, are increasingly adopted for urban mobility due to their environmental advantages, affordability, and suitability for short-distance, low-speed applications [1]. These vehicles require propulsion systems that are compact, efficient, and easy to control. Among various motor technologies, direct current (DC) motors particularly in brushed, brushless, and permanent magnet configurations remain a practical choice for LEVs due to their simple construction, smooth torque control, and minimal electronic complexity [2].

DC motors are widely used not only in propulsion systems, such as wheel hub motors [3], but also in auxiliary functions including electric power steering, ventilation, heating, and actuators for windows and seats [4]. For high-performance and larger electric vehicles (e.g., cars and buses), alternating current (AC) machines such as induction motors and permanent magnet synchronous motors (PMSMs) are preferred for their higher efficiency, speed capabilities, and regenerative braking support [5,6]. Nevertheless, for light-duty and cost-sensitive applications, DC motors continue to offer an optimal balance between functionality and system simplicity.

Recent advancements in motor drive control have emphasized optimization strategies for improving system performance under dynamic conditions [13–19]. For example, Xu et al. proposed an adaptive velocity particle swarm optimization (AVPSO) method to enhance flux-weakening control in permanent magnet synchronous motor drives [13]. The technique demonstrated reduced steady-state error and improved transient performance compared to gain parameters tuned using trial-and-error methods [13]. Other works have employed similar techniques to reduce torque ripple in surface-mounted PMSMs and optimize controller parameters using genetic algorithms [20]. Beyond motor control, AVPSO has also been successfully applied to maximum power point tracking in photovoltaic systems under no uniform irradiation conditions, as demonstrated by Zhang et al., yielding superior tracking speed and accuracy [21].

The main contribution of this study is the development of an off-line optimization strategy based on AVPSO for tuning the cascaded PI controllers of permanent magnet DC (PMDC) motor drives in light electric vehicles. Unlike conventional tuning methods that rely on manual adjustments and perform poorly under dynamic conditions, the proposed approach simultaneously optimizes four PI gains two for the inner current loop and two for the outer speed loop. The proposed AVPSO algorithm adaptively adjusts its search parameters to enhance convergence and solution quality. This results in notable improvements in dynamic response, tracking accuracy, and robustness over the traditional PSO algorithm presented in this work.

Table 1. *Comparison of DC motor types for electric vehicle applications*

Motor Type	Advantages	Disadvantages	Suitability for EV
Permanent Magnet DC [2]	- Compact size - High efficiency - High torque at low speed - Simple construction	- Limited power range - Poor heat dissipation - Expensive rare-earth magnets	Highly suitable for light EVs due to simplicity, easy control, and efficiency
Separately Excited DC Motor [3]	- Independent control of field and armature - Excellent speed and torque control - High dynamic response	- Requires two power sources - Complex - Higher cost and maintenance	Partially suitable for research or high-performance EVs, but less practical for consumer use
Shunt Self-Excited DC Motor [8]	- Stable speed regardless of load - simple speed regulation	- Lower starting torque - Not ideal for variable load	Not suitable, not ideal for frequent start-stop or acceleration requirements in EVs
Series Self-Excited DC Motor [9]	- Very high starting torque - Simple construction - simple for traction	- Poor speed regulation - Risk of overspeed in no load - Maintenance-intensive (brushes)	Historically used in older EVs or trams, but not preferred now due to control and safety limitations
Short Compound DC Motor [10]	- Balanced between torque and speed regulation - Better load handling than shunt	- More complex than shunt/series - Still not efficient for EVs	Limited suitability, complexity not justified for LEVs
Long Compound DC Motor [11]	- Improved torque and regulation - Better handling under dynamic loads	- Bulkier - More losses - Costlier	Rarely used in EVs today due to inefficiency and size
Cumulative Compound Motor [11]	- Combines advantages of series and shunt - Better performance under varying loads	- Moderate complexity - Larger footprint	Some suitability for heavier EVs, but still less efficient than PMDC or brushless types
Differential Compound Motor [11]	-Not suitable for EV applications	- Risk of instability under load - Torque drops with increased load (not desirable)	Not suitable for EV applications

The method is particularly effective in four-quadrant operation and torque disturbance scenarios, offering a computationally efficient and scalable solution for advanced LEV propulsion systems.

The remainder of this paper is organized as follows. Section 2 reviews related works relevant to motor drive optimization and PI controller tuning. Section 3 presents the modeling of the proposed LEV system, including the PMDC motor and the cascaded PI control strategy. Section 4 describes the methodology, focusing on the development of the Adaptive Velocity Particle Swarm Optimization (AVPSO) algorithm for optimal PI parameter tuning. Section 5 provides the results and discussion, including the simulation setup, performance evaluation, and comparative analysis between AVPSO, classical PSO, and existing PSO-based methods. Finally, Section 6 concludes the paper by summarizing the main findings and outlining potential directions for future research.

2. RELATED WORK

Tables 1 and 2 provide a comparative overview of DC motor types relevant to electric vehicle applications. Table 1 highlights technical attributes such as efficiency, torque capability, and maintenance requirements, while Table 2 outlines common real-world use cases and defining features like cost, reliability, and control complexity. Together, these tables support informed motor selection based on application-specific requirements in the EV industry.

3. METHODOLOGY

3.1. Modeling the proposed drive system of the PMDC motor. The objective of this study is to enhance the control system of LEVs to achieve better performance and reliability. Figure 1 presents the overall LEV architecture, which con-

Table 2. *Comparison of DC Motor types in EV applications*

Motor Type	Common Use	Key Features
Brushed DC (BDC)	Low-cost LEVs, toys, small EVs	Simple, low cost, high maintenance [11]
Brushless DC	E-bikes, scooters, hub motors	Efficient, reliable, requires electronic controller [12]
PMDC	Lightweight EVs, power assist systems	High torque, compact, simple [12]
Series DC Motor	Older EV designs	High starting torque, less efficient [8]

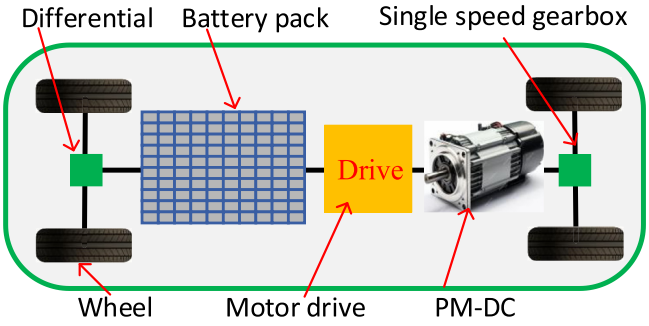


Figure 1. Main components of LEV system

sists of a motor drive, electric machine, battery pack, fixed-speed gearbox, and driving wheels. The traction motor, selected as a PMDC type powered by a Li-ion battery pack, is the primary focus of this work 19. The motor drive also incorporates the electrical circuit and charging system of the EV, where a DC/DC boost converter regulates power flow to ensure efficient energy conversion. This section therefore outlines the complete structure of the proposed LEV drive system.

The control system for the PMDC motor employs a cascaded PI control strategy consisting of two loops: an outer speed loop and an inner current loop, as shown in Figure 2. Meanwhile, the outer PI controller is responsible for tracking the reference angular velocity by generating a reference current signal. In contrast, the inner PI controller regulates the armature current by driving the voltage applied to the motor armature.

For beginning with, the electrical dynamics of the PMDC motor are derived from Kirchhoff Voltage law, expressed as:

$$v_a = R_a i_a + L_a \frac{d}{dt} i_a + K_e \omega \tag{1}$$

where v_a is the applied armature voltage, R_a and L_a are the armature resistance and inductance respectively, i_a is the armature current, and ω is the angular velocity of the motor. K_e is the back electromotive force constant.

In parallel, the inner PI controller generates a control voltage v_a based on the difference between the reference and actual armature currents:

$$v_a = K_{pi}(i_{a\text{-ref}} - i_a) + K_{ii} \int (i_{a\text{-ref}} - i_a)\, dt \tag{2}$$

where K_{pi} and K_{ii} are the proportional and integral gains of the current controller, and i_{a_ref} is the current reference signal provided by the speed loop.

In sequence, the outer PI controller generates the current reference i_{a_ref} by comparing the desired and actual angular velocities:

$$i_{a\text{-ref}} = K_{p\omega}(\omega_{ref} - \omega) + K_{i\omega} \int (\omega_{ref} - \omega)\, dt \tag{3}$$

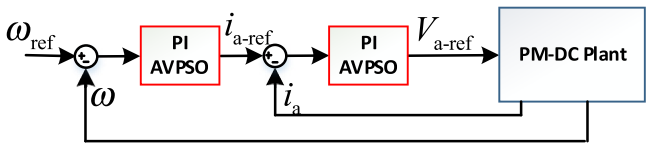


Figure 2. Cascaded PI controls scheme of PMDC motor

By substituting Equation (3) into Equation (2) and equating it to the voltage equation (1) allows the entire cascaded PI structure to be merged with the motor dynamics. The result is a single differential equation that represents the closed-loop system:

$$K_{pi}(K_{p\omega}(\omega_{ref} - \omega) + K_{i\omega} \int (\omega_{ref} - \omega)dt - i_a) + K_{ii}(K_{p\omega} \int (\omega_{ref} - \omega)dt + K_{i\omega} \int (\omega_{ref} - \omega)\, dt - i_a)\, dt = R_a i_a + L_a \frac{d}{dt} i_a + K_e \omega \tag{4}$$

This expression combines the dynamic response of the electrical system, the current controller, and the speed controller into a unified formulation. The mechanical dynamics of the motor are also considered through the rotational equation:

$$J \frac{d}{dt} \omega + B\, \omega = K_t\, i_a \tag{5}$$

where J is the rotor inertia, B is the viscous damping coefficient, and K_t is the torque constant.

Therefore, the equations above represent the complete nonlinear model of the cascaded PI-controlled PMDC motor system. This formulation is particularly suitable for simulation and control design purposes, including linearization and state-space representation in subsequent analysis stages.

3.2. Proposed AVPSO strategy . The AVPSO algorithm is designed to balance exploration and exploitation by dynamically updating its key parameters. The position and velocity of each particle in the swarm are iteratively updated according to the following equations:

$$v_i^{k+1} = w^k \cdot v_i^k + c_1^k \cdot r_1 \circ (p_i^k - x_i^k) + c_2^k \cdot r_2 \circ (g^k - x_i^k) \tag{6}$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \tag{7}$$

Here, x_i^k and v_i^{k+1} are the position and velocity vectors of the i^{th} particle at iteration k, respectively; $p_i^k$is the personal best position of the ith particle up to k; $g^k$is the global best position found by the entire swarm up to k; $r_1, r_2 \sim U(0,1)$ are vectors of random scalars that are sampled uniformly; $\circ$ denotes element-wise multiplication; w^k is the inertia weight controlling the trade-off between global and local exploration; and c_1^k and c_2^k are the adaptive cognitive and social acceleration coefficients, respectively.

The w^k term is linearly decreased at each iteration according to:

$$w^k = w_{max} - \left(\frac{w_{max} - w_{min}}{k_{max}} \right) \cdot k \tag{8}$$

Where w_{max} =0.8, w_{min} =0.3, and k_{max} denoting the maximum number of iterations. The cognitive and social acceleration coefficients are also updated dynamically at each iteration using:

$$c_1^k = 2.5 - 2.0 \frac{k}{k_{max}} \tag{9}$$

$$c_2^k = 0.5 + 2.0 \cdot \frac{k}{k_{max}} \tag{10}$$

Therefore, this adaptive scheduling gradually shifts the swarm behavior from individual exploration (c_1^k) toward global convergence (c_2^k) as the iteration progresses [22,23].

The updated position of each particle is constrained by boundary conditions as follows:

$$x_i^{k+1} = min\left(max\left(x_i^{k+1}, lb\right), ub\right) \tag{11}$$

where lb and ub represent the lower and upper bounds of the search space, respectively. The optimization process terminates when either the maximum number of iterations $k_{\max}$ is reached or the change in the global best cost $|J^k - J^{k-1}|$ falls below a predefined threshold ε over several consecutive iterations, indicating convergence.

In AVPSO, the optimization is guided by a scalar objective function that quantifies the performance of each candidate solution. For PI controller tuning in the PMDC drive system, the objective function is defined to penalize both steady-state and transient errors in the electrical and mechanical responses. Specifically, the objective function is expressed as:

$$J(\theta) = E_\omega + E_{i_a} \tag{12}$$

where, $\Theta \in R^n$ represents the decision variables, typically control parameters such as gains or time constants, E_ω is the accumulated error in angular velocity, and $E_{i_a}$is the accumulated error in armature current, both error terms are computed over a fixed simulation or control horizon. However, the weighting factors in the cost function are set equally (both equal to one) in this study. This reflects driving scenarios where speed regulation and current control are given equal priority to achieve accurate tracking and maintain motor efficiency. Since the objective is to enhance overall dynamic and steady-state performance, both E_ω, and E_{ia} are treated with the same importance in the optimization process. Therefore, each component is defined as:

$$E_\omega = \int \left(\omega_{ref} - \omega\right)^2 dt \tag{13}$$

$$E_{i_a} = \int \left(i_{a-ref} - i_a\right)^2 dt \tag{14}$$

Therefore, the scalar cost function directs AVPSO toward parameter sets that achieve fast and accurate dynamic responses by minimizing overshoot and steady-state error [16].

Constraints on system parameters and dynamic behavior are handled within the AVPSO framework using a combination of bounded search space enforcement and simulation-based feedback. Each decision variable is confined to a predefined range:

$$\theta_j \in \left[\theta_j^{min}, \theta_j^{max}\right], j = 1, 2, \ldots, n \tag{15}$$

For ensuring the search remains within the feasible solution space, each position of the particle is adjusted after every update using boundary projection:

$$\theta_j^{k+1} = min\left(max\left(\theta_j^{k+1}, \theta_j^{min}\right), \theta_j^{max}\right) \tag{16}$$

Therefore, this constraint-handling strategy effectively prevents the optimizer from exploring invalid regions of the search space

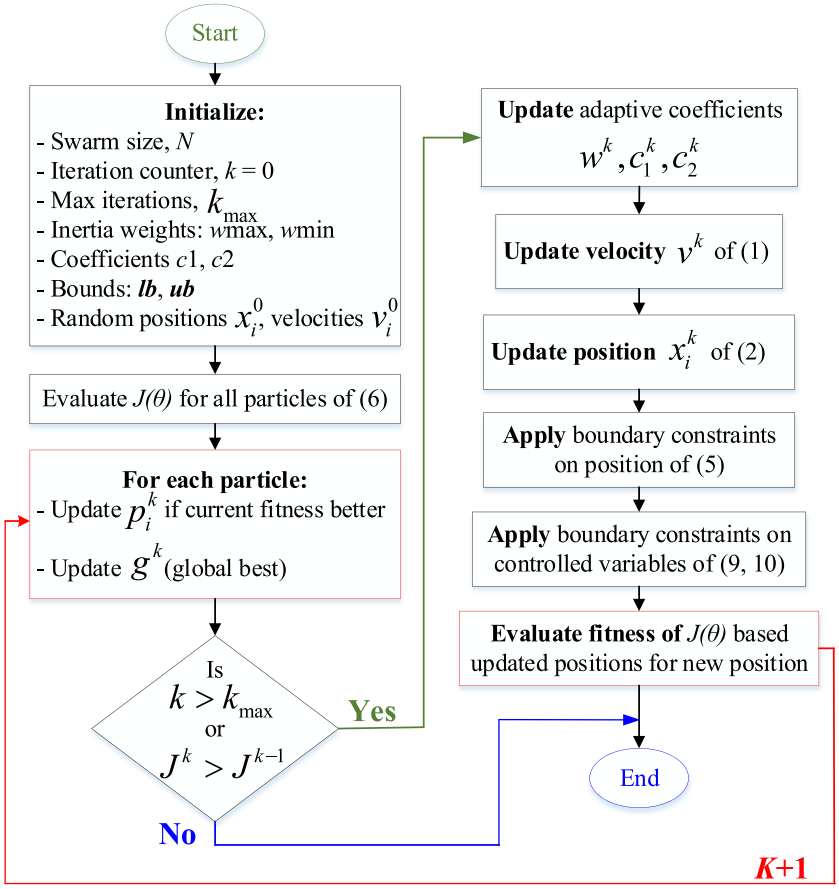


Figure 3. Flowchart of the adaptive variable PSO algorithm

Table 3. *Comparison of the proposed adaptive PSO and classical PSO Features*

Feature	Adaptive PSO	Classical PSO
Adaptive inertia w	Yes	No
Adaptive c1, c2	Yes	No
Cost history	Yes	No
Early stopping	Yes	No
More exploration	Yes	No
Practical convergence	High	Low

and contributes to the overall stability and convergence of the optimization process [24,25].

The dynamic system model evaluates the fitness function while implicitly satisfying physical and operational constraints such as current limits, voltage saturation, and speed thresholds. These constraints are not enforced explicitly but are reflected through simulation feedback. If a given parameter set leads to instability or violates system limits, the objective function $J(\theta)$ yields a high cost. This formulation enables AVPSO to reject infeasible solutions and favor parameter sets that result in stable and accurate system behavior [22,25].

Figure 3 shows the flowchart of the AVPSO algorithm. The process begins with initialization of particle positions and velocities within defined bounds. At each iteration, inertia weight and acceleration coefficients are updated adaptively. The fitness of each particle is evaluated using the objective function, and personal and global best positions are updated accordingly. This process continues until either the cost improvement becomes negligible over successive iterations or the maximum number of iterations is reached.

Table 3 summarizes the main differences between the proposed APSO and the classical PSO approach. The adaptive version integrates features such as dynamic adjustment of inertia weight, cognitive and social coefficients, and cost history tracking. These enhancements improve its capability to explore the solution space effectively and achieve faster, more reliable convergence.

In contrast, the classical PSO lacks these adaptive mechanisms and typically exhibits lower convergence speed and limited exploration space.

The proposed AVPSO and classical PSO methods were compared via a performance evaluation in which a PMDC drive system was modeled using the MATLAB/Simulink environment. The key parameters of the motor and the PSO techniques are listed in Table 4. The initial gains of the four PI controllers, namely, the proportional and integral gains for both the velocity and armature current loops, were randomly generated for both optimization methods. These gains were selected within the same predefined ranges. Specifically, the lower limits were set to 1, 0.1, 1, and 1, while the upper limits were set to 300, 5, 50, and 200, respectively. These limits ensured that both techniques were evaluated over an identical search space, providing a fair and consistent basis for performance comparison.

Meanwhile, the cost function defined in Equation (12) guides the optimization process. Therefore, the procedure is executed

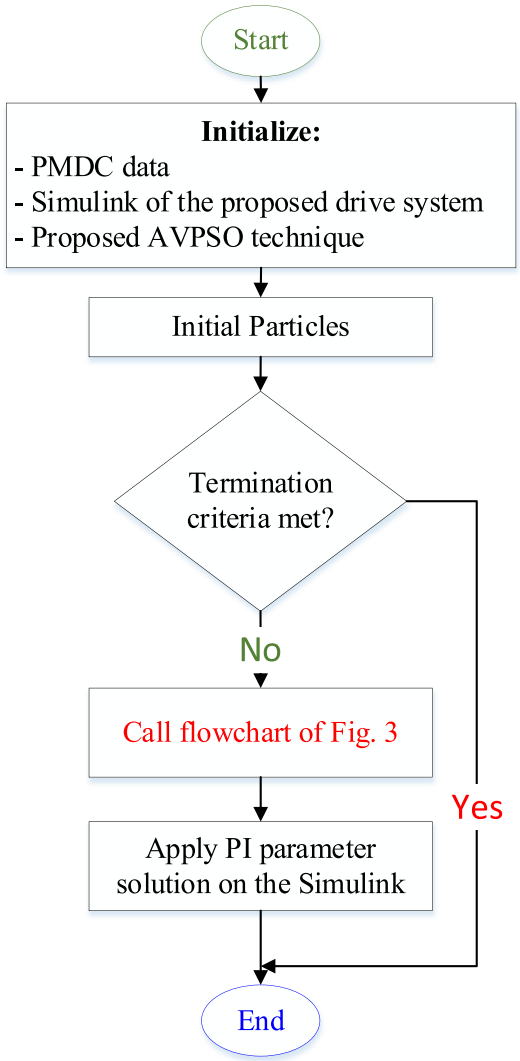


Figure 4. Implementation of the off-line AVPSO technique for optimizing the PI Parameters

with a population size of 50 and a maximum of 100 generations, as illustrated in Figure 4.

4. RESULTS AND DISCUSSION

4.1. Comparison between AVPSO and PSO. A comparison between the proposed AVPSO technique and the classical PSO technique is presented in this section. Figure 5 presents the convergence behavior of the cost function for both the classical PSO and the proposed AVPSO. Although both methods achieved the same final cost value of 1,000,000, AVPSO converged faster, terminating at the 51st iteration.

Figure 6 illustrates the variations in the cognitive acceleration coefficient, social acceleration coefficient, and inertia weight throughout the AVPSO process. The adaptive behavior of these parameters enhances the balance between exploration and exploitation during tuning. Thus, the PI controller parameters obtained using each method differ, as listed in Table 5.

Although the cost function converges to the same value for both the AVPSO and classical PSO, this does not imply that the resulting PI controller parameters are identical. The cost

Table 4. *Key parameters of the motor drive system and PSO techniques*

Symbols	Name	Value	Unit
R_a	armature resistance	0.24	Ω
L_a	armature inductance	18	mH
K_e	back electromotive fore constant	0.7237	V · s/rad
K_t	torque constant	0.7237	N · m/A
B	viscous damping coefficient	0.01	N · m · s/rad
J	moment of inertia	0.5	$Kg.m^2$
N	number of particles (swarm size)	50	particles
C_1	cognitive acceleration coefficient of classical PSO	0.5	-
C_2	social acceleration coefficient of classical PSO	0.5	-
w	inertia weight of classical PSO	0.2	-
w_{max}	maximum inertia weight of proposed AVPSO	0.8	-
w_{min}	minimum inertia weight of proposed AVPSO	0.3	-

function evaluates the overall system performance based on predefined criteria such as tracking error, control effort, and response time. Because the optimization landscape may contain multiple parameter sets that yield similar cost values, each algorithm can converge to a different solution that satisfies the objective equally well. In the case of AVPSO, the dynamic adjustment of the inertia weight and acceleration coefficients enhances search-space exploration. This adaptive behavior enables AVPSO to identify distinct, yet equally optimal, PI parameter sets compared to classical PSO. The following section presents two numerical scenarios for a further comparison of the PMDC motor performance under both tuning methods.

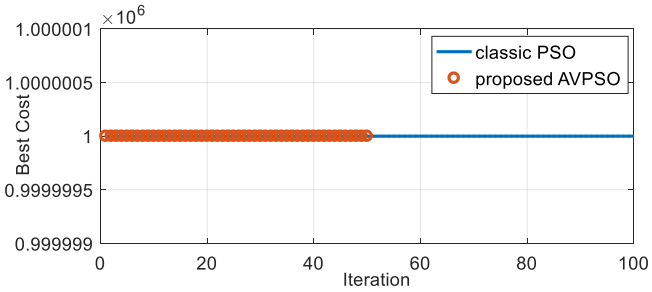


Figure 5. Cost function convergence of classical PSO and proposed AVPSO

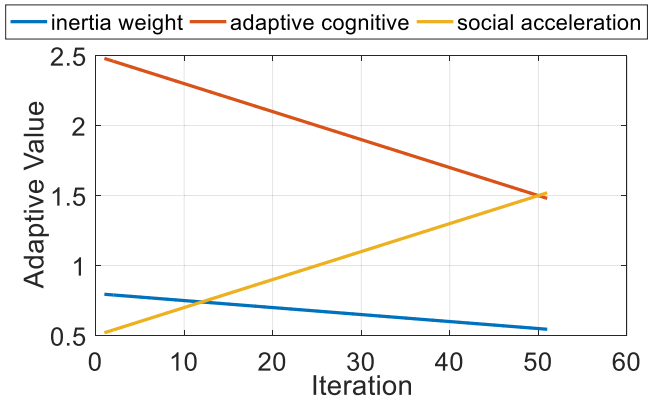


Figure 6. Implementation of the off-line AVPSO technique for optimizing the PI Parameters

4.2. Four quadrant operation scenario. Figures 7–9 compare the PMDC motor performance under a four-quadrant operation scenario using the proposed AVPSO-based PI tuning method and the classical PSO approach. In this setup, the reference traction torque is maintained at 7 Nm, while the angular velocity alternates between 600 RPM (62.83 rad/s) in both forward and reverse directions. The analysis focuses on system behavior during quadrant transitions and dynamic responses.

As shown in Figure 7, the armature voltage maintains the correct polarity and magnitude across all quadrants. In forward operation, the voltage reaches a peak of 48.5 V for both tuning methods. In the reverse operation, the AVPSO method yields a peak of –48.8 V, compared to –42.5 V for classical PSO. These results indicate that AVPSO provides more symmetrical and consistent voltage regulation throughout the operating range. The AVPSO method showed a sharper voltage spike at 12 s during the quadrant transition (Figure 7(a)). This is attributed to the rapid response to error changes. Despite the higher transient, the system remains stable, as evidenced by the smooth settling in the angular velocity (Figure 7(b)), armature current (Figure 7(c)), and torque (Figure 7(d)), with no signs of instability or oscillatory behavior.

The angular-velocity response (Figure 7(b)) confirms that the improved dynamic performance of the proposed AVPSO-based tuning method is comparable to that of PSO. In the forward rotation, the settling time was reduced by 15.6%, reaching approximately 6.2 s. In the reverse rotation, the settling time decreased

Table 5. *Optimized PI controller parameters*

	Proposed AVPSO	Classic PSO
$K_{p\omega}$	227.8	123.2
$K_{i\omega}$	1.1	1.1
K_{pi}	23.8	6.9
K_{ii}	182.4	5.1

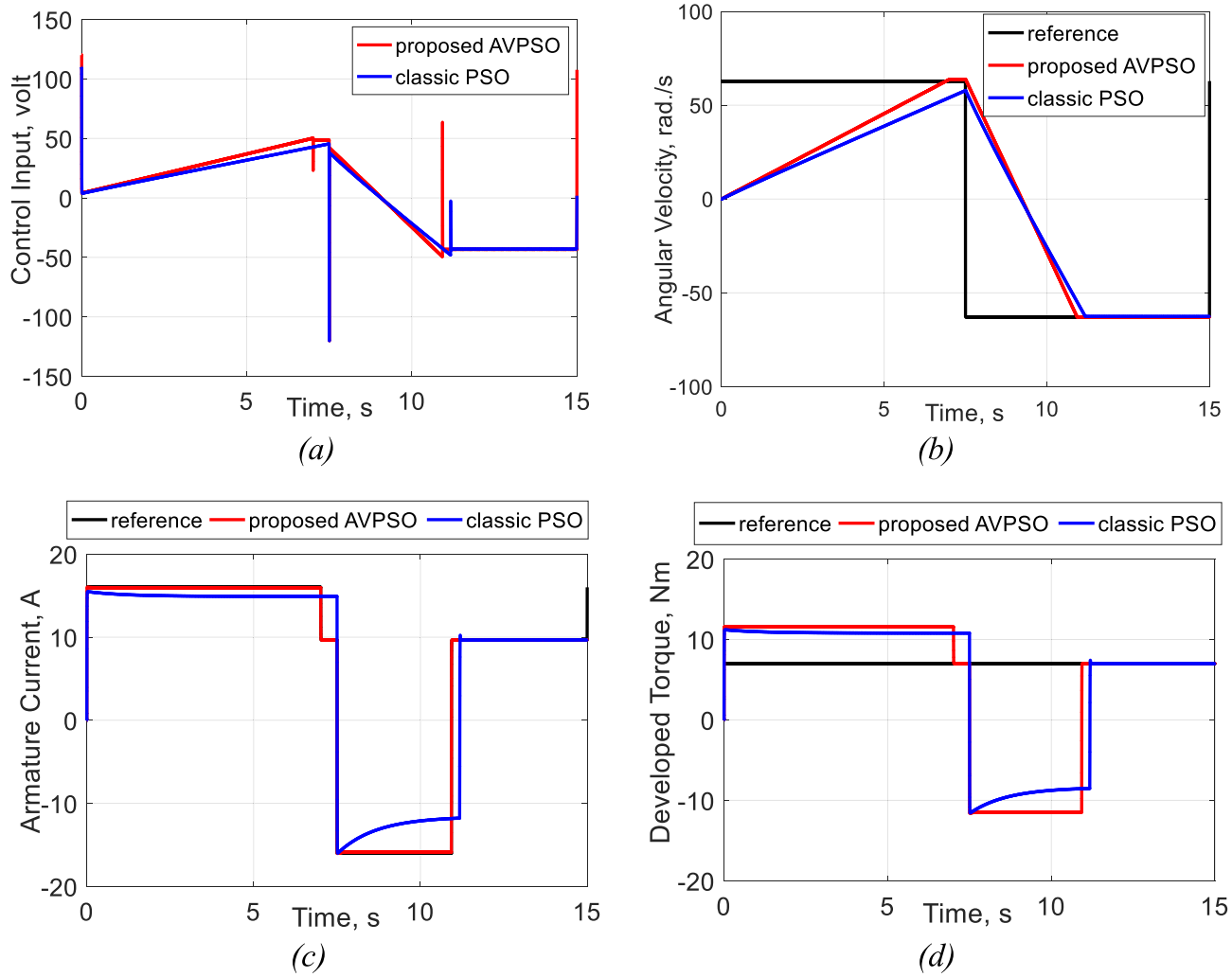


Figure 7. (a) Armature voltage, (b) angular velocity, (c) armature current, and (d) developed torque under four-quadrant operation

by 2.1%, achieving a value of approximately 3.26 s. Additionally, the AVPSO method eliminates steady-state error in both directions, whereas the classical PSO method shows a 6.2 rad/s error during forward operation.

The armature current response (Figure 7(c)) further highlights the advantages of the proposed method. The outer PI controller generated reference currents of 16 and 9.7 A for the forward and reverse motions, respectively. Under classical PSO tuning, the actual current failed to accurately track the reference during transients and exhibited a steady-state error of 1.02 A in forward operation. In contrast, AVPSO ensures precise current tracking with no steady-state errors, even during dynamic transitions. These improvements in current and voltage behaviors enhance both the electrical input power and mechanical output power. In motoring mode, AVPSO ensures smooth energy delivery, whereas in regenerative mode, it handles the reverse energy flow more effectively, as evidenced by the higher and more stable negative voltage.

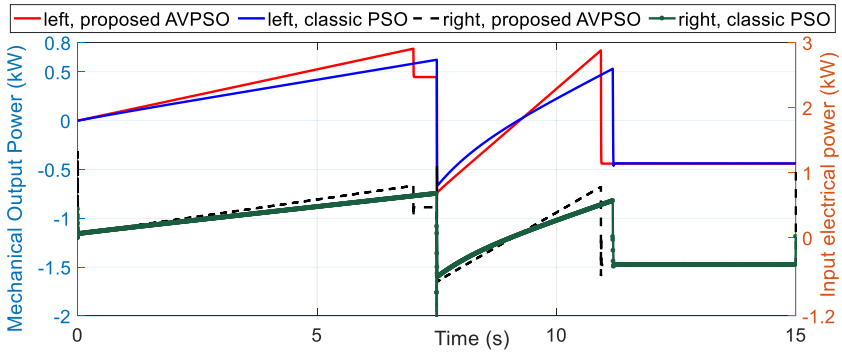


Figure 8. Power consumption under four-quadrant operation

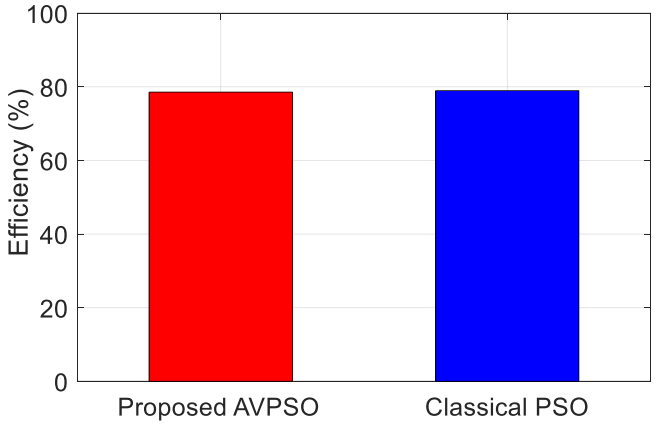


Figure 9. Motor efficiency under four-quadrant operation

As shown in Figure 7(d), the developed torque reaches the reference value more quickly for AVPSO than PSO, with faster responses by 0.5 and 1 s in the forward and reverse directions, respectively. Figure 8 further demonstrates that AVPSO improves energy efficiency with reduced power consumption during both motoring and regenerative operations. Moreover, Figure 9 shows that the AVPSO method achieves better dynamic performance with faster settling, improved tracking, and elimination of steady-state errors, while maintaining a motor efficiency comparable to classical PSO at 78.8%, indicating no significant gain in steady-state efficiency. Finally, Table 6 summarizes the performance indicators for both tuning techniques under a four-quadrant operation scenario.

4.3. Torque disturbance operation scenario.

Figures 10–12 illustrate the PMDC motor response under a torque-disturbance scenario by comparing the performance of the proposed AVPSO-based PI tuning with that of classical PSO. Following the standard-load case, this scenario was used to evaluate motor performance under more demanding torque conditions to assess the dynamic response and energy-conversion efficiency of both tuning methods. As shown in Figure 10(a), the armature voltage during forward operation remains positive, peaking at approximately 95 V for AVPSO and 97 V for classical PSO. Although the PSO method yields a slightly higher peak voltage, AVPSO results in a faster and more stable response to sudden changes in torque demand. In addition, a more pronounced voltage spike at 12 s was observed using AVPSO because of its faster corrective action during the sudden torque reduction. This aggressive response improves error minimization but leads to a transient overshoot. However, as seen in the corresponding angular velocity (Figure 10(b)), armature current (Figure 10(c)), and torque (Figure 10(d)) plots, the voltage stabilizes quickly without introducing oscillations, confirming that the system stability was maintained.

Figure 10(b) shows the angular velocity response, where the proposed AVPSO method demonstrates enhanced performance under high-torque conditions. Compared to PSO, the settling time using AVPSO was reduced by 10.7%, reaching approximately 10.3 s. In addition, the steady-state error was significantly reduced . The classical PSO resulted in a steady-state error of 7.7 rad/s, whereas AVPSO reduced this value to 3.7 rad/s, indicating superior speed regulation.

Figures 10–12 illustrate the PMDC motor response under a torque-disturbance scenario by comparing the performance of the proposed AVPSO-based PI tuning with that of classical PSO. Following the standard-load case, this scenario was used to evaluate motor performance under more demanding torque conditions to assess the dynamic response and

Table 6. *Performance comparison under four-quadrant operation*

Performance Metric	AVPSO-Based Tuning	Classical PSO-Based Tuning	Observation
Peak Armature Voltage (Forward)	48.5 V	48.5 V	Same performance
Peak Armature Voltage (Reverse)	–48.8 V	–42.5 V	AVPSO shows better symmetry in voltage regulation
Settling Time (Forward)	6.2 s	~7.35 s	AVPSO reduced time by 15.6%
Settling Time (Reverse)	3.26 s	~3.33 s	AVPSO reduced time by 2.1%
Speed Steady-State Error (Forward)	0 rad/s	6.2 rad/s	AVPSO eliminates steady-state error
Armature Current Tracking Error	0 A	1.02 A (forward)	AVPSO achieves accurate current tracking
Torque Response Time (Forward)	Faster by 0.5 s	Slower	AVPSO responds quicker
Torque Response Time (Reverse)	Faster by 1.0 s	Slower	AVPSO responds quicker
Power Consumption	Higher during transitions	Lower	AVPSO enhanced motor performance without compromising the efficiency of classic PSO technique
Motor Efficiency	78.8%	78.8%	Comparable for both techniques

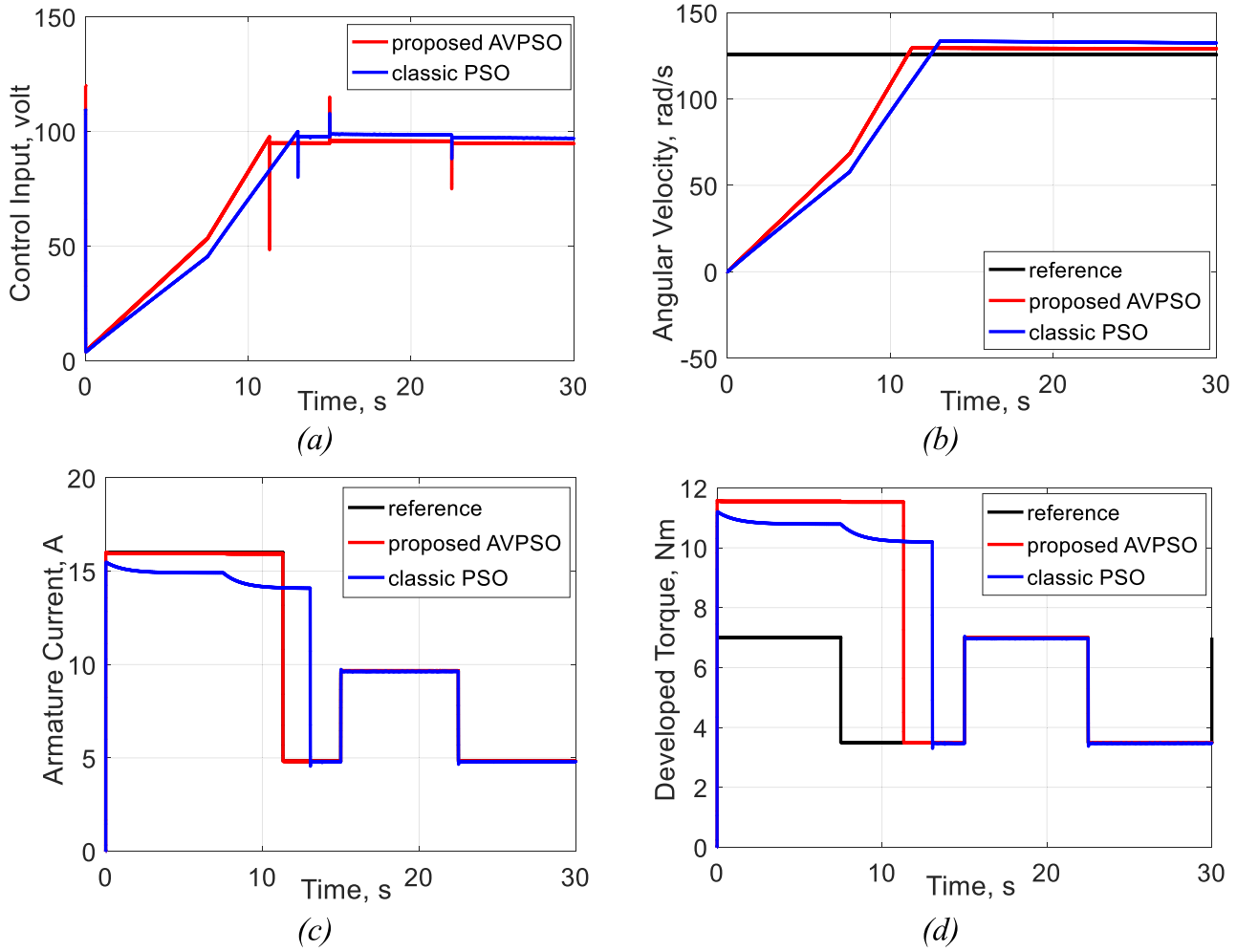


Figure 10. (a) Armature voltage, (b) angular velocity, (c) armature current, and (d) developed torque under torque-disturbance operation

energy-conversion efficiency of both tuning methods. As shown in Figure 10(a), the armature voltage during forward operation remains positive, peaking at approximately 95 V for AVPSO and 97 V for classical PSO. Although the PSO method yields a slightly higher peak voltage, AVPSO results in a faster and more stable response to sudden changes in torque demand. In addition, a more pronounced voltage spike at 12 s was observed using AVPSO because of its faster corrective action during the sudden torque reduction. This aggressive response improves error minimization but leads to a transient overshoot. However, as seen in the corresponding angular velocity (Figure 10(b)), armature current (Figure 10(c)), and torque (Figure 10(d)) plots, the voltage stabilizes quickly without introducing oscillations, confirming that the system stability was maintained.

Figure 10(b) shows the angular velocity response, where the proposed AVPSO method demonstrates enhanced performance under high-torque conditions. Compared to PSO, the settling time using AVPSO was reduced by 10.7%, reaching

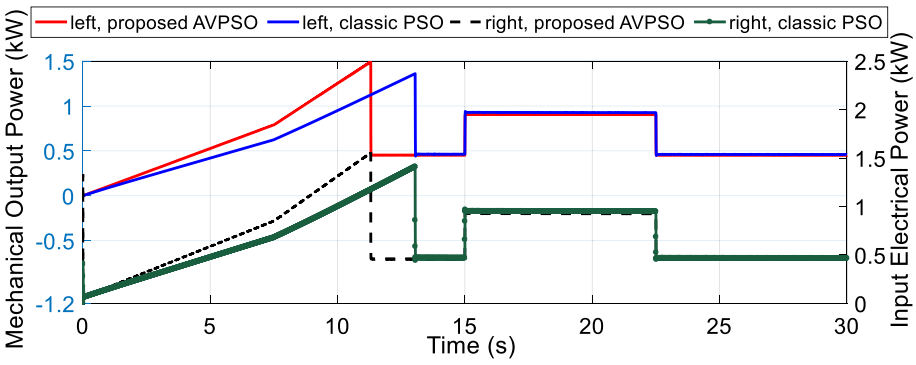


Figure 11. Power consumption under torque disturbance operation

approximately 10.3 s. In addition, the steady-state error was significantly reduced. The classical PSO resulted in a steady-state error of 7.7 rad/s, whereas AVPSO reduced this value to 3.7 rad/s, indicating superior speed regulation.

The good performance of AVPSO is further supported by the armature current response, as shown in Figure 10(c). The outer PI controller assigned reference current values of 16 and 4.7 A in the forward and reverse directions, respectively. Under classical PSO tuning, the motor current deviates from the reference value during transients and exhibits a steady-state error of approximately 1 A. In contrast, AVPSO achieved accurate current tracking with no steady-state error, even under dynamic conditions.

Based on the behavior of both the voltage and current, the electrical input power and mechanical output power can be inferred. During motoring mode, AVPSO ensures a smooth and consistent power profile, contributing to efficient energy transfer. Although the regenerative operation is less dominant in this scenario, the controller maintains a stable reverse energy flow and operational reliability when required.

As shown in Figure 10(d), the developed torque reaches the reference value 1.7 s faster with AVPSO than with the classical PSO, and exhibits reduced oscillations during convergence. Consequently, the power consumption response in Figure 11 is more efficient and stable under AVPSO control, indicating improvements in both transient and steady-state operations. Moreover, Figure 12 shows that the AVPSO method achieves superior dynamic performance with faster settling, improved tracking, and reduced steady-state error, while maintaining a motor efficiency comparable to that of classical PSO at approximately 91.0%, i.e., no significant gain in steady-state efficiency. Finally, Table 7 summarizes the key performance indicators for both tuning techniques under the torque–disturbance scenario.

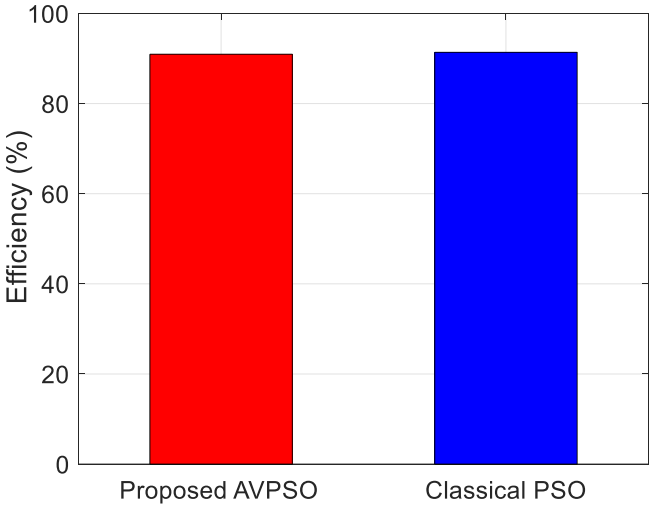


Figure 12. Motor efficiency under torque disturbance operation

4.4. Comparison with existing PSO-based PI tuning methods. Table 8 compares PSO-based PI/PID tuning methods for PMDC drives. Previous studies have shown that traditional PSO improves dynamic tracking, reduces overshoot, and handles load transients more effectively than manual tuning or GA-based optimization [26–28]. However, these approaches rely on static PSO parameters and do not consider four-quadrant operations, transient load variations, or the trade-off between speed and current errors through weighted tuning. In contrast, our AVPSO method introduces adaptive velocity update parameters, achieving faster convergence and improved tracking while maintaining comparable motor efficiency (~78–91%) while ensuring stability under regenerative-braking and torque-disturbance conditions.

Table 7. *Performance Comparison under Torque Disturbance Scenario*

Performance Metric	AVPSO-Based Tuning	Classical PSO-Based Tuning	Observation
Peak Armature Voltage (Forward)	~95 V	~97 V	Classical yields slightly higher peak, but less stable
Settling Time (Angular Velocity)	10.3 s	~11.54 s	AVPSO reduces time by 10.7%
Speed Steady-State Error	3.7 rad/s	7.7 rad/s	AVPSO shows better speed regulation
Armature Current Tracking Error	0 A	~1 A	AVPSO achieves accurate current tracking
Torque Response Time	Faster by 1.7 s	Slower	AVPSO reaches reference torque sooner
Torque Oscillation	Reduced	Noticeable	AVPSO shows smoother convergence
Power Consumption	Higher during transitions	Lower	AVPSO technique enhanced motor performance without compromising the efficiency of classic PSO technique
Motor Efficiency	91.0%	91.0%	Comparable efficiency for both techniques

Table 8. *Summary of selected PSO-based PI/PID tuning methods for PMDC motor drives*

Reference	Control Structure	Objective (Cost)	Gains over classical tuning	How it differs from your AVPSO
[26]	Triloop PID (hysteretic control of firing angle)	Minimize $e_\omega + e_I + e_R$ (singleweight sum of speed, current deviation, ripple)	Smoother start torque, fast acceleration, stable trajectory tracking vs classical tuning	No adaptation of PSO coefficients (fixed w,c_1,c_2); optimization of three loops only; no fourquadrant effort; no explicit efficiency reporting
[27]	Triloop errordriven PID	Integral of timesquared error; dynamic error criterion	Improved transient, fast settling (~4 s), zero overshoot under load/disturbance	Standard PSO; no adaptive velocity; performance verified in simulation only; various transient scenarios, but no regenerative mode
[28]	Cascade P + PI (no ripple loop, only current and speed)	Integral time absolute error over speed and position errors under load conditions	PSO eliminated position overshoot (vs ~7 % class., ~4 % genetic algorithm) and suppressed deviation ~12.0°under loaded start	Only two PI loops; constant PSO; no fourquadrant or regenerative analysis; emphasis on overshoot; no torque disturbance scenario

5. CONCLUSIONS

This study developed an AVPSO approach for tuning PI controllers of a PMDC motor-drive system for LEVs. This method introduces an adaptive adjustment of the inertia weight and acceleration coefficients to improve convergence and parameter tuning. The simulation results demonstrate that AVPSO-based tuning enhances the key system performance parameter. The classical PSO method exhibited residual speed and current errors during forward operation, whereas the AVPSO achieved a 15.6% reduction in settling time and eliminated steady-state errors in both rotational directions. In addition, the AVPSO approach improves the quality of the power consumption response during transient events, particularly under regenerative braking. However, it is important to note that both methods maintained comparable motor efficiencies, with no significant efficiency gains with the proposed method. These findings confirm that the AVPSO algorithm provides an accurate and responsive control strategy for PI parameter optimization in LEV drive systems. Its effectiveness in reducing tracking errors and enhancing the dynamic behavior supports its applicability for intelligent controller designs.

In future work, the proposed cost function will be further developed to enable online computation of the drive controller parameters. This enhancement will enable the control system to adapt in real-time under varying loads and operating conditions in practical LEV applications. Additionally, experimental validation through HIL testing will be performed to assess the system stability and implementation feasibility.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Mohammed Aldhaif Allah – *Control and Instrumentation Engineering Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia;* 0009-0003-7275-5824
Email: s202274560@kfupm.edu.sa

Corresponding Author

Moustafa Magdi Ismail – *Research Mentor, Interdisciplinary Research Center for Sustainable Energy Systems (IRC-SES), King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia;* 0000-0001-5554-6084
Email: moustafa.mohamed@kfupm.edu.sa

ACKNOWLEDGEMENTS

This research is a direct outcome of an internal project (INSE2521), conducted at the Interdisciplinary Research Center for Sustainable Energy Systems (IRC-SES), King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia. The authors sincerely appreciate the support and resources provided by the IRC-SES, which contributed significantly to the successful completion of this work.

REFERENCES

(1) D. Rimpas, S. D. Kaminaris, D. D. Piromalis, G. Vokas, K. G. Arvanitis, and C.-S. Karavas, "Comparative Review of Motor Technologies for Electric Vehicles Powered by a Hybrid Energy Storage System Based on Multi-Criteria Analysis," Energies, vol. 16, p. 2555, 2023.

(2) A. Eldho Aliasand and F. T. Josh, "Selection of Motor foran Electric Vehicle: A Review," Materials Today: Proceedings, vol. 24, pp. 1804–1815, 2020/01/01/ 2020.

(3) A. Joseph Godfrey and V. Sankaranarayanan, "A new electric braking system with energy regeneration for a BLDC motor driven electric vehicle," Engineering Science and Technology, an International Journal, vol. 21, pp. 704–713, 2018/08/01/ 2018.

(4) A. Isah, A. Mohammed, and A. Hamza, "Electric Power-Assisted Steering: A Review," in 2019 2nd International Conference of the IEEE Nigeria Computer Chapter (NigeriaComputConf), 2019, pp. 1–6.

(5) E. Sangeetha and V. Ramachandran, "Different topologies of electrical machines, storage systems, and power electronic converters and

their control for battery electric vehicles—a technical review," Energies, vol. 15, p. 8959, 2022.

(6) T. Sutikno, N. R. N. Idris, and A. Jidin, "A review of direct torque control of induction motors for sustainable reliability and energy efficient drives," Renewable and Sustainable Energy Reviews, vol. 32, pp. 548–558, 2014/04/01/ 2014.

(7) W. Słowik, P. Piątek, T. Dziwiński, and J. Baranowski, "Selected current sensing circuits for motor control application," Pomiary Automatyka Robotyka, vol. 21, 2017.

(8) Z. Bitar, A. Sandouk, and S. Al Jabi, "Testing the performances of DC series motor used in electric car," Energy Procedia, vol. 74, pp. 148–159, 2015.

(9) N. M. Morris and N. M. Morris, "DC Machines," Electrical Principles III, pp. 44–63, 1978.

(10) E. Soressi, "New life for old compound DC motors in industrial applications?," in 2012 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES), 2012, pp. 1–6.

(11) Y. S. Chang, Y. C. Luo, and P. C. Shih, "AIR: Agent and Ontology-Based Information Retrieval Architecture for Mobile Grid," in 2008 IEEE Asia-Pacific Services Computing Conference, 2008, pp. 650–655.

(12) M. Onda, M. Misawa, T. Kojima, N. Aya, A. Seto, and T. Yamane, "A stratospheric LTA stationary platform for telecommunication and environmental protection," in SICE '99. Proceedings of the 38th SICE Annual Conference. International Session Papers (IEEE Cat. No.99TH8456), 1999, pp. 1227–1232.

(13) W. Xu, M. M. Ismail, Y. Liu, and M. R. Islam, "Parameter optimization of adaptive flux-weakening strategy for permanent-magnet synchronous motor drives based on particle swarm algorithm," IEEE Transactions on Power Electronics, vol. 34, pp. 12128–12140, 2019.

(14) M. M. Ismail, M. Al-Dhaifallah, H. Rezk, H. U. Rahman Habib, and S. A. Hamad, "Optimizing battery discharge management of PMSM vehicles using adaptive nonlinear predictive control and a Generalized Integrator," Ain Shams Engineering Journal, p. 103169, 2024/11/19/ 2024.

(15) Y. Yu, Y. Pan, Q. Chen, Y. Hu, J. Gao, Z. Zhao, et al., "Multi-Objective Optimization Strategy for Permanent Magnet Synchronous Motor Based on Combined Surrogate Model and Optimization Algorithm," Energies, vol. 16, p. 1630, 2023.

(16) L. Kong, H. Zhang, T. Zhang, J. Wang, C. Yang, and Z. Zhang, "Adaptive Control Parameter Optimization of Permanent Magnet Synchronous Motors Based on Super-Helical Sliding Mode Control," Applied Sciences, vol. 14, p. 10967, 2024.

(17) H. Xu, J. Zhang, J. Liu, Y. Cao, and A. Ma, "Optimization of Ship Permanent Magnet Synchronous Motor ADRC Based on Improved QPSO," Applied Sciences, vol. 15, p. 1608, 2025.

(18) G. Yuan, Y. Tao, S. Bozhko, P. Wheeler, T. Dragicevic, and C. Gerada, "Neural Network aided PMSM multi-objective design and optimization for more-electric aircraft applications," Chinese Journal of Aeronautics, vol. 35, pp. 233–246, 2022.

(19) C. Sun, F. Wen, W. Xiong, H. Wang, and H. Shang, "Multi-objective comprehensive teaching algorithm for multi-objective optimisation design of permanent magnet synchronous motor," IET Electric Power Applications, vol. 14, pp. 2564–2576, 2020.

(20) M. M. Ismail, W. Xu, X. Wang, A. K. Junejo, Y. Liu, and M. Dong, "Analysis and optimization of torque ripple reduction strategy of surface-mounted permanent-magnet motors in flux-weakening region based on genetic algorithm," IEEE Transactions on Industry Applications, vol. 57, pp. 4091–4106, 2021.

(21) N. Pragallapati, T. Sen, and V. Agarwal, "Adaptive Velocity PSO for Global Maximum Power Control of a PV Array Under Nonuniform Irradiation Conditions," IEEE Journal of Photovoltaics, vol. 7, pp. 624–639, 2017.

(22) M. P. Aghababa, A. M. Shotorbani, and R. M. Shotorbani, "An Adaptive Particle Swarm Optimization Applied to Optimum Controller Design for AVR Power Systems," International Journal of Computer Applications, vol. 11, pp. 22–29, 2010.

(23) X. Li, K. Mao, F. Lin, and X. Zhang, "Particle swarm optimization with state-based adaptive velocity limit strategy," Neurocomputing, vol. 447, pp. 64–79, 2021.

(24) A. R. Jordehi, "A review on constraint handling strategies in particle swarm optimisation," Neural Computing and Applications, vol. 26, pp. 1265–1275, 2015.

(25) M. S. Innocente and J. Sienz, "Constraint-handling techniques for particle swarm optimization algorithms," arXiv preprint arXiv:2101.10933, 2021.

(26) A. M. Sharaf and A. A. A. El-Gammal, "A novel Particle Swarm Optimization PSO tuning scheme for PMDC motor drives controllers," in Proc. Int. Conf. Power Eng., Energy and Electr. Drives, 2009, pp. 134–139.

(27) P. G. Medewar and R. K. Munje, "PSO-based PID controller tuning for PMDC motor," in Proc. Int. Conf. Energy Systems and Applications (ICESA), 2015, pp. 522–526.

(28) K. G. Abdulhussein, N. M. Yasin, and I. J. Hasan, "Comparison of cascade P-PI controller tuning methods for PMDC motor based on intelligence techniques," Int. J. Electr. Comput. Eng. (IJECE), vol. 12, p. 1, 2022.

KFUPM
Journal of Undergraduate Research International

Diamond Open Access 94

Article Electrical Engineering/Physics

Design and Modeling of a High-Efficiency Unit Concentrated Solar Thermoelectric Generator

Md. Habibur Rahman Aslam[1], Foyzul Karim[1] and Anisul Islam Suva[2]*

Cite https://doi.org/10.64589/juri/207994

Submitted: May 27, 2025 **Revised:** June 30, 2025 **Accepted:** July 07, 2025

ABSTRACT

A concentrated solar thermoelectric generator (CSTEG) is a low-maintenance, silent system that converts focused solar heat energy into electricity. However, its applications are constrained by low efficiency, high material costs, and complex thermal management. In this study, a CSTEG was designed and simulated in COMSOL to operate between 473 K and 353 K. Each leg of the CSTEG was segmented into two parts, and materials with a high dimensionless figure of merit (ZT) value were used to enhance efficiency. Compounds melt-spun with excess Te (Te-MS) ($Bi_{0.5}Sb_{1.5}Te_3$) and zone-melted $Bi_{0.4}Sb_{1.6}Te_3$ after hot deformation (HD-A-Sb1.6) were used as the p-type materials. Bi_2Te_3 -10 wt% nanocomposites and polycrystalline $Bi_2Te_{2.3}Se_{0.7}$ alloy were used as the n-type materials. The simulation yielded an open-circuit voltage of 46.2 mV, a maximum power of 28.32 mW, and a conversion efficiency of 5.69%. These results highlight the potential of material segmentation and selection in enhancing CSTEG performance for concentrated solar applications.

Keywords: solar TEG, COMSOL, thermoelectric, CSTEG

1. INTRODUCTION

Recent challenges such as global warming, climate change, and environmental degradation have driven the transition to renewable energy[1]. The sun, the source of all energy forms, enables electricity generation through photovoltaic (PV) and concentrating solar power (CSP) methods[2]. While PV converts sunlight directly into electricity, CSP converts it into heat, which can then be transformed into electricity via a thermoelectric generator (TEG) using the thermoelectric effect. Several concentrated solar thermoelectric generator (CSTEG) designs have been proposed. D. Kraemer developed a solar TEG with 4.6% efficiency operating between 473 K and 353 K[3]. L. Long et al. designed a CSTEG with 4.3% efficiency and 11.2 W maximum power at 106 × suns[4]. D. Kraemer also demonstrated 7.2% efficiency at 600 °C[5].

Optical concentrators have shown potential for improving solar thermoelectric generator (STEG) performance. Moh'd A et al. reported a 19.13% increase in power development using 20-sun concentration in a flat plate collector–TEG system[6]. Baranowski et al. modeled a STEG, achieving 15.9% efficiency under 100 suns and a hot-side temperature of 1000 °C[7].

Another effective approach involves using high-efficiency solar absorber surfaces that capture solar radiation, convert it to heat, and direct it to the TEG[8]. STEG performance can also be enhanced by electroforming carbon and copper layers on the hot side to improve thermal conductivity[9]. Kraemer et al. achieved 7.4% efficiency using spectrally selective and black paint absorbers at 211-sun concentration[10].

Despite such advances, CSTEG efficiency remains low. Improvements can be achieved by using advanced materials and optimizing geometry[11]. Segmentation further enhances performance by combining materials tailored to different temperature ranges, maximizing output across the thermal gradient[12]. Yang et al. developed and validated a segmented TEG model combining $Mg_3(Sb,Bi)_2$ and $Bi_{0.5}Sb_{1.5}Te_3$–GeTe, achieving 10.4% efficiency and 0.41 W output at $\Delta T = 440$ K[13].

In this study, a segmented CSTEG was designed based on the temperature gradient. COMSOL simulations show that it produces an open-circuit voltage of 46.2 mV and a maximum electrical power output of 28.32 mW, and a conversion efficiency of 5.69%. The use of thermally stable materials ensures uniform heat distribution and improved energy conversion.

2. SET UP SCHEMATIC AND DESIGN STRUCTURE

A hybrid solar energy conversion system combining a PV cell and thermoelectric module (Figure 1) was used to capture different parts of the solar spectrum. Parallel sunlight beams contain electromagnetic waves across various ranges, including ultraviolet (UV), visible (VIS), and infrared (IR) radiation. A concave mirror is a curved mirror that focuses parallel light rays onto a concentrated area. A selective concave mirror enhances the energy conversion by reflecting specific wavelengths. Sunlight is widely used in solar concentrators for efficient thermal and electrical conversion. A PV cell absorbs UV, VIS, and NIR light to generate electricity. An IR absorber behind the cell captures the

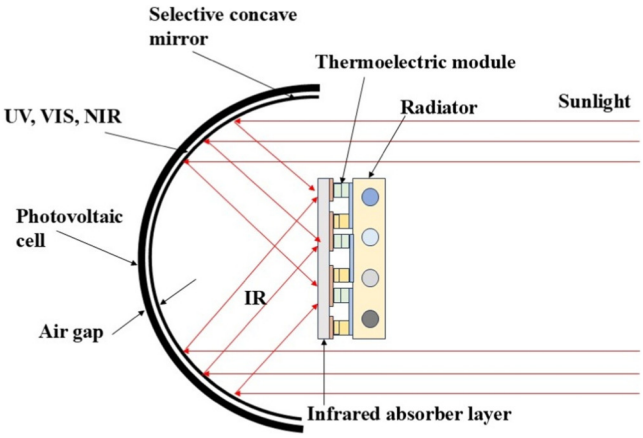


Figure 1. Schematic of hybrid solar energy converter attached with CSTEG

IR portion of sunlight channeled through a central opening or selectively transparent mirror section, heating the thermoelectric module's hot side. This module directly generates electricity from heat energy via the Seebeck effect. The radiator dissipates heat from the cold side and maintains a temperature gradient for better efficiency.

2.1. Design Structure. The designed CSTEG comprises thermoelectric materials sandwiched between hot ceramic (Al_2O_3) and cold ceramic surfaces. Ceramics exhibit excellent thermal stability, electrical insulation, and durability under temperature gradients. Heat flows from the hot-side heat source to the cold-side heat sink through thermoelectric materials. P- and n-type semiconductors are used as energy conversion materials. Two copper connectors positioned at the ends of each semiconductor leg near the cold side function as output terminals, facilitating the extraction of the direct current (DC) electrical power generated by the thermoelectric modules. Each thermoelectric leg was segmented into two parts, each of which was selected with an appropriate figure of merit (ZT) to match the operating temperature, thereby enhancing the overall conversion efficiency and performance. Figure 2 shows a schematic of the designed CSTEG module.

2.2. Boundary Conditions. The CSTEG is designed to operate within a temperature gradient, with the cold side maintained between 353 and 473 K. On the hot side, the p-type thermoelectric material functions effectively between 450 and 473 K, whereas the n-type material operates between 410 and 473 K. On the cold side, the p-type material covers 353 K to 450 K, and the n-type material operates between 353 and 410 K. These segmented temperature ranges are selected to optimize the performance and compatibility of the thermoelectric materials under realistic operating conditions. Table 1 presents the structural dimensions of the designed CSTEG unit.

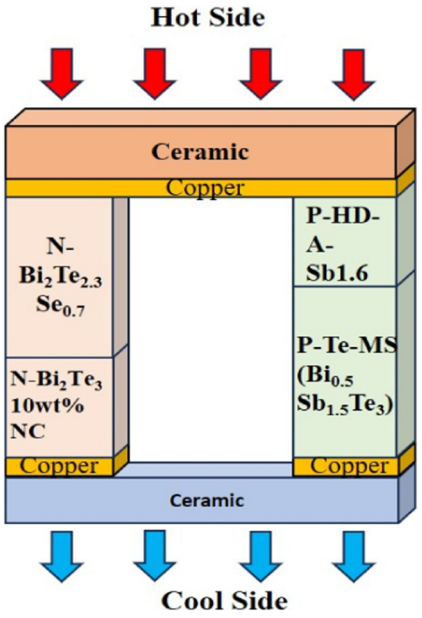


Figure 2. Schematic of designed unit CSTEG module

2.3. Materials Properties. The thermoelectric properties of materials, such as their Seebeck coefficient, electrical conductivity, and thermal conductivity, shown in Fig. 3, are critical for effective segmentation in TEGs. Selecting materials with a high ZT within a specific temperature range ensures efficient energy conversion [14]. Proper matching of these properties across segments minimizes interfacial losses and enhances overall TEG performance.

Figure 3(a) illustrates the ZT profiles of p-type $Bi_{0.5}Sb_{1.5}Te_3$ synthesized via liquid-phase compaction (Te-MS) and p-type $Bi_{0.4}Sb_{1.6}Te_3$. The ZT curves intersect at approximately 410 K, indicating a crossover in thermoelectric performance. Below this temperature, $Bi_{0.5}Sb_{1.5}Te_3$ exhibits a superior ZT, whereas above 410 K, $Bi_{0.4}Sb_{1.6}Te_3$ demonstrates enhanced performance. To

Table 1. *Geometrical measurement of designed CSTEG*

Materials	Height (mm)	Width (mm)	Depth (mm)	Operating Temperature (K)
Hot Heat exchanger (Al_2O_3)	0.5	3.9	1.4	573
Hot-side Copper	0.25	3.9	1.4	573
Hot-side p-type semiconductor ($Bi_{0.4}Sb_{1.6}Te_3$)	0.5	1.4	1.4	450–473
Cool side p-type semiconductor ($Bi_{0.5}Sb_{1.5}Te_3$)	1.0	1.4	1.4	353–450
Hot-side n-type semiconductor ($Bi_2Te_{2.3}Se_{0.7}$)	0.9	1.4	1.4	410–473
Cool side n-type semiconductor (Bi_2Te_3, 10 wt% NC)	0.6	1.4	1.4	353–410
Hot side (Terminal) Copper	0.2	1.4	1.4	353
Cool Heat exchanger (Al_2O_3)	0.3	3.9	1.4	353

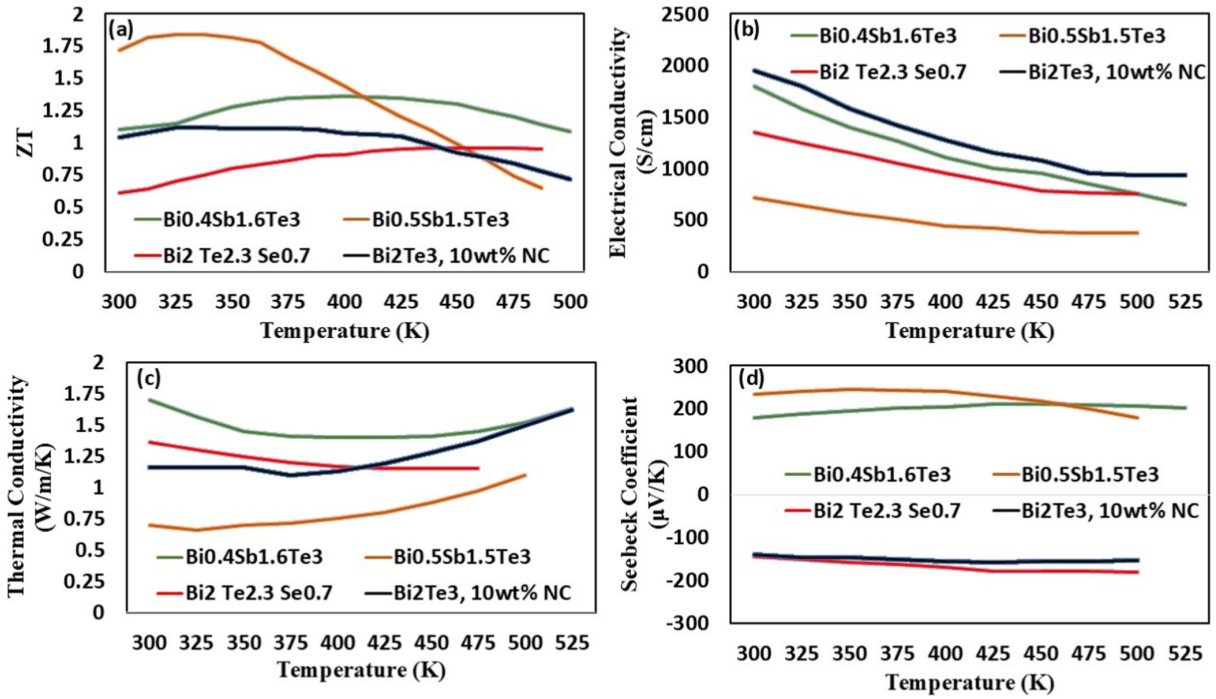


Figure 3. (a) ZT, (b) Electrical conductivity, (c) Thermal conductivity, (d) Seebeck coefficient graph against temperature for $Bi_{0.4}Te_{1.6}Te_3$ [15], $Bi_{0.5}Te_{1.5}Te_3$ [16], $Bi_2Te_{2.3}Se_{0.7}$ [17], Bi_2Te_3– wt% Nanocompound (NC) [18]

optimize the overall efficiency of the p-type leg, a segmented configuration is adopted, employing $Bi_{0.5}Sb_{1.5}Te_3$ at temperatures below 410 K and $Bi_{0.4}Sb_{1.6}Te_3$ at temperatures above this point.

Similarly, the ZT curves of n-type $Bi_2Sb_{2.3}Te_{0.7}$ and n-type Bi_2Te_3-10 wt% nanocompound (NC) intersect at approximately 445 K, signifying a transition in thermoelectric performance. Below this temperature, Bi_2Te_3-10 wt% NC delivers a higher ZT, whereas $Bi_2Sb_{2.3}Te_{0.7}$ exhibits improved performance at elevated temperatures. To enhance the overall performance of the n-type leg, a segmented design is employed, with Bi_2Te_3-10 wt% NC below 445 K and $Bi_2Sb_{2.3}Te_{0.7}$ above this temperature.

The figure also shows the Seebeck coefficient, thermal conductivity, and electrical conductivity, which are key parameters in TEG design. The Seebeck coefficient influences voltage generation, electrical conductivity affects power output, and thermal conductivity governs heat retention. Optimizing their balance is critical for achieving high conversion efficiency and overall device performance.

Ceramic (Al_2O_3) is employed as a heat exchanger in TEGs because of its excellent thermal stability, electrical insulation, and resistance to oxidation, which ensure reliable thermal transfer without short-circuiting. Copper is used as the conductor and terminal point because of its superior electrical and thermal conductivities, which enable efficient current flow and minimal energy loss. Thus, both materials are ideal for high-performance TEG applications.

3. MATHEMATICAL MODEL OF SIMULATION AND PERFORMANCE EVALUATION

The TEG was simulated in COMSOL Multiphysics 5.5 using both heat transfer and AC/DC modules to accurately model the coupled thermal and electrical behavior. The heat transfer module captures the temperature gradients and heat flow, whereas the AC/DC module simulates the electrical potential and current distribution. Their integration is essential for evaluating thermoelectric performance under realistic operating conditions.

3.1. Necessary Equations of Simulation. Heat flows from the heat source and is absorbed by the thermoelectric materials, governed by the energy conservation equation [19]:

$$\rho C_p \cdot u \nabla T + \nabla q = Q + Q_{ted} \quad (1)$$

where ρ denotes material density (kg/m^3), C_p denotes heat capacity ($J/kg \cdot K$), u indicates velocity vector (m/s), ∇T represents temperature gradient, Q is volumetric heat generation from internal sources (W/m^3), Q_{ted} denotes thermoelectric heat source or sink associated with the Thomson and Peltier effects (W/m^3), q indicates divergence of heat flux (W/m^3) and heat flux vector (W/m^2), q is $-k \nabla T$, and k denotes thermal conductivity of the material ($W/m \cdot K$).

This equation describes the conductive heat transfer, including thermoelectric heat sources such as the Joule and Peltier effects.

The electric current interface models the current conservation using Ohm's law, where the magnitude of electric potential is the main variable. This enables the calculation of the electric potential, current density, and electric field of the conductive materials [19].

$$\nabla J = Q_j \quad (2)$$

where J denotes induced current density and Q_j is the current source.

$$J = \sigma E + J_e \quad (3)$$

where σ is electrical conductivity, E is the electric field, and J_e represents external current density.

$$E = -\nabla V \tag{4}$$

where V denotes electric potential.

Eqs. 5, 6, and 7 merge the electric currents and heat transfer in Solids modules [19] to capture the interrelated Peltier, Seebeck, and Thomson effects in thermoelectric devices, as described by the following equations:

$$Q = P \cdot J \tag{5}$$

where Q is volumetric heat generation from internal sources (W/m^3), P is the Peltier coefficient, and J is the induced current density.

$$P = S \cdot T \tag{6}$$

where S is the Seebeck coefficient (V/K) and T is temperature (K).

$$J_e = -\sigma S \nabla T \tag{7}$$

where J_e is external current density, σ is electrical conductivity, and ∇T is temperature difference.

These equations capture the interaction between heat and charge transport in thermoelectric materials, supporting accurate simulations of performance under nonuniform temperature distributions.

3.2. Performance Evaluation. A TEG has an inherent internal resistance that stems from the properties of its thermoelectric legs. The internal resistance is determined using the following equation:

$$R_{int} = \sum_{i=1}^{n} \left(\frac{L}{\sigma A}\right)_i \tag{8}$$

This internal resistance leads to a voltage decrease within the device, thereby reducing the terminal voltage available at the load. The output voltage of the TEG at the load end can be expressed as

$$V = \frac{R_L}{R_L + R_{int}} \sum_{i=1}^{N} \int_{Tc}^{Th} \alpha \, dT \tag{9}$$

When an external load resistance is connected in between terminal points, the amount of current according to Ohm's law for electrical circuits is

$$I = \frac{V}{R_L + R_{int}} \tag{10}$$

This current flow generates power at the load terminals, representing the usable output power. This power can be determined using the following equation:

$$P = I^2 R_L \tag{11}$$

Electrical power is the converted form of heat energy absorbed from a heat source. The amount of absorbed heat energy is calculated using the following equation:

$$Q_h = \alpha T_h I - 0.5(I^2 R) + K(T_h - T_c) \tag{12}$$

The efficiency of this energy conversion process is expressed as a ratio of electrical power and absorbed heat.

$$\eta = \frac{P}{Q_h} \tag{13}$$

The material interface effects, contact resistances, and radiative boundary losses are generally negligible under typical operating conditions because conduction and thermoelectric effects dominate the heat and charge transport in most thermoelectric devices [20,21].

The preceding analysis clarifies the concepts involved in the energy conversion process within a TEG, which transforms heat into electricity and ultimately enables the determination of the conversion efficiency.

4. RESULTS AND DISCUSSION

The TEG performance was evaluated using the following key parameters: open-circuit voltage, load terminal voltage, load current, output power, and energy conversion efficiency. The open-circuit voltage reflects the Seebeck effect, whereas the load voltage and current indicate the response of the system to external loads. The output power indicates the energy delivery to the load, and the conversion efficiency quantifies the thermal-to-electrical energy conversion.

4.1. Thermal Distribution and Voltage Generation. In TEG, semiconductor legs contain excess charge carriers. When a temperature gradient is applied, they absorb thermal energy from the high-temperature side and diffuse toward the cold side, converting heat into kinetic energy. In n-type materials, electrons diffuse from high- to low-temperature regions and accumulate a negative charge on the lower-temperature side. Conversely, in p-type materials, holes migrate in the same direction, leading to the build-up of positive charges on the same side. This charge separation causes a potential difference across the terminals of the TEG. The COMSOL simulation reveals that a temperature gradient of $\Delta T = 200$ K ($T_h = 473$ K and $T_c = 353$ K) generates a voltage difference of 46.2 mV, depicted in Fig. 4(b), in the TEG.

The generated voltage increases linearly with the temperature gradient, and increasing the gradient increases the voltage. Additionally, the high-temperature side, which is affected by the sun's position, seasonal changes, and weather conditions, significantly influences the voltage output. For the same material, at the high temperature of 463 K, the voltage is 42.3 mV, whereas at a 483 K temperature, the voltage increases to 50 mV. These findings describe the critical role of thermal conditions in TEG performance.

4.2. Terminal Voltage and Load Current. When the two terminals of a TEG are connected by an external load, current flows owing to the generated potential difference. The magnitude of this current depends on the load resistance, assuming that the ohmic losses are negligible. The TEG has an internal electrical resistance that causes a voltage drop, thereby limiting the voltage available at the load terminals. For an optimal power transfer from TEG, the external load should match the internal resistance. The internal resistance is calculated using Eq. 8, the terminal

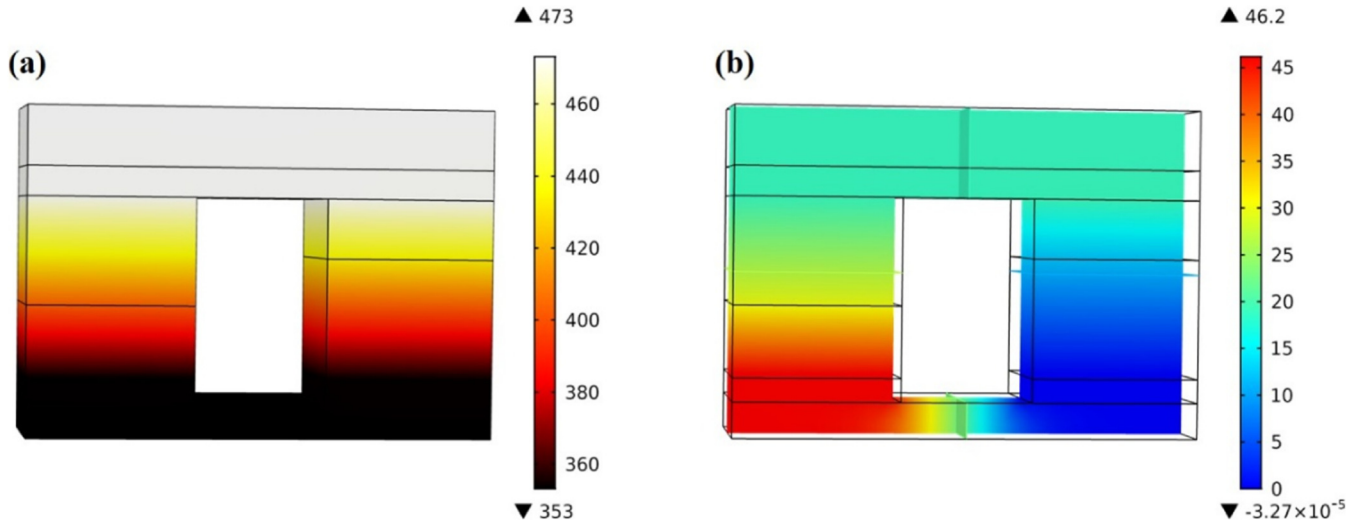


Figure 4. (a) Thermal distribution and (b) voltage distribution of the designed TEG

voltage across the load is determined using Eq. 9, and the current flowing through the load is expressed by Eq. 10.

Figure 5(a) shows the variation in the terminal voltage with the load resistance. Higher temperatures generate higher open-circuit voltages, resulting in an increased voltage across the load. Consequently, this leads to a higher load current, as demonstrated in Figure 5(b) and 5(c). These results show a direct relationship among the temperature gradient, open-circuit voltage, and current of the TEG.

A high terminal voltage and current boost the power output and efficiency in TEGs but can increase heat losses and material wear. Low voltage and current improve stability and longevity by reducing thermal stresses but result in lower power output. Balancing these factors is crucial for achieving optimal performance and durability in thermoelectric applications.

4.3. Power. Power output is a key performance metric in TEG applications because it reflects the ability of the system to convert heat energy into usable power. It is governed by the product of current and voltage across the terminal resistance. The highest amount of power transfer occurs when the external load matches the internal resistance of the TEG, ensuring optimal energy conversion. Using Eq. 8, the internal resistance of the designed device was 0.0188 Ω. The output power was subsequently evaluated using Eq. 11. Fig. 6 shows the power development of the designed TEG.

The graphs show variation in power output with load (a), terminal voltage (b), and load current (c) at three different hot-side temperatures (T_h = 463, 473, and 483 K). Higher hot-side temperatures (483 K) lead to higher power output across all load parameters. The maximum power output for a heat source temperature of 483 K is 33.17 mW, while for 473 K, it is 28.32 mW, and for 463 K, 23.74 mW. Throughout these measurements, the heat sink temperature is consistently maintained at 353 K. All power output curves exhibit an initial increase with increasing load resistance, reaching a peak when the value of load resistance satisfies the maximum power transfer theorem, which states that the power delivered is maximized when the external load matches the internal resistance of the source. Beyond this optimal point, as the load resistance continues to increase, the current through the circuit decreases significantly, reducing the power output. This behavior highlights the importance of load matching in the TEG design to ensure efficient energy conversion. The corresponding voltage and current values at the point of maximum power are illustrated in Fig. 6(b) and Fig. 6(c), respectively. At a hot-side temperature of 463 K, the highest power is obtained at a terminal voltage of 21.12 mV and a load current of 1.12 A. For hot-side temperatures of 473 and 483 K, the terminal voltages increase to 23.10 and 24.95 mV, with corresponding load currents of 1.22 and 1.33 A, respectively. The optimal voltage, current, and power outputs are selected based on the specific load demand.

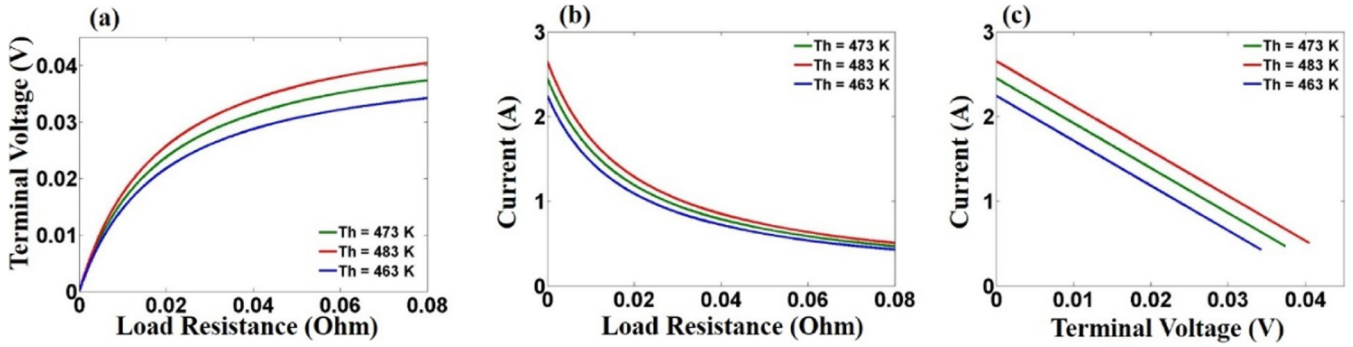


Figure 5. Graphs of (a) load terminal voltage vs. load resistance, (b) load current vs. resistance (c) load Current vs. terminal voltage

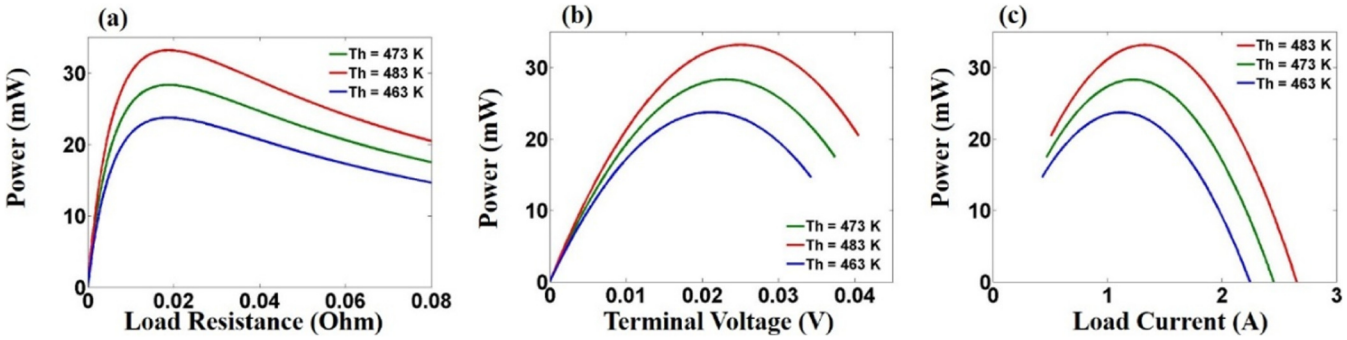


Figure 6. Graphs of (a) power vs. load resistance, (b) power vs. terminal voltage, and (c) power vs. load current

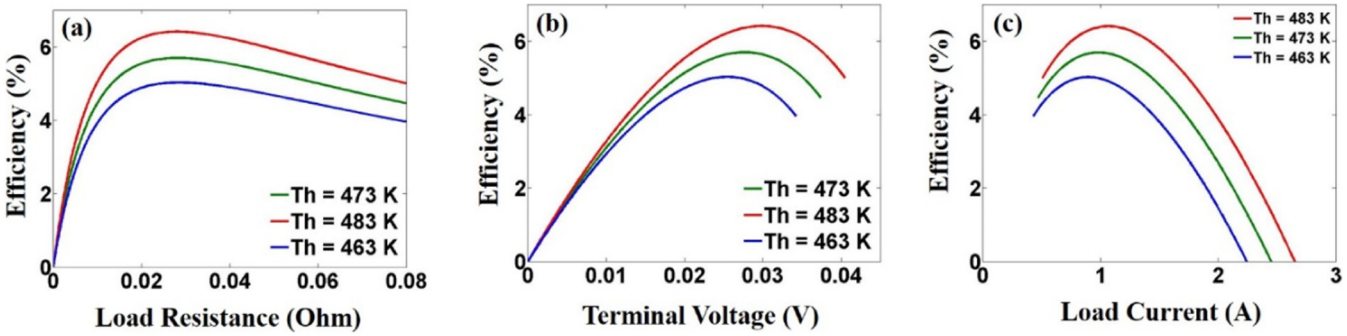


Figure 7. Graphs of (a) efficiency vs. load resistance, (b) efficiency vs. terminal voltage, and (c) efficiency vs. load current

This trend indicates that both the voltage and current at the maximum efficiency increase with increasing temperature gradients, thereby enhancing the overall performance of the TEG. An increment in hot-side temperature leads to a larger temperature gradient across the TEG, which increases the Seebeck voltage and drives more charge-carrier diffusion, resulting in greater electrical power.

4.4. Efficiency. Efficiency is crucial for TEG performance because it measures the effectiveness of heat conversion to electricity. High efficiency ensures practical viability, optimal material use, and an improved design. Figure 7 shows the efficiency trends of the designed TEG units under varying conditions. It shows how efficiency changes with resistance [Figure 7(a)], terminal voltage [Figure 7(b)], and load current [Figure 7(c)], each tested at high-temperature sides (T_h) of 463, 473, and 483 K. The outcomes show that the maximum efficiency is 6.4% (for T_h = 483 K), 5.69% (for T_h = 473 K), and 5.02% (for T_h = 463 K). Maximum efficiency is achieved at the same point where maximum power is obtained—that is, when the external load resistance matches the internal resistance of the TEG, which is 0.0188 Ω. The corresponding voltage and current values at the point of maximum efficiency are shown in Fig. 7(b) and Fig. 7(c). At a hot-side temperature of 463 K, the maximum efficiency is achieved at a terminal voltage of 21.12 mV and a load current of 1.12 A. As the hot-side temperature increases to 473 and 483 K, the terminal voltages increase to 23.10 and 24.95 mV, with corresponding load currents of 1.22 and 1.33 A, respectively. In each case, the heat sink temperature was 353 K.

This trend indicates that both the voltage and current at maximum efficiency increase with higher temperature gradients owing to the corresponding increase in output power. These results highlight the significant influence of the hot-side temperature on the TEG efficiency, emphasizing the critical role of operating conditions in maximizing the energy conversion performance.

Telkes employed a 50× optical concentration using a lens and achieved a 247 °C temperature gradient across thermoelectric elements composed of p- ZnSb alloy and n- Bi-based alloy, reporting a conversion efficiency of 3.35%[22]. D. Kraemer et al. reported that an efficiency of 8.6% is achievable by imposing a temperature gradient of 200 °C across an ideal thermoelectric device with $(ZT)_M = 1$ and T_c = 20 °C. However, they developed a solar TEG that achieved a highest efficiency of 4.6%[3]. Candadai et al. achieved 1.2% efficiency at a hot-side temperature of 280 °C for a STEG under 62-sun optical concentration, while simulations predicted 4.6% theoretical efficiency at a temperature gradient of 270 °C[23]. Pereira et al. investigated a solar TEG with $Si_{80}Ge_{20}$ and a $TiAlN/SiO_2$ PV absorber, achieving a maximum efficiency of 1.8%[24]. The analysis was limited to the thermal model of the solar TEG, excluding thermoelectric effects, owing to the significantly low energy conversion efficiency (approximately 3%) of Bi_2Te_3-based commercial thermoelectric modules[25,26]. However, our model obtains a peak efficiency of 5.69% at a 120 °C temperature difference (T_h = 473 K, T_c = 353 K), demonstrating improved performance over conventional modules.

5. CONCLUSIONS

A new model of a solar TEG unit was designed and simulated to enhance the efficiency via segmentation. The proposed approach optimizes the temperature gradient across a thermoelectric converter to achieve higher output energy parameters. Each leg of the

CSTEG was segmented into two parts, and materials with high ZT values were used to enhance the efficiency. The melt-spun with excess Te (Te-MS) compound ($Bi_{0.5}Sb_{1.5}Te_3$) and zone-melted $Bi_{0.4}Sb_{1.6}Te_3$ after hot deformation (HD-A-Sb1.6) were used as the p-type materials. Bi_2Te_3 -10 wt% nanocomposites and polycrystalline $Bi_2Te_{2.3}Se_{0.7}$ alloy were used as the n-type materials.

The newly designed COMSOL-simulated concentrated solar TEG model demonstrated a promising electrical performance at low temperature gradients. The model developed an open circuit voltage of 46.2 mV at 120 K ($T_h = 573$ K, $T_c = 453$ K), achieving a maximum power output of 28.32 mW and an efficiency of 5.69%. The effects of increasing and decreasing temperature gradients were also investigated, and a 10 K increment was observed to increase the voltage to 49.9 mV and power to 33.17 mW. Conversely, a 10 K decrease significantly decreased voltage to 42.25 mV and power to 23.74 mW, resulting in an efficiency of 5.02%.

The simulation results demonstrate that thermoelectric energy conversion performance can be significantly enhanced through strategic segmentation and appropriate material selection. Thus, segmented CSTEGs are promising candidates for efficient solar thermal energy harvesting, particularly for low-temperature-range applications.

This study is limited to idealized simulations, excluding real-world factors like contact resistance, material degradation, and environmental variations. Future work should include experimental validation, practical integration, and advanced segmentation/material strategies to improve efficiency.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Md. Habibur Rahman Aslam – *Electrical & Electronic Engineering, Chittagong University of Engineering and Technology, Bangladesh;* 0009-0008-0751-9645
Email: 06habib05@gmail.com

Author

Foyzul Karim – *Electrical & Electronic Engineering, Chittagong University of Engineering and Technology, Bangladesh;* 0009-0007-8318-0519
Email: u1902169@student.cuet.ac.bd

Corresponding Author

Anisul Islam Suva – *Research Mentor, Institute of Energy Technology, Chittagong University of Engineering and Technology, Bangladesh;* 0009-0000-7967-7548
Email: anisulislam.me@cuet.ac.bd

ACKNOWLEDGEMENTS

The authors acknowledge the support and resources provided by the COMSOL Community. The extensive user forums, tutorials, and shared simulation insights greatly facilitated the modeling and validation process in this study. The collaborative knowledge base offered by COMSOL users worldwide significantly contributed to overcoming technical challenges and improving simulation accuracy.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

FINANCIAL DISCLOSURE

The authors received no specific funding for this work.

REFERENCES

(1) Adrian, Maisarah, et al. "Energy transition towards renewable energy in Indonesia." *Heritage and Sustainable Development* 5.1 (2023): 107–118.

(2) Desideri, Umberto, and Pietro Elia Campana. "Analysis and comparison between a concentrating solar and a photovoltaic power plant." *Applied energy* 113 (2014): 422–433.

(3) Kraemer, Daniel, et al. "High-performance flat-panel solar thermoelectric generators with high thermal concentration." *Nature materials* 10.7 (2011): 532–538.

(4) Li, Long, et al. "Combined solar concentration and carbon nanotube absorber for high performance solar thermoelectric generators." *Energy Conversion and Management* 183 (2019): 109–115. Fan, Shufen, et al. "Influence of nanoinclusions on thermoelectric properties of n-type Bi_2 Te_3 nanocomposites." *Journal of Electronic materials* 40 (2011): 1018–1023.

(5) Kraemer, Daniel, et al. "Concentrating solar thermoelectric generators with a peak efficiency of 7.4%." *Nature Energy* 1.11 (2016): 1–8.

(6) Al-Nimr, M. A., B. M. Tashtoush, M. A. Khasawneh, and I. Al-Keyyam. "A Hybrid Concentrated Solar Thermal Collector/Thermo-Electric Generation System." *Energy*, vol. 134, 2017, pp. 1001–1012. *Elsevier*, https://doi.org/10.1016/J.ENERGY.2017.06.093.

(7) Baranowski, L. L., G. J. Snyder, and E. S. Toberer. "Concentrated Solar Thermoelectric Generators." *Energy & Environmental Science*, vol. 5, no. 9, 2012, pp. 9055–9067. *Royal Society of Chemistry*, https://doi.org/10.1039/C2EE22248E

(8) Kraemer, D., B. Poudel, H. P. Feng, J. C. Caylor, B. Yu, X. Yan, Y. Ma, et al. "High-Performance Flat-Panel Solar Thermoelectric Generators with High Thermal Concentration." *Nature Materials*, vol. 10, 2011, pp. 532–538. *Nature Publishing Group*, https://doi.org/10.1038/nmat3013.

(9) Piarah, W. H., Z. Djafar, A. S. Rosali, A. Halim, and B. H. J. Mustofa. "The Effect of Copper Coating on the Hot-Side on the Performance of a Thermoelectric Generator Using the Electroforming Method." *International Journal of Design & Nature and Ecodynamics*, vol. 17, no. 6, 2022, pp. 823–830. https://doi.org/10.18280/IJDNE.170602.

(10) Kraemer, D., Q. Jie, K. McEnaney, F. Cao, W. Liu, L. A. Weinstein, J. Loomis, Z. Ren, and G. Chen. "Concentrating Solar Thermoelectric Generators with a Peak Efficiency of 7.4%." *Nature Energy*, vol. 1, 2016, pp. 1–8. *Nature Publishing Group*, https://doi.org/10.1038/nenergy.2016.153.

(11) Chen, Wei-Hsin, and Yi-Bin Chiou. "Geometry design for maximizing output power of segmented skutterudite thermoelectric generator by evolutionary computation." *Applied Energy* 274 (2020): 115296.

(12) Snyder, G. J., and T. S. Ursell. "Thermoelectric Efficiency and Compatibility." *Physical Review Letters*, vol. 91, no. 14, 2003, p. 148301. https://doi.org/10.1103/PhysRevLett.91.148301.

(13) Yang, Kai-Yu, et al. "General Screening Rules and Segmented Optimization Strategy for Efficient Thermoelectric Devices Validated by $Mg_3(Sb,Bi)_2$ and $Bi_{0.5}Sb_{1.5}Te_3$-GeTe Module." *Advanced Science*, first published 5 May 2025, https://doi.org/10.1002/advs.202502832.

(14) Chen, Wei-Hsin, et al. "Optimization of a segmented thermoelectric generator with various doping amounts using central composite design, multi-objective genetic algorithm, and artificial neural network." *Energy* 316 (2025): 134469.

(15) Xu, Z. J., et al. "Enhanced thermoelectric and mechanical properties of zone melted p-type (Bi, Sb) 2Te3 thermoelectric materials by hot deformation." *Acta Materialia* 84 (2015): 385–392.

(16) Kim, Sang Il, et al. "Dense dislocation arrays embedded in grain boundaries for high-performance bulk thermoelectrics." *Science* 348.6230 (2015): 109–114.

(17) Hu, Lipeng, et al. "Point defect engineering of high-performance bismuth-telluride-based thermoelectric materials." *Advanced Functional Materials* 24.33 (2014): 5211–5218.

(18) Fan, Shufen, et al. "Influence of nanoinclusions on thermoelectric properties of n-type Bi2Te3 nanocomposites." *Journal of Electronic materials* 40 (2011): 1018–1023.

(19) Prasad, Asutosh, and Raj CN Thiagarajan. "Multiphysics modeling and multilevel optimization of thermoelectric generator for waste heat recovery." *Proceedings of the COMSOL Conference*. 2018.

(20) Zebarjadi, M., et al. "Perspectives on Thermoelectrics: From Fundamentals to Device Applications." *Energy & Environmental Science*, vol. 5, no. 1, 2012, pp. 5147–5162. https://doi.org/10.1039/C1EE02497C.

(21) Baranowski, L. L., G. J. Snyder, and E. S. Toberer. "Concentrated Solar Thermoelectric Generators." *Energy & Environmental Science*, vol. 5, no. 7, 2012, pp. 9055–9067. https://doi.org/10.1039/C2EE22248E.

(22) Telkes, M. "Solar Thermoelectric Generators." *Journal of Applied Physics*, vol. 25, 1954, pp. 765–777.

(23) Candadai, A. A., V. P. Kumar, and H. C. Barshilia. "Performance Evaluation of a Natural Convective-Cooled Concentration Solar Thermoelectric Generator Coupled with a Spectrally Selective High Temperature Absorber Coating." *Solar Energy Materials and Solar Cells*, vol. 145, 2016, pp. 333–341. https://doi.org/10.1016/J.SOLMAT.2015.10.040.

(24) Pereira, A., et al. "High Temperature Solar Thermoelectric Generator–Indoor Characterization Method and Modeling." *Energy*, vol. 84, 2015, pp. 485–492. https://doi.org/10.1016/J.ENERGY.2015.03.053.

(25) Jaziri, N., A. Boughamoura, J. Müller, B. Mezghani, F. Tounsi, and M. Ismail. "A Comprehensive Review of Thermoelectric Generators: Technologies and Common Applications." *Energy Reports*, vol. 6, 2020, pp. 264–287. https://doi.org/10.1016/j.egyr.2019.12.011.

(26) Maksymuk, M., B. Dzundza, O. Matkivsky, I. Horichok, R. Shneck, and Z. Dashevsky. "Development of the High Performance Thermoelectric Unicouple Based on Bi_2Te_3 Compounds." *Journal of Power Sources*, vol. 530, 2022, Article 231301. https://doi.org/10.1016/j.jpowsour.2022.231301.

KFUPM
Journal of Undergraduate Research International

Diamond Open Access 102

Article Environmental Science and Engineering/Earth Sciences

Detection of Urban Changes in Mumbai, Jakarta, Hong Kong, Dhaka, and Beijing

Latifa Alhabeeb[1] and Muhammad Bilal [1,2]*

Cite https://doi.org/10.64589/juri/209736

Submitted: June 04, 2025 **Revised:** July 28, 2025 **Accepted:** August 20, 2025

ABSTRACT

This study presents a multi-city assessment of urban expansion across five Asian megacities—Mumbai, Jakarta, Hong Kong, Dhaka, and Beijing—between 2014 and 2024. Using Landsat 8 and 9 surface reflectance imagery, four spectral indices—Normalized Difference Vegetation Index (NDVI), Normalized Difference Built-up Index (NDBI), Built-up Index (BUI), and Index-Based Built-up Index (IBI)—were computed within Google Earth Engine (GEE). A harmonized classification framework was applied to detect annual transitions between built-up and non-built-up areas. Urban dynamics were quantified through six key metrics: built-up area, non-built-up area, annual gain, annual loss, total change, and net transition. Results reveal heterogeneous urbanization trajectories. Jakarta exhibited the most substantial net increase (~22%), reflecting rapid sprawl, while Hong Kong showed compact but significant growth (~14%), consistent with vertical densification. Dhaka experienced rapid but spatially inconsistent expansion, with peripheral encroachment and inner-city restructuring. Beijing displayed stable, regulated growth shaped by planning policies, whereas Mumbai exhibited a unique decline of ~9% in built-up areas, suggesting a shift toward vegetated land cover. These findings highlight how harmonized remote sensing indices can capture nuanced urban patterns across diverse ecological and governance contexts. The study contributes practical insights for sustainable urban planning and provides a reproducible, low-cost framework for long-term urban monitoring in data-scarce regions..

Keywords: urban growth, remote sensing, NDVI, NDBI, built-up index, megacities

1. INTRODUCTION

Urban expansion is one of the defining processes of the 21st century. Globally, more than half the population resides in urban areas, and by 2050, this figure is projected to approach 70% [1,2]. Asian megacities, particularly those situated in coastal and floodplain areas, are undergoing an unprecedented transformation due to rapid population growth, economic dynamism, and climate-related pressures, including flooding and sea-level rise [3–5]. These processes reshape spatial structures, intensify land-use conflicts, and exacerbate sustainability challenges.

Among Asian cities, Mumbai, Jakarta, Hong Kong, Dhaka, and Beijing exemplify contrasting yet comparable cases.

- *Mumbai* demonstrates unregulated sprawl driven by informal settlements and constrained land availability.
- *Jakarta* is a transitional city where informal expansion interacts with state-led relocation and reclamation projects, compounded by subsidence and flooding.
- *Hong Kong* reflects compact vertical growth shaped by terrain and policies prioritizing density and green preservation.
- *Dhaka* epitomizes hyper-urbanization in a deltaic environment, where informal settlements expand into wetlands and floodplains.
- *Beijing* represents regulated peri-urban expansion with strong planning interventions under the "Beautiful China" and Jing-Jin-Ji strategies.

Together, these cities illustrate diverse governance, ecological, and socioeconomic contexts, offering a comparative lens into Asia's urban transformation.

Satellite-based remote sensing has emerged as a reliable and cost-effective method for detecting changes in urban land-cover change. Spectral indices provide objective classification of vegetation and built-up areas across time and space. The four indices used in this study—NDVI, NDBI, BUI, and IBI—are widely validated for distinguishing urban from non-urban surfaces.

Most existing studies, however, focus on single cities **or** short-term analyses**.** Few have harmonized thresholds across multiple cities and applied consistent methodologies to allow meaningful cross-comparisons. This study addresses this gap by implementing a standardized multi-index, multi-city approach across five megacities for the decade 2014–2024. This study aims to: (i) Quantify annual changes in built-up and non-built-up land across five Asian megacities. (ii) Compare multi-index trajectories across diverse ecological and governance contexts. (iii) Identify gains, losses, and net transitions in the built-up area. (iv) Evaluate index performance in detecting urban changes under

varied urban landscapes. (v) Provide transferable insights for sustainable urban planning and monitoring.

2. METHODOLOGY

2.1. Study Areas. Figure 1 shows the AOIs for Mumbai, Jakarta, Hong Kong, Dhaka, and Beijing. These five cities were chosen for their contrasting urbanization pressures: coastal reclamation (Mumbai, Jakarta, Hong Kong), floodplain expansion (Dhaka), and planned inland development (Beijing).

2.2. Data Sources. Annual Landsat 8 OLI/TIRS and Landsat 9 OLI/TIRS surface reflectance imagery (30 m resolution) were used for the period 2014–2024[6–8]. Landsat data were accessed through the Google Earth Engine (GEE) platform, ensuring consistency and free availability.

- Temporal selection: January–March median composites were generated for each year to minimize phenological variations, reduce seasonal bias, and avoid monsoon effects (significant in tropical cities).
- Spatial selection: Administrative boundaries were digitized for each city to define Areas of Interest (AOIs).
- Projection: All datasets were reprojected to WGS84 for global consistency.

This section outlines the satellite datasets, index-based classification techniques, and annual urban change-detection methods used in this study, which employed a harmonized remote sensing framework to evaluate urban expansion in five major Asian cities (Mumbai, Jakarta, Hong Kong, Dhaka, and Beijing) between 2014 and 2024.

2.3. Spectral Indices. Four indices were computed using standard band equations:

1. ***Normalized Difference Vegetation Index (NDVI):***

$$NDVI = \frac{(NIR - RED)}{(NIR + RED)} \tag{1}$$

Initially introduced by Rouse et al. (1974)[9], NDVI highlights vegetated vs. non-vegetated land.

2. ***Normalized Difference Built-up Index (NDBI):***

$$NDBI = \frac{(SWIR - NIR)}{(SWIR + NIR)} \tag{2}$$

Proposed by Zha et al. (2003)[10], NDBI enhances built-up detection.

3. ***Built-up Index (BUI):***

$$BUI = NDBI - NDVI \tag{3}$$

Combines vegetation and built-up indices for clearer separation[11].

4. ***Index-Based Built-up Index (IBI):***

$$IBI = \frac{NDBI - (NDVI - MNDWI)}{NDBI + (NDVI + MNDWI)} \tag{4}$$

Where:

$$MNDWI = \frac{(Green - SWIR)}{(Green + SWIR)} \tag{5}$$

Introduced by Xu (2008)[12], IBI incorporates water index information to reduce misclassification in mixed urban environments.

2.4. Thresholding and Classification. For each city, visually calibrated thresholds were applied consistently across all years to classify pixels into built-up and non-built-up. These thresholds (Table 1) accounted for environmental differences, such as stricter NDVI thresholds in vegetated cities (e.g., Hong Kong, Dhaka) compared to relaxed ones in arid Beijing.

3. RESULTS AND DISCUSSION

This section presents the spatial and temporal dynamics of urban changes in the five cities, analyzed using NDVI, NDBI, BUI, and IBI indices.

3.1. Urban Index Maps.

3.1.1. Mumbai. Figure 2 illustrates Mumbai's land-cover classifications using RGB, NDVI, NDBI, BUI, and IBI. Built-up surfaces expanded northward and eastward, with visible growth along reclaimed coastal zones and transport corridors. However, compared with other cities, Mumbai exhibited intermittent reversals where vegetated land appeared to recover.

- NDVI emphasized vegetation loss across peri-urban and coastal zones, with sharp declines between 2016 and 2020.
- NDBI effectively captured rooftops and dense built-up areas, especially in northern suburbs. Major expansion was observed in 2015, 2018, and 2022.
- BUI confirmed steady intensification across central and northeastern districts.
- IBI provided balanced detection, reducing misclassification in shadowed areas.

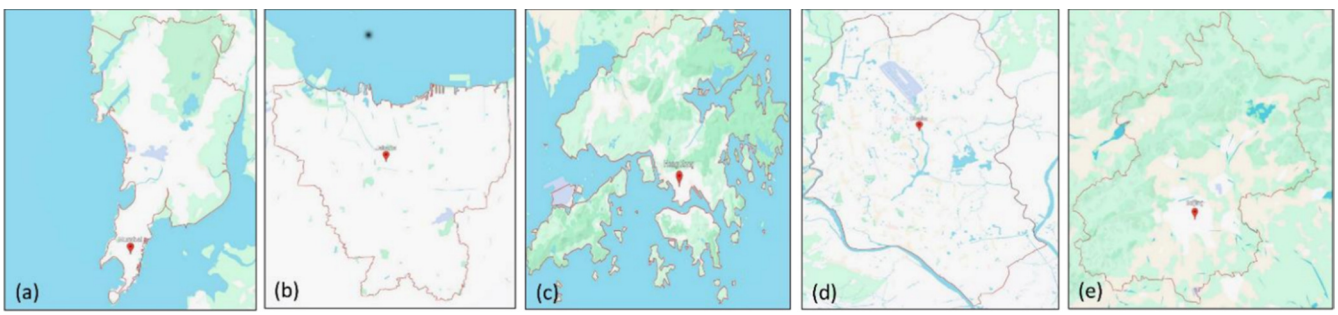


Figure 1. Locations and AOIs of the five study cities.

Table 1. *Threshold values for NDVI, NDBI, BUI, and IBI (2014–2024).*

Index	Mumbai (2014–2024)	Jakarta (2014–2024)	Hong Kong (2014–2024)	Dhaka (2014–2024)	Beijing (2014–2024)
NDVI	0.25 (2014), 0.35 (2024)	0.30 (2014), 0.30 (2024)	0.30 (2014), 0.30 (2024)	0.30 (2014), 0.30 (2024)	0.20 (2014), 0.20 (2024)
NDBI	0.00 (2014), 0.05 (2024)	0.05 (2014), 0.07 (2024)	0.00 (2014), 0.00 (2024)	0.00 (2014), 0.00 (2024)	0.00 (2014), 0.00 (2024)
BUI	−0.25 (2014), −0.20 (2024)	−0.22 (2014), −0.20 (2024)	−0.25 (2014), −0.20 (2024)	−0.20 (2014), −0.20 (2024)	−0.30 (2014), −0.30 (2024)
IBI	−0.60 (2014), −0.55 (2024)	−0.58 (2014), −0.55 (2024)	−0.50 (2014), −0.65 (2024)	−0.70 (2014), −0.70 (2024)	−0.50 (2014), −0.50 (2024)

Year-to-year differences reveal alternating gains and losses. Periods such as 2015–2016 and 2017–2018 show moderate built-up gains, while 2016–2017 and 2018–2019 exhibit losses or vegetation recovery. After 2020, Mumbai's transitions show a bias toward net loss, reinforcing the overall decline noted earlier. Despite episodic construction surges, Mumbai displayed an overall 9% decline in net built-up area by 2023. This reversal may be attributed to coastal protection programs, mangrove restoration efforts, and enhanced environmental enforcement.

3.1.2. Jakarta. As shown in Figure 3, Jakarta underwent extensive outward growth during the decade. Built-up expansion was especially prominent in eastern and southern districts, while vegetation fragmentation occurred in central and northern areas.

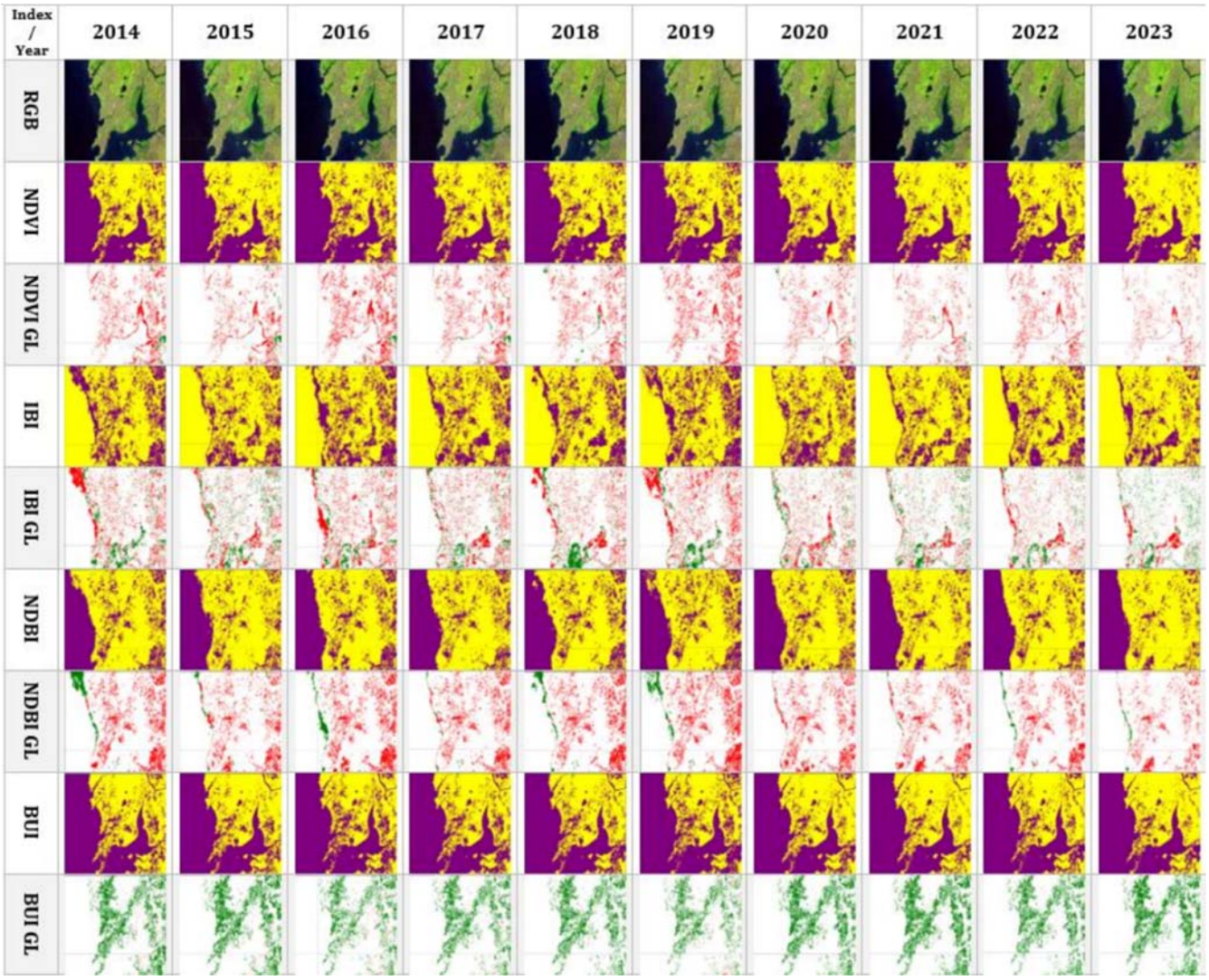


Figure 2. Annual classification and gain/loss maps for Mumbai (2014–2023). Built-up areas are shown in purple and non-built-up areas in yellow. Green indicates new urban development (gain), while red highlights areas of built-up land loss.

- NDVI revealed steady vegetation loss after 2017, particularly in peri-urban zones.
- NDBI captured linear growth along industrial corridors and highways.
- BUI highlighted concentric outward growth around the core.
- IBI depicted densification along the urban fringe.

The results show consistent positive transitions, particularly during the periods of 2015–2016, 2018–2019, and 2022–2023. Negative transitions were minimal and scattered, indicating a strong net expansion trajectory. Jakarta recorded the most significant net increase (~22%) in built-up area among the five cities. Land subsidence and flood-prone conditions may have influenced land-use restrictions in the north, redirecting expansion southward.

3.1.3. Hong Kong. Figure 4 shows Hong Kong's compact but intensive urban development. Expansion was concentrated in reclaimed zones (West Kowloon, Tseung Kwan O, and parts of Lantau).

- NDVI detected gradual vegetation fragmentation along urban fringes.
- NDBI revealed consistent intensification in Kowloon and Hong Kong Island.
- BUI confirmed clustering around transit hubs and public housing estates.
- IBI effectively detected vertical and infill development, reflecting Hong Kong's densification strategy.

Hong Kong shows smaller but steady positive bars in most years, reflecting controlled densification and reclamation. Occasional minor negative values (e.g., 2016–2017) may represent temporary redevelopment or vegetation misclassification. Hong Kong exhibited a 14% net gain in built-up land, primarily through vertical development and reclamation, consistent with transit-oriented planning.

3.1.4. Dhaka. Figure 5 depicts Dhaka's rapid but inconsistent expansion. While built-up areas increased across peri-urban zones, the core displayed cycles of redevelopment.

- NDVI highlighted consistent vegetation loss after 2016.
- NDBI revealed strong expansion along northern and eastern corridors.
- BUI indicated outward sprawl, consuming agricultural land.

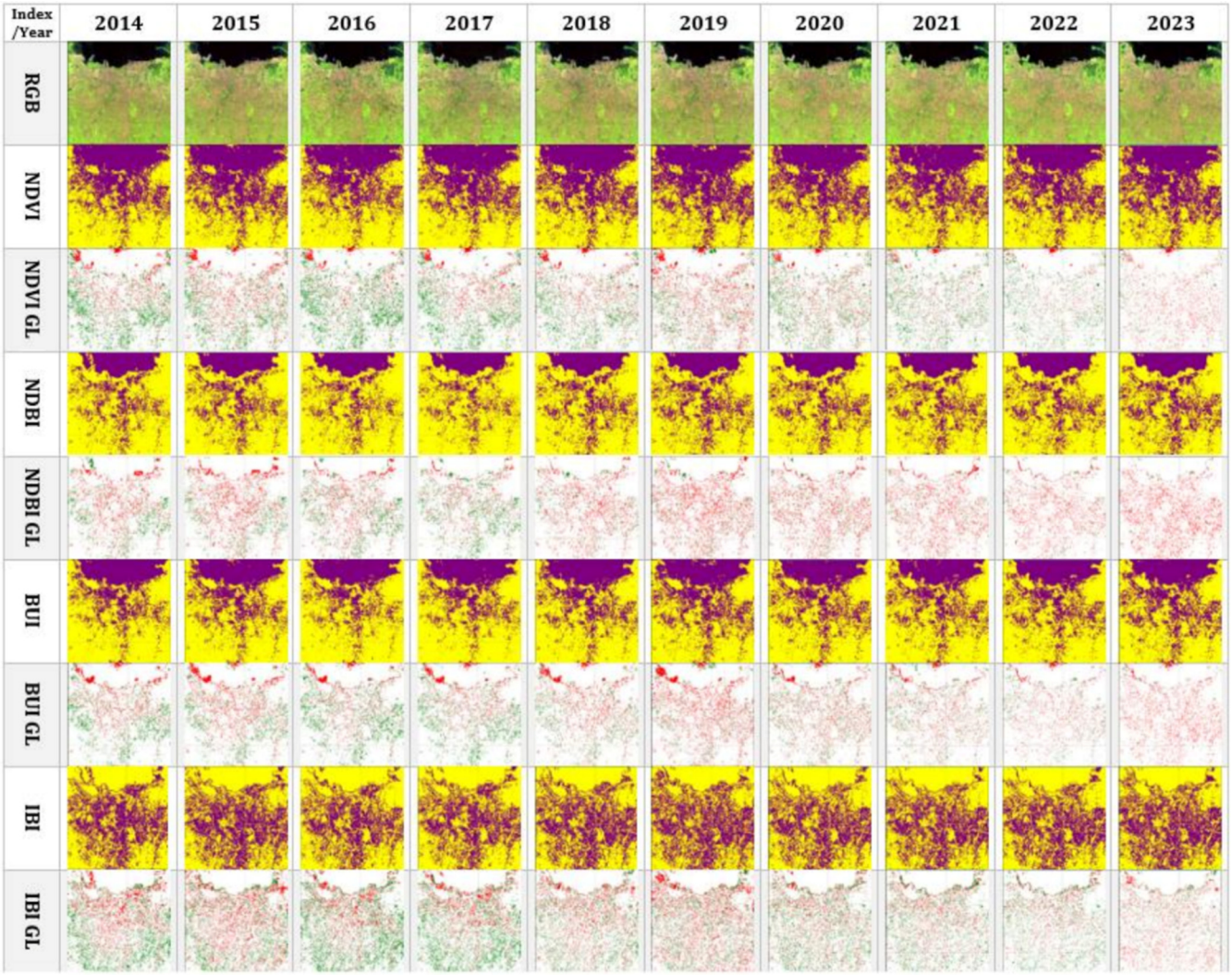


Figure 3. Annual classification and gain/loss maps for Jakarta (2014–2023). Built-up areas are shown in purple and non-built-up areas in yellow. Green indicates new urban development (gain), while red highlights areas of built-up land loss.

- IBI emphasized saturation within the urban core, particularly near transport corridors.

Dhaka's yearly transitions were volatile. Gains were evident in 2014–2017 and 2018–2020, but sharp negative bars occurred after 2020, suggesting redevelopment phases, seasonal floodplain effects, or index misclassification. Dhaka's expansion was rapid but unstable. Built-up growth was offset by localized redevelopment and misclassification during flooding seasons. Overall, the city demonstrated net positive growth, though less consistent than Jakarta or Hong Kong.

3.1.5. Beijing. Figure 6 highlights Beijing's structured outward growth. Expansion occurred primarily in suburban districts such as Tongzhou, Shunyi, and Fangshan.

- NDVI indicated gradual vegetation loss in peri-urban zones, offset by greening campaigns (e.g., afforestation).
- NDBI captured industrial and residential developments in the southern and eastern districts.
- BUI identified suburban growth along expressways and rail corridors.
- IBI emphasized densification of older urban zones (e.g., Haidian, Chaoyang).

Beijing's transitions were mostly positive until 2020, reflecting suburban expansion. Bars after 2021 show smaller or negative changes, aligning with policies promoting ecological restoration and stabilization of peri-urban growth. Beijing exhibited steady, regulated growth, consistent with its polycentric planning model and "Beautiful China" ecological policies. Unlike Jakarta or Dhaka, expansion was tightly controlled and integrated with transport planning.

3.2. Consolidated Trends, Comparative Analysis, and Limitations.

3.2.1. Mumbai. The consolidated charts for Mumbai (Figure 7) highlight a net decline in built-up area across NDVI, NDBI, BUI, and IBI.

- Built-up area: NDVI and BUI show a gradual reduction after 2018, while NDBI fluctuates, capturing temporary construction cycles.
- Non-built-up area: NDVI indicates steady vegetation recovery, particularly in coastal and wetland zones.

Figure 4. Annual classification and gain/loss maps for Hong Kong (2014–2023). Built-up areas are shown in purple and non-built-up areas in yellow. Green indicates new urban development (gain), while red highlights areas of built-up land loss.

- Gain/loss: Small gains in 2015–2016 were outweighed by consistent losses after 2020.

Mumbai is the only city in this study to show a long-term negative trajectory, possibly due to stricter coastal zone regulations, mangrove restoration, and reclassification of semi-urban land. This makes it a unique case among Asian megacities, where most are expanding.

3.2.2. Jakarta. The consolidated trends for Jakarta (Figure 13) confirm it as the fastest-growing city in the dataset.

- Built-up area: All four indices show strong upward growth, with NDBI and BUI capturing rapid sprawl along the eastern and southern peripheries.
- Non-built-up area: NDVI reveals progressive vegetation decline, particularly in agricultural zones converted to housing or industry.
- Gain/loss: Peaks in 2018–2019 and 2022–2023 indicate construction booms linked to large-scale housing projects.

Jakarta's gains far outweigh its losses. Despite flood risks and land subsidence, the city is expanding rapidly into peri-urban land, underscoring the challenges of balancing growth with environmental vulnerability.

Hong Kong

Hong Kong's consolidated results (Figure 9) reveal a steady upward trajectory in built-up area despite limited land availability.

- Built-up area: Gains are consistent across indices, strongest in IBI (vertical redevelopment and infill).
- Non-built-up area: Declined steadily, reflecting reclamation and loss of green belts in suburban districts.
- Gain/loss: Positive gains dominate most years; losses are minimal and largely transitional.

Hong Kong's growth model is compact and vertical, enabled by reclamation projects and redevelopment of high-density zones. The city shows one of the most stable positive trends.

3.2.3. Dhaka. Dhaka's consolidated charts (Figure 10) highlight its volatile and unstable urban growth pattern.

- Built-up area: Increased during 2014–2019, peaking around 2020, then declined slightly in subsequent years.

Figure 5. Annual classification and gain/loss maps for Dhaka (2014–2023). Built-up areas are shown in purple and non-built-up areas in yellow. Green indicates new urban development (gain), while red highlights areas of built-up land loss.

- Non-built-up area: NDVI shows substantial recovery after 2020, possibly due to seasonal flooding or re-greening.
- Gain/loss: Gains were large in 2016–2018, but sharp losses in 2021–2023 suggest redevelopment cycles or classification challenges in floodplains.

Dhaka exemplifies the instability of unregulated urbanization in flood-prone environments. Although net growth was positive, the magnitude of losses in later years undermines long-term sustainability.

3.2.4. Beijing. Beijing's consolidated results (Figure 11) confirm steady, regulated expansion.

- Built-up area: Grew consistently across all indices until 2020, then stabilized.
- Non-built-up area: Declined gradually, but NDVI shows partial recovery after 2021 due to afforestation.
- Gain/loss: Peaks in 2016–2018 reflect major suburban expansion; subsequent years show balanced gains and losses, aligning with policy-driven stabilization.

Beijing's growth is controlled and polycentric, guided by planning frameworks such as the Jing-Jin-Ji regional integration strategy. Unlike Jakarta or Dhaka, Beijing demonstrates long-term regulation of urban form.

3.3. Comparative Analysis. Comparing across all five cities revealed that:

- Rapid Expansion: Jakarta showed the strongest outward growth (~22% increase), driven by housing and industrial corridors.
- Compact Growth: Hong Kong displayed controlled densification (~14% increase), relying on reclamation and vertical redevelopment.
- Unstable Growth: Dhaka recorded high variability, with periods of strong gains followed by significant losses, reflecting its unregulated expansion in a flood-prone delta.
- Regulated Expansion: Beijing grew steadily, reflecting central planning, then stabilized under ecological and decentralization policies.

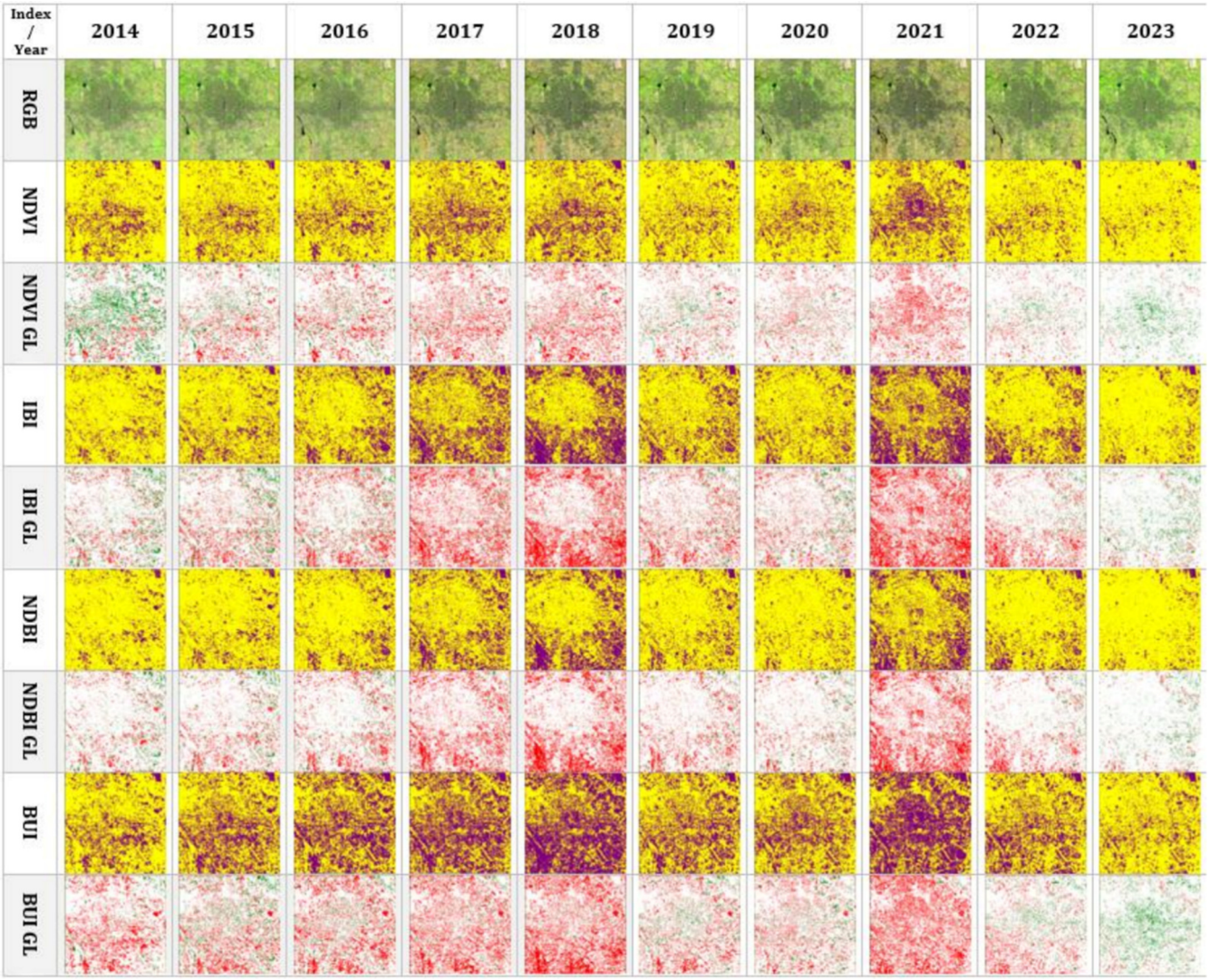


Figure 6. Annual classification and gain/loss maps for Beijing (2014–2023). Built-up areas are shown in purple and non-built-up areas in yellow. Green indicates new urban development (gain), while red highlights areas of built-up land loss.

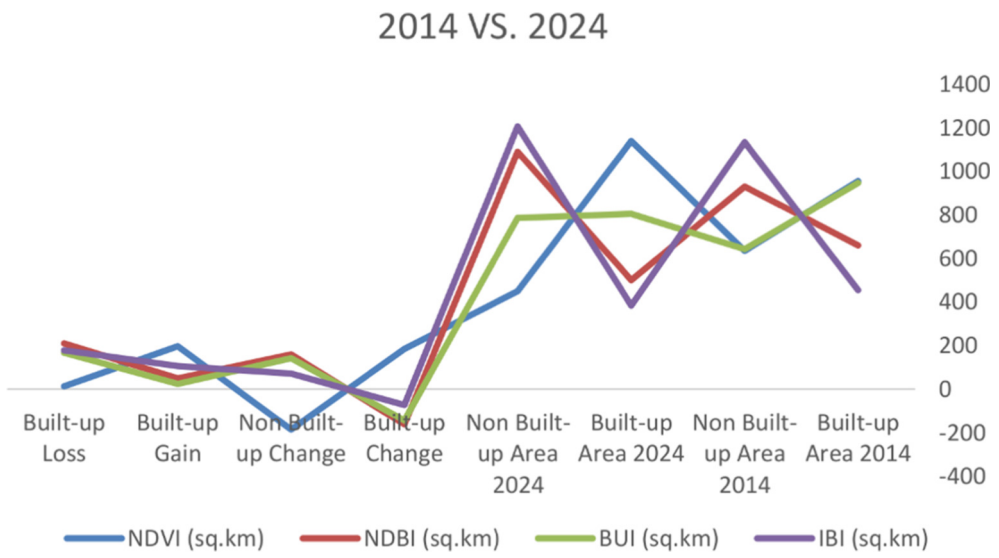


Figure 7. Consolidated charts of built-up and non-built-up trends, gains, and losses in Mumbai (2014–2023).

- Unique Decline: Mumbai stood out as the only city with a net decline (~9%) in built-up area, likely due to environmental restoration and stricter regulations.

These findings underscore the influence of policy frameworks, ecological contexts, and governance capacity on urban growth patterns in Asia.

3.4. Methodological Constraints and Limitations.

Despite rigorous processing, several limitations must be acknowledged:

1. Threshold Sensitivity: Threshold calibration was manual and visually guided. While consistent across years, subjective thresholds may misclassify mixed pixels, especially in transitional zones (e.g., urban-vegetation interfaces in Dhaka or Hong Kong).
2. Vegetation Interference: NDVI misclassified green roofs, wetlands, and peri-urban greenery as non-urban in highly vegetated cities (Hong Kong, Dhaka). Conversely, sparse vegetation in Mumbai occasionally confused NDVI into labeling built-up land as "green."
3. Temporal Variability: Using annual composites reduced seasonality but may not fully capture intra-annual fluctuations (e.g., flood-related land-use changes in Dhaka, construction pauses in Mumbai).
4. Spatial Resolution: Landsat's 30 m resolution smooths fine-scale urban features (narrow streets, small plots). While suitable for regional comparisons, it underrepresents micro-scale redevelopment.
5. Index Behavior: Each index had distinct strengths and weaknesses:
   - NDBI was effective for dense urban cores but underestimated in arid zones (Mumbai).
   - NDVI was unsuitable for heavily vegetated cities.
   - BUI provided balanced detection but smoothed variations.

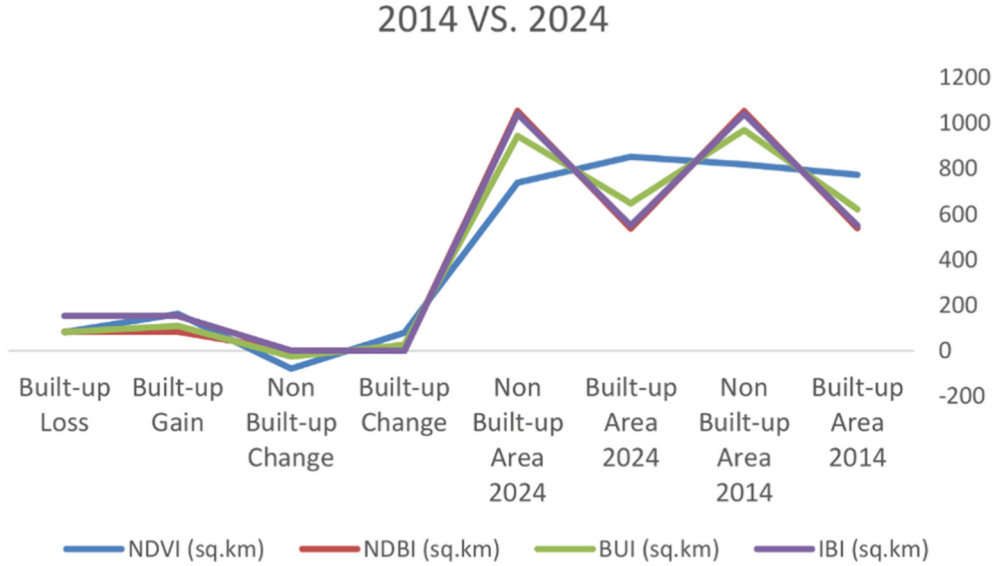


Figure 8. Consolidated charts of built-up and non-built-up trends, gains, and losses in Jakarta (2014–2023).

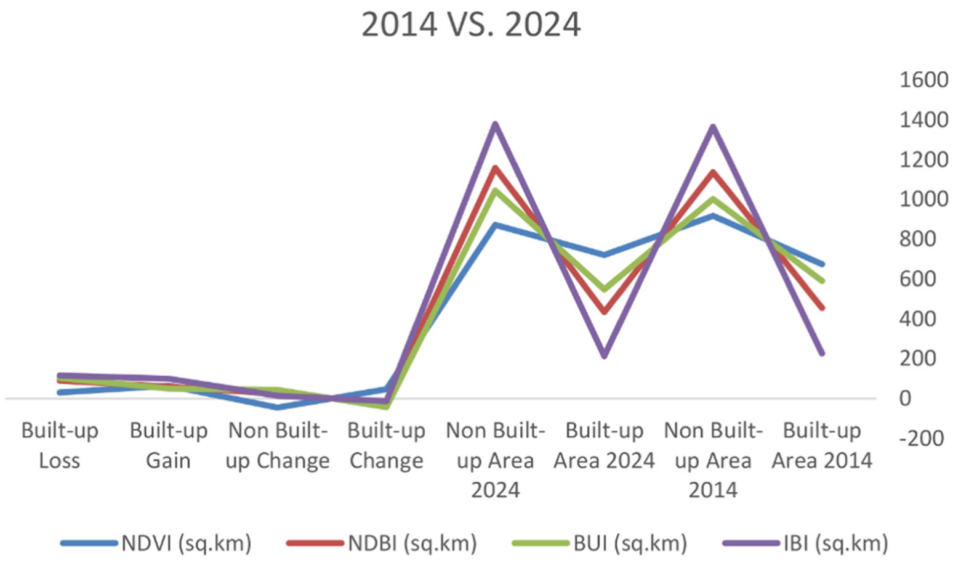


Figure 9. Consolidated charts of built-up and non-built-up trends, gains, and losses in Hong Kong (2014–2023).

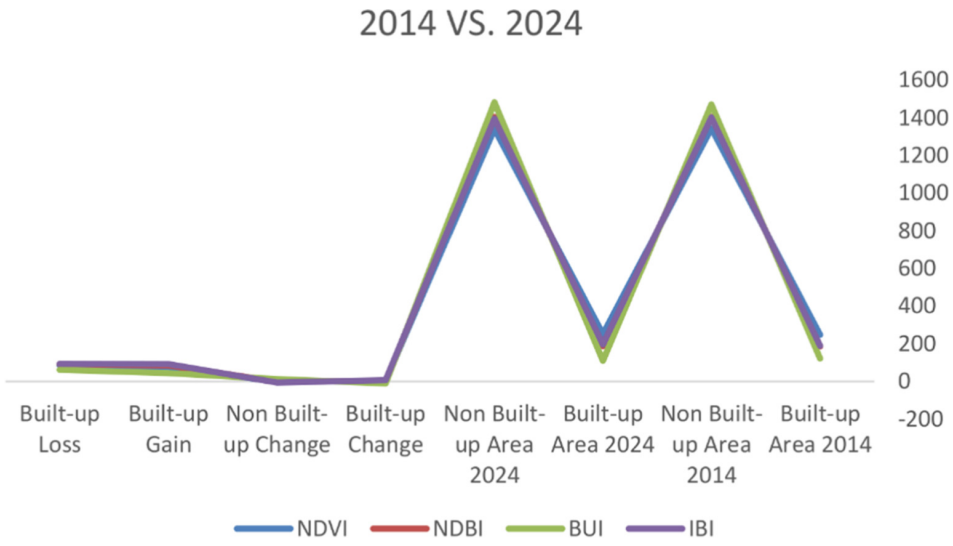


Figure 10. Consolidated charts of built-up and non-built-up trends, gains, and losses in Dhaka (2014–2023).

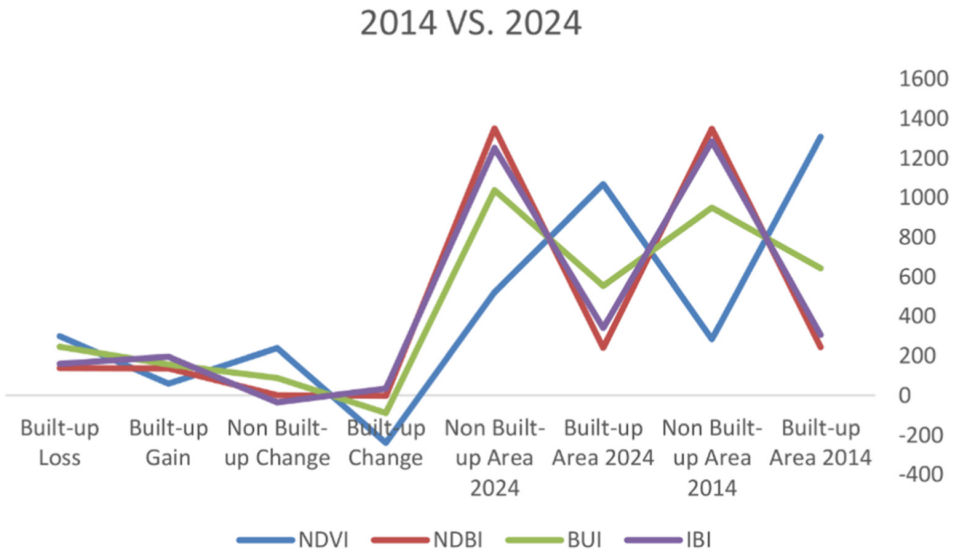


Figure 11. Consolidated charts of built-up and non-built-up trends, gains, and losses in Beijing (2014–2023)

- IBI captured compact/vertical growth but was overly sensitive in redevelopment phases (Seoul, Dhaka).

6. Validation Limitations: No ground-truth accuracy assessment was performed due to data availability. Visual verification against RGB composites ensured general validity but lacked quantitative accuracy metrics.

The harmonized multi-index framework enabled cross-city comparability, but improvements could include higher-resolution imagery, automated thresholding, or machine learning classifiers.

4. CONCLUSIONS

This study investigated urban growth dynamics in five Asian megacities—Mumbai, Jakarta, Hong Kong, Dhaka, and Beijing—over the period 2014–2024 using Landsat 8/9 imagery and four widely used remote sensing indices (NDVI, NDBI, BUI, IBI). By applying harmonized thresholds across all cities, we quantified annual changes in built-up and non-built-up areas, examined year-to-year transitions, and consolidated long-term growth trajectories. Key findings include: (i) Jakarta recorded the most rapid and sustained expansion (~22% net increase), reflecting unregulated sprawl and suburbanization, especially in eastern and southern corridors. (ii) Hong Kong showed compact, vertical growth (~14% increase), consistent with reclamation projects and transit-oriented redevelopment. (iii) Dhaka experienced volatile expansion, with strong early growth followed by significant losses after 2020, highlighting the instability of unregulated growth in flood-prone regions. (iv) Beijing displayed steady, regulated expansion until 2020, then stabilized due to ecological restoration and decentralization policies. (v) Mumbai was unique in showing a net decline (~9% reduction), likely influenced by coastal conservation policies, mangrove restoration, and land-use reclassification. The comparative results confirm that policy frameworks and governance strongly influence urban outcomes. Cities with strong planning regimes (Beijing, Hong Kong) experienced controlled growth, while those with weaker regulation (Jakarta, Dhaka) faced rapid and unstable sprawl. The ecological context also mattered: floodplains (Dhaka, Jakarta) exhibited volatile patterns, while arid Mumbai reflected unique reversals linked to conservation efforts. A multi-index approach was essential for reducing bias. NDVI underperformed in green cities, while NDBI performed better in arid ones, whereas BUI and IBI provided a more balanced classification.

The harmonized multi-index framework demonstrated that medium-resolution Landsat imagery, processed on GEE, can provide replicable and cost-effective monitoring of megacities across diverse contexts. This approach is particularly valuable for developing regions where ground data are scarce.

Overall, this study shows that Asian megacities exhibit diverse, context-driven urban trajectories and that remote sensing indices can capture both the pace and direction of these changes. These insights are crucial for urban planners and policymakers seeking to strike a balance between growth and environmental sustainability.

5. RECOMMENDATIONS

5.1. Recommendations for Future Work.

- Integrating higher-resolution imagery (e.g., Sentinel-2, PlanetScope) for improved detection of fine-scale changes.
- Automating thresholding using machine learning or time-series classification models.
- Incorporating socio-economic and policy datasets to strengthen interpretation of remote sensing results.
- Extending comparative frameworks to additional cities for a broader regional perspective.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Latifa Alhabeeb – *Architecture and City Design Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia;* 0009-0009-6967-6491
Email: s202243320@kfupm.edu.sa

Corresponding Author

Muhammad Bilal – *Research Mentor, Architecture and City Design Department, College of Design and Built Environment, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia; Center for Aviation and Space Exploration,;* 0000-0003-1022-3999
Email: muhammad.bilal@kfupm.edu.sa

ACKNOWLEDGEMENTS

This report was completed as part of an undergraduate program in the Department of Architecture and City Design at King Fahd University of Petroleum and Minerals (KFUPM). The author thanks the faculty advisors and the JURI editorial team for their valuable guidance and support throughout the project.

The author also acknowledges the use of AI-based tools to improve the clarity and structure of the manuscript's English. The analysis, results, and interpretations remain entirely the original work of the student researcher.

REFERENCES

(1) Elmqvist, T., Fragkias, M., Goodness, J., Güneralp, B., Marcotullio, P. J., McDonald, R. I., Seto, K. C. (2013). *Urbanization, biodiversity and ecosystem services: Challenges and opportunities.* Springer. (See chapter on urban expansion and land-use conflicts). https://doi.org/10.1007/978-94-007-7088-1

(2) Seto, K. C., Güneralp, B., & Hutyra, L. R. (2012). Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proceedings of the National Academy of Sciences, 109*(40), 16083–16088. https://doi.org/10.1073/pnas.1211658109

(3) Becker, M., Seeger, K., Paszkowski, A., Loebl, M., Pham, T. T. H., Weerts, A. H., Minderhoud, P. S. J. (2024). Coastal flooding in Asian megadeltas: Recent advances, persistent challenges, and call for actions amidst local and global changes. *Reviews of Geophysics*, https://doi.org/10.1029/2024RG000846

(4) Chan, F. K. S., Chuah, C. J., & Gao, X. (2014). Impacts of climate change: Challenges of flooding in coastal East Asia. In M. Beeson & S. Hameiri (Eds.), *Routledge Handbook of Asian Regionalism* (pp. 1–20). Routledge. https://eprints.whiterose.ac.uk/id/eprint/81151/

(5) Neumann, B., Vafeidis, A. T., Zimmermann, J., & Nicholls, R. J. (2015). Future coastal population growth and exposure to sea-level rise

and coastal flooding: A global assessment. *PLoS ONE, 10*(3), e0118571. https://doi.org/10.1371/journal.pone.0118571

(6) Bilal, M., Nazeer, M., Nichol, J.E., Bleiweiss, M.P., Qiu, Z., Jäkel, E., Campbell, J.R., Atique, L., Huang, X., Lolli, S., 2019. A Simplified and Robust Surface Reflectance Estimation Method (SREM) for Use over Diverse Land Surfaces Using Multi-Sensor Data. Remote Sensing 11.

(7) Nazeer, M., Ilori, C.O., Bilal, M., Nichol, J.E., Wu, W., Qiu, Z., Gayene, B.K., 2021. Evaluation of atmospheric correction methods for low to high resolutions satellite remote sensing data. Atmospheric Research 249.

(8) Vermote, E., Justice, C., Claverie, M., Franch, B., 2016. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. Remote Sens Environ Volume 185, 46–56.

(9) Huang, S., Tang, L., Hupy, J.P. et al. A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing. *J. For. Res.* **32**, 1–6 (2021). https://doi.org/10.1007/s11676-020-01155-1

(10) Zha, Y., Gao, J. & Ni, S. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. Int. J. Remote Sens. 24, 583–594 (2003).

(11) Kaimaris, D.; Patias, P. Identification and area measurement of the built-up area with the built-up index (bui). Int. J. Adv. Remote Sens. GIS 2016, 5, 1844–1858.

(12) Xu, H. A new index for delineating built-up land features in satellite imagery. Int. J. Remote Sens. 29, 4269–4276 (2008).

KFUPM
Journal of Undergraduate Research International

Diamond Open Access 113

Article Environmental Science and Engineering/Earth Sciences

Urban Change Detection and Growth Analysis (2014–2024): A Remote Sensing Study of Riyadh, London, and Seoul

Zahra Alhaddad[1] and Muhammad Bilal[1,2]*

Cite https://doi.org/10.64589/juri/209731

Submitted: June 04, 2025 **Revised:** August 16, 2025 **Accepted:** August 20, 2025

ABSTRACT

This study examined urban expansion in Riyadh, London, and Seoul between 2014 and 2024 using Landsat 8 and 9 surface reflectance imagery. Four spectral indices – the normalized difference vegetation index (NDVI), normalized difference built-up index (NDBI), built-up index (BUI), and index-based built-up index (IBI) – were utilized in Google Earth Engine (GEE) to categorize built-up and non-built-up areas. An annual threshold-based classification was conducted, and the resulting maps were visualized through image collages and time-series charts to monitor gains, losses, and overall trends. Riyadh demonstrated rapid horizontal growth, with significant built-up expansion in line with national development plans. Conversely, London experienced minimal urban changes, primarily due to planning policies focusing on densification. Seoul exhibited a dual trend of peripheral expansion and inner-city redevelopment driven by population demand and infrastructure projects. The findings revealed substantial variations in urban growth dynamics among the three cities, highlighting the necessity for context-specific remote sensing thresholds. Despite challenges such as vegetation interference and threshold sensitivity, the multi-index approach offers a robust and replicable method for long-term urban change monitoring. This study provides practical insights for urban planners and policymakers aiming to navigate sustainable development amidst increasing urban pressures.

Keywords: urban expansion, remote sensing, Landsat imagery, built-up classification, Google Earth Engine, sustainable development

1. INTRODUCTION

1.1. Urbanization as a Global Trend. Urbanization represents a prominent global trend in the 21st century. In 1950, only approximately 29% of the global population resided in cities; today, over half of the world's inhabitants live in urban areas. Projections indicate that by 2050, nearly 70% of people will be urban dwellers[1]. This rapid urbanization has led to the emergence of numerous megacities; by 2030, approximately 43 cities are expected to exceed 10 million residents[1]. While growing urban populations drive economic progress and innovation, they also present significant challenges. Unplanned or sprawling urban development strains infrastructure and land resources, resulting in unsustainable growth. Currently, urban areas account for about two-thirds of global energy consumption and more than 70% of greenhouse gas emissions[2]. Given the rapid and extensive urban expansion worldwide – the global urban land area is projected to nearly triple from 2000 to 2030[3] – it is imperative to monitor urban development patterns to inform sustainable planning and policy decisions.

1.2. Remote Sensing for Monitoring Urban Changes. Remote sensing is essential for observing and quantifying urban changes. Satellite imagery offers consistent spatial data over large areas with high detail and frequent revisits, enabling the creation of continuous time series of land-cover changes[4,5]. This synoptic perspective captures the evolution of urban landscapes and is commonly used to detect built-up land expansion at local, regional, and global scales. Monitoring changes in built-up areas – the land covered by buildings and infrastructure – is crucial because they directly reflect city growth. Tracking the expansion of built-up area is vital for urban planning and governance, as the built environment underpins economic and social activities. Accurate long-term monitoring of urban footprints helps planners assess growth patterns, infrastructure needs, and environmental impacts over time. With the growing availability of open satellite data and advanced analytical methods (e.g. cloud-based platforms and time-series analyses), researchers can regularly measure changes in urban extent and built-up density on an annual basis. Remote sensing-based change detection thus offers an objective basis for comparing urbanization trends across cities and decades.

1.3. Study Focus: Riyadh, London, and Seoul. From a comparative perspective, this study analyzes urban change from 2014 to 2024 in three major metropolitan areas–Riyadh, London, and Seoul–representing diverse geographic and developmental contexts. These cities were chosen for their global importance and differing urbanization patterns.

Table 1. *Urban Demographic and Environmental Characteristics of Riyadh, London, and Seoul*

City	Country	Population	Density	Median Age	Income	PM2.5	Waste	Transport	Governance	Climate Risk	RESI (%)	COMM (%)	Green (%)	Water (%)
London	UK	9000000	5600	40.5	48000	13	Advanced	Excellent	London Auth.	Low	38	35	25	2
Seoul	South Korea	9700000	16400	43.7	42000	23	Advanced	Excellent	Metro	Low	40	35	20	5
Riyadh	Saudi Arabia	7600000	4000	30.8	22000	60	Moderate	Moderate	Governorate	Moderate	45	30	15	10

Note: Population, density, income, pollution, transport, governance, climate risk, and land-use distribution figures are approximate and intended to provide context for each city.

Table 1 summarizes key demographic and environmental characteristics of the three study cities, including population size, urban density, median age, income, $PM_{2.5}$ pollution levels, waste management status, public transport quality, governance structure, climate risk level, and land-use distribution. These city-specific indicators provide crucial context for interpreting land-cover change patterns.

1.4. Objectives . This study examined urban expansion in Riyadh, London, and Seoul from 2014 to 2024 by employing remote sensing indices. The primary objectives are as follows:

- Quantification of built-up changes using NDVI, NDBI, BUI, and IBI.
- Comparison of urban growth trends across the three cities with different development contexts.
- Evaluation of index performance in capturing urbanization under varying environmental and urban conditions.
- Analysis of gain and loss in built-up areas to identify dynamic changes.
- Support for urban planning through visual tools and data that inform sustainable development strategies.

2. DATASET AND TOOLS

2.1. Dataset. This study used yearly Landsat imagery from 2014 to 2024 to examine land-cover transformations in Riyadh, London, and Seoul. The dataset included Landsat 8 and 9 Surface Reflectance images at 30-meter resolution, accessed via GEE using specific path/row identifiers corresponding to each city's location.

Table 2 provides the Landsat 8 and 9 band designations used in this study. Notably, bands 4 (red) and 5 (near-infrared) were used for NDVI calculation, and bands 5 (NIR) and 6 (SWIR1) for NDBI calculation, among others (e.g., a green band and SWIR for the water index component of IBI).

3. METHODOLOGY

This study employed a systematic workflow to monitor and analyze urban growth in the three cities.

3.1. Path/Row Selection and Imagery Access. Using USGS EarthExplorer, the appropriate WRS-2 path and row values were identified for each city's Landsat scenes. Landsat 8 and 9 surface reflectance images were then obtained through GEE. One nearly cloud-free image per year (2014–2024) was selected for each city to represent annual conditions. To reduce seasonal variability, images were chosen from the same month each year (and on similar dates when feasible), ensuring consistency of vegetation status and minimizing seasonal bias in the index values.

Table 2. *Landsat 8 OLI and TIRS Band Designations and Resolutions*

Bands	Wavelength (micrometers)	Resolution (meters)
Band 1 - Coastal aerosol	0.43 - 0.45	30
Band 2 - Blue	0.45 - 0.51	30
Band 3 - Green	0.53 - 0.59	30
Band 4 - Red	0.64 - 0.67	30
Band 5 - Near Infrared (NIR)	0.85 - 0.88	30
Band 6 - SWIR 1	1.57 - 1.65	30
Band 7 - SWIR 2	2.11 - 2.29	30
Band 8 - Panchromatic	0.50 - 0.68	15
Band 9 - Cirrus	1.36 - 1.38	30
Band 10 (TIRS) 1 - Thermal Infrared	10.60 - 11.19	100
Band 11 (TIRS) 2 - Thermal Infrared	11.50 - 12.51	100

Note: NDVI = (Band 5 – Band 4) / (Band 5 + Band 4); NDBI = (Band 6 – Band 5) / (Band 6 + Band 5); BUI = (NDBI – NDVI) / (NDBI + NDVI); IBI uses a combination of NDBI, NDVI, and a Modified NDWI (which uses Band 3 and Band 6). All indices were calculated on GEE with these band assignments.

Table 3. *Formulas for Spectral Indices Used in Urban Classification*

Index	Equation
NDVI (Normalized Difference Vegetation Index)	(NIR - Red) / (NIR + Red) = (Band 5 - Band 4) / (Band 5 + Band 4)
NDBI (Normalized Difference Built-up Index)	(SWIR - NIR) / (SWIR + NIR) = (Band 6 - Band 5) / (Band 6 + Band 5)
BUI (Built-up Index)	(NDBI - NDVI) / (NDBI + NDVI)
IBI (Index-based Built-up Index)	(NDBI - (NDVI + MNDWI)) / (NDBI + (NDVI + MNDWI)) MNDWI = (Green - SWIR) / (Green + SWIR) = (Band 3 - Band 6) / (Band 3 + Band 6)

3.2. Index Calculation in GEE. Four spectral indices were computed using the selected Landsat bands for each image (Table 3).

Threshold values were applied to each index to classify pixels as built-up or non-built-up. From 2014 to 2024, each city used a fixed threshold per index, determined through visual calibration to minimize misclassification. For example, Seoul's dense vegetation cover required a stricter NDVI threshold to avoid labeling vegetated areas as built-up, whereas Riyadh's arid environment required adjusting NDVI to detect sparse vegetation over bare soil. The thresholds (detailed in Table 3) were maintained at a constant level for all years to ensure temporal consistency and comparability. Table 4 lists the mathematical formulas of these indices as implemented. GEE's expression syntax was used to calculate each index from the Landsat bands, and annual index maps were exported for analysis.

These four indices have well-established uses in urban remote sensing: the NDBI was originally proposed as an automated means to map built-up areas from Landsat TM imagery[10], while the IBI was developed to delineate built-up land features by combining built-up and vegetation indices[11]. In this study, the chosen indices and thresholds were tailored to each city's conditions (e.g., accounting for Riyadh's sparse vegetation and Seoul's abundant greenery) to improve classification accuracy.

3.3. Classification and Area Estimation. Using the binary masks derived from each index (built-up vs. non-built-up), the pixel count of built-up area was calculated and converted to area (square kilometers) for each city and year. This provided annual estimates of built-up and non-built-up land area under each index.

4. RESULTS AND DISCUSSION

This section presents the spatial and quantitative outcomes of the urban change analysis for Riyadh, London, and Seoul from 2014 to 2024. The results derive from the spectral index classification maps (NDVI, NDBI, BUI, IBI), gain/loss spatial comparisons, and time-series graphs. The findings highlight the direction and magnitude of urban transformation in each city, illustrating how built-up and non-built-up areas evolved over the decade. Each city is discussed in turn, through visual interpretation of the classification maps and detailed analysis of the index-based trends, supported by the corresponding figures and graphs.

4.1. Riyadh.

4.1.1. Index-Based Urban Change (2014--2024). RGB and index maps (NDVI, NDBI, BUI, IBI) show Riyadh's steady outward expansion, especially in the north, east, and southeast (Figure 1a). Uniform thresholds ensured that changes reflect real growth. NDVI underestimated built-up land in earlier years due to sparse desert vegetation, while NDBI captured expansion

Table 4. *Threshold Values for Built-up Classification (2014 vs. 2024)*

City	Index	2014 Threshold (y1)	2024 Threshold (y2)
Riyad	NDVI	0.20	0.20
Riyadh	NDBI	0.05	0.05
Riyadh	BUI	-0.22	-0.22
Riyadh	IBI	-0.62	-0.62
London	NDVI	0.15	0.15
London	NDBI	0.04	0.04
London	BUI	-0.22	-0.22
London	IBI	-0.55	-0.55
Seoul	NDVI	0.19	0.19
Seoul	NDBI	0.04	0.04
Seoul	BUI	-0.20	-0.20
Seoul	IBI	-0.55	-0.55

Note: For each city and index, a single fixed threshold is used for all years (2014–2024).

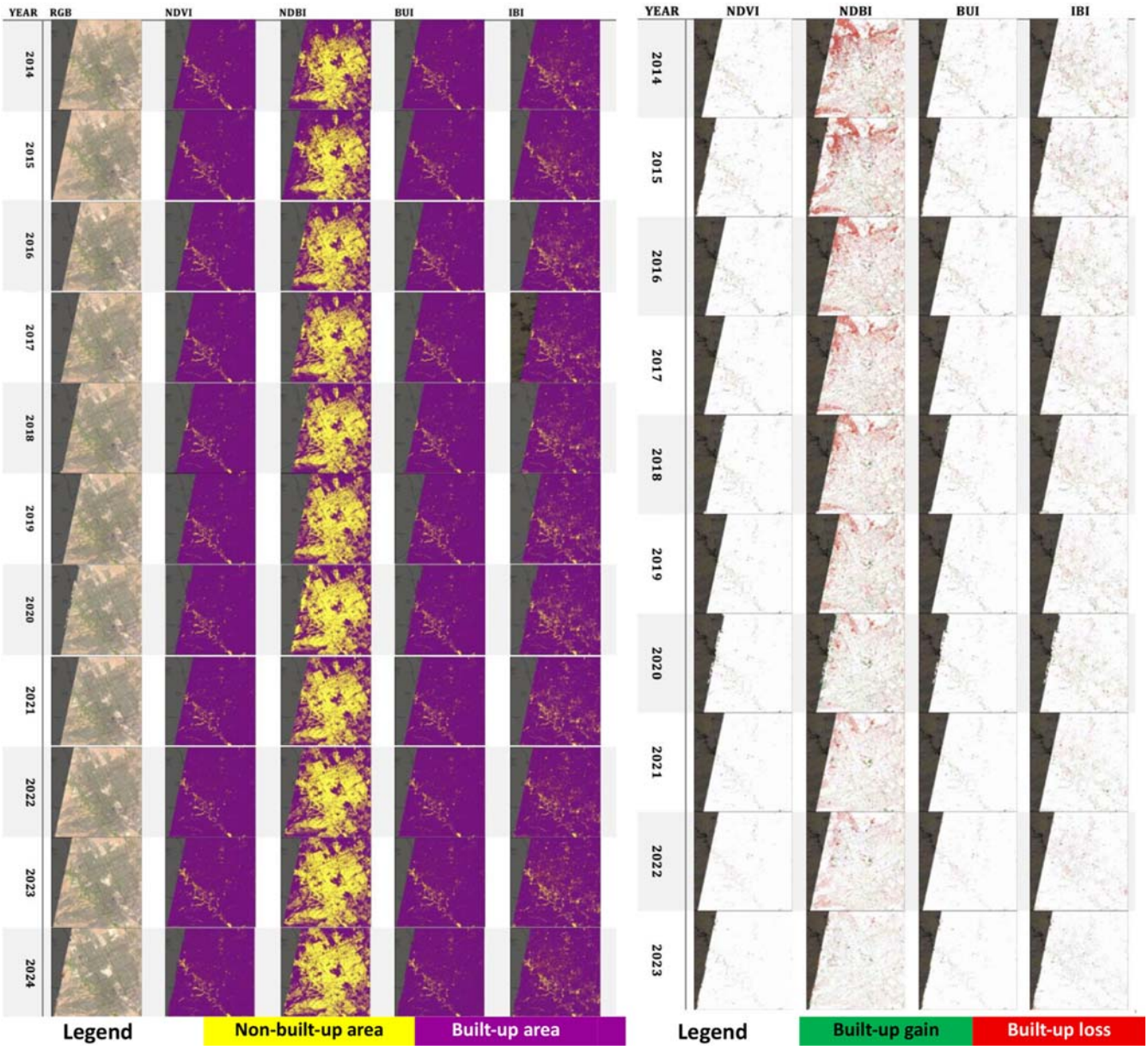


Figure 1. (*a*) Annual built-up vs. non-built-up classification maps and (b) built-up gain/loss maps for Riyadh (2014–2024). In (a), purple areas indicate built-up land and yellow areas indicate non-built-up land for each year. In (b), green areas show new development (gain) and red areas show reclassified loss of built-up land, each year compared to the 2024 baseline.

effectively but showed minor spectral fluctuations. BUI, integrating NDVI and NDBI, clearly outlined urban boundaries and highlighted infill within the core. IBI offered a more conservative but less noisy depiction, emphasizing stable urban centers. Overall, the maps indicate consistent horizontal growth, with new development zones emerging every 1–2 years at the city's fringes.

4.1.2. Gain and Loss Analysis (2014--2024). Figure 1b shows built-up gains (green) concentrated on Riyadh's periphery—north, east, and southeast—linked to suburban growth, ring-road extensions, and industrial projects. Major surges occurred in 2016–2018 and 2021–2023, aligning with Vision 2030 housing and infrastructure initiatives. Losses (red) were minimal, primarily in transitional construction or redevelopment sites that were cleared before rebuilding. The largest net expansion occurred from 2018 to 2020, with further outward growth sustained after 2021. Overall, gains far outweighed losses, confirming a persistent outward expansion with only minor, temporary contractions.

4.1.3. Time-Series Analysis (2014--2024). Figure 2 summarizes Riyadh's temporal dynamics.

- **Built-up area:** NDVI and BUI recorded the largest extents, increasing from ~1,250 km^2 in 2014 to over 1,350 km^2 in 2024, with a sharp peak in 2015, a decline until 2019, and recovery thereafter. IBI followed the same trend at slightly lower values (~1,210 to 1,270 km^2). In contrast, NDBI declined from ~792 to ~702 km^2, underestimating growth due to SWIR sensitivity to bright desert soils.
- **Non-built-up area:** All indices indicated a decline. NDVI/BUI dropped from ~340 to 250 km^2, and NDBI

from ~380 to 320 km^2. IBI reported larger values (~800–890 km^2) but still showed a downward trend, consistent with conservative classification of urban land.

- **Annual gain:** NDBI recorded the highest additions (>60 km^2 in most years, peaking near 90 km^2 in 2020), but much of this reflected temporary spectral changes. IBI and BUI showed steadier gains (10–40 km^2), while NDVI detected <20 km^2 annually, affected by vegetation variability.
- **Annual loss:** NDBI reported unrealistically high losses (60–75 km^2 annually), revealing over-sensitivity. IBI and BUI showed stable values (15–30 km^2), and NDVI only minor, steady losses (~10–20 km^2), likely from seasonal vegetation shifts.

Except for NDBI artifacts, the indices confirm that once land was converted to urban use, it rarely reverted. Riyadh's urban growth during 2014–2024 was effectively irreversible and outward-oriented.

4.2. London.

4.2.1. Index-Based Urban Change (2014--2024).

Figure 3a shows London's stable urban footprint with only localized changes. NDVI underestimated built-up land in earlier years due to the abundance of vegetation (parks, gardens, and tree-lined streets). However, from 2017, small increases appear, linked to redevelopment and brownfield housing projects. NDBI consistently highlighted dense urban cores (Central London, Canary Wharf, Thames corridors) with little outward change. BUI emphasized slight intensification rather than sprawl, reflecting infill within existing boundaries. IBI mapped the broadest extent, occasionally overestimating built-up areas in suburban zones where vegetation and structures intermingle. Overall, London's changes reflect infill, regeneration, and densification, not horizontal expansion.

4.2.2. Gain and Loss Analysis (2014--2024).

Figure 3b shows dispersed but modest gains, mainly in central and inner-city boroughs (e.g., King's Cross, Stratford, Crossrail corridors). Peaks occurred in 2016, 2020, and 2023, aligning with major redevelopment completions. Losses were minimal, often temporary clearances during regeneration or minor classification shifts (e.g., vegetation regrowth). Across all indices, gains outweighed losses. BUI and NDBI captured sharper growth in dense districts, while NDVI and IBI reflected subtle changes in mixed-use or vegetated areas. Together, these results confirm London's incremental, controlled transformation, driven by redevelopment and density increases rather than sprawl.

4.2.3. Time-Series Analysis (2014--2024).

- **Built-up area trends:** Figure 4 (top-left) shows London's built-up area remained stable, with only minor fluctuations. IBI reported the largest extents (220–330 km^2), peaking in 2019 during redevelopment surges, then stabilizing near 225 km^2 by 2024. BUI averaged 130–180 km^2, dipping slightly in 2016–2018 before recovering. NDBI, the lowest (50–140 km^2), declined to 2018 then rose to its maximum in 2024. NDVI ranged from 60–150 km^2, peaking in 2024, reflecting improved classification of redeveloped zones. Collectively, these patterns confirm limited net growth, driven by infill and regeneration rather than sprawl.

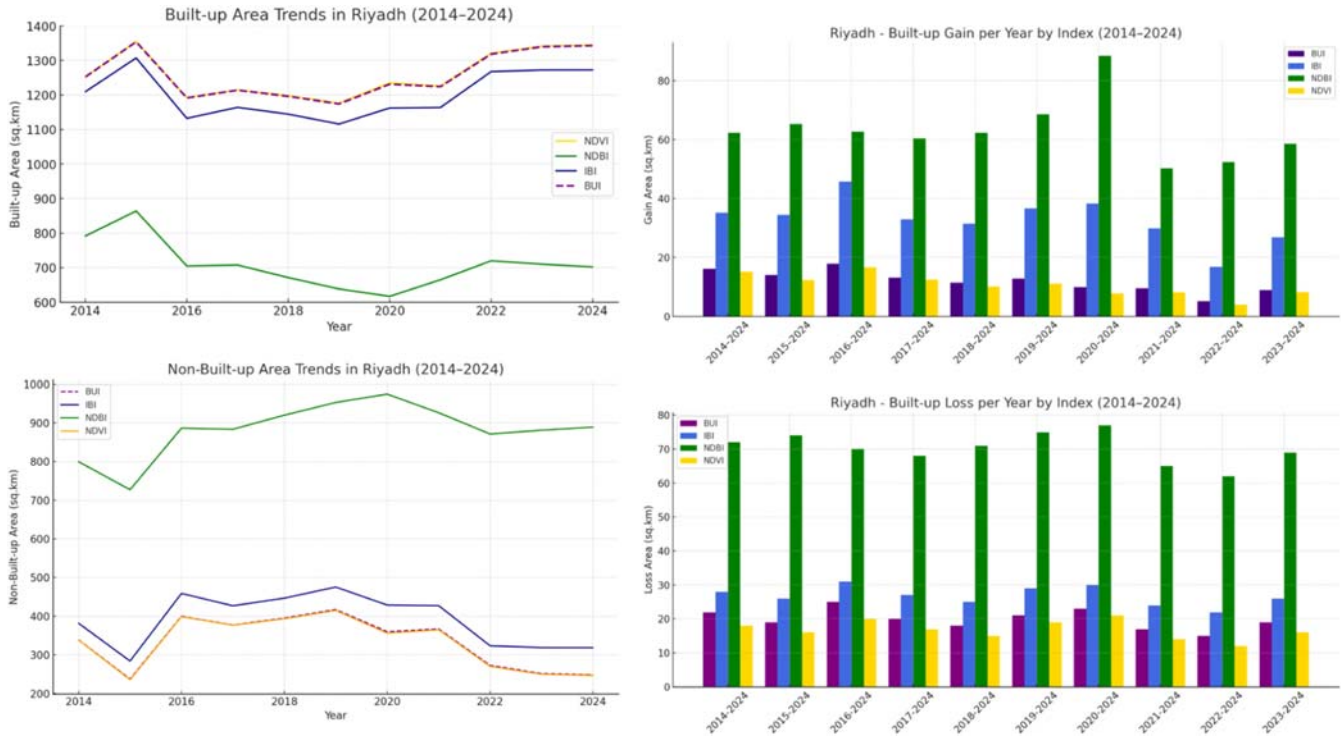


Figure 2. Time-series plots of Riyadh's built-up dynamics (2014–2024): (top-left) Total built-up area; (bottom-left) total non-built-up area; (top-right) annual built-up area gain; (bottom-right) annual built-up area loss. These plots compare results from four indices (NDVI, NDBI, BUI, IBI), highlighting overall growth in built-up extent, the corresponding loss of open land, and the magnitude of yearly changes. (Note: NDBI's higher gain/loss fluctuations reflect its index sensitivity in arid conditions.)

- **Non-built-up area trends:** Figure 4 (bottom-left) indicates little overall change. NDVI and NDBI consistently showed ~1,490–1,540 km^2, reflecting London's preserved parks and green belt. IBI dropped from ~1,380 to 1,300 km^2 during densification (2014–2018), then recovered slightly. BUI remained 1,400–1,470 km^2, with a gentle decline after 2017 linked to incremental housing infill. In summary, vegetation-sensitive indices highlight stable green cover, while built-up–focused indices captured subtle infill-driven reductions.
- **Annual built-up gain:** Figure 4 (top-right) shows modest yearly gains across all indices, consistent with London's slow growth. IBI and NDVI registered 75–90 km^2 per year, BUI ~55–70 km^2, and NDBI <90 km^2. The absence of sharp peaks reflects incremental redevelopment projects (e.g., brownfield conversions, small housing estates).
- **Annual built-up loss:** Figure 4 (bottom-right) shows consistently low losses. IBI reported the largest area (~80–100 km^2 mid-decade, declining to ~70 km^2 by 2024), which was mostly attributed to temporary demolition before redevelopment. BUI losses averaged 30–40 km^2, with a small rise in 2017–2018. NDBI losses stayed under 40 km^2, with slight late increases linked to vegetation initiatives or temporary clearances. NDVI showed the smallest (10–25 km^2), reflecting transient vegetation regrowth. Overall, losses were temporary and offset quickly by redevelopment, with no lasting contraction of the urban footprint.

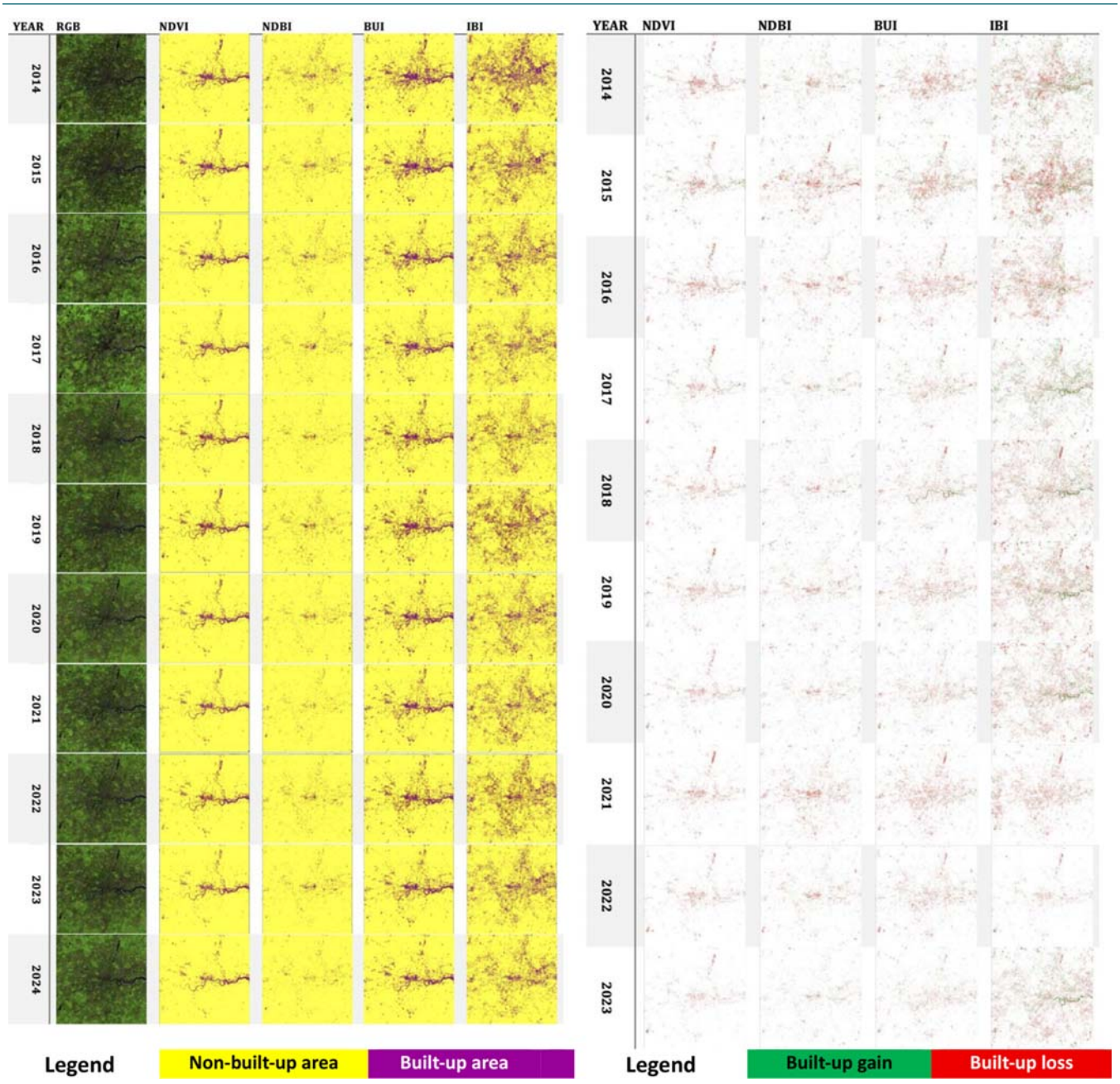


Figure 3. (a) Annual built-up vs. non-built-up classification maps and (b) built-up gain/loss maps for London (2014–2024). In (a), built-up areas are shown in purple and non-built areas in yellow. In (b), green areas denote where new development occurred (gains) and red areas denote where previously built-up areas were reclassified as non-built (losses), relative to 2024.

London's time series confirms a mature, stable urban system, characterized by incremental redevelopment, strong green space preservation, and negligible permanent losses. Growth occurred primarily through densification and regeneration, reinforcing the city's controlled planning framework.

4.3. Seoul.

4.3.1. Index-Based Urban Change (2014--2024). Figure 5a shows Seoul's dynamic transformation, with shifts in both central districts and expanding peripheries. NDVI initially underestimated built-up land due to the city's extensive vegetation and green infrastructure, but after 2017 it captured clearer urban delineation as pressure on fringe green areas intensified. NDBI consistently highlighted dense cores (e.g., Gangnam, Yeouido, downtown) and later detected suburban expansion into Goyang and Hanam after 2018. BUI outlined the built environment well, revealing stable cores and outward suburban growth post-2018. IBI showed sharp expansion from 2014–2017, then a decline during 2017–2019 linked to demolition and redevelopment cycles, before stabilizing after 2020. Together, the indices illustrate Seoul's dual trajectory: outward suburban expansion coupled with vertical redevelopment and densification of central districts, reflecting national housing strategies, new town projects, and transport-led growth.

4.3.2. Gain and Loss Analysis (2014--2024). As shown in Figure 5b, gains (green) were concentrated in outer districts and satellite cities—especially Songpa, Gangdong, and Gyeonggi Province—peaking in 2016–2017 and again in 2023, coinciding with housing drives and delayed infrastructure projects. Losses (red) were more pronounced than in Riyadh or London, largely between 2016–2019 in central areas under redevelopment. IBI registered the greatest losses, often exaggerating demolition phases by classifying cleared or debris-covered land as non-built. By 2020 and especially 2023, strong net gains reappeared, led by BUI and NDBI, as suburban projects were completed and redevelopment areas reoccupied. Overall, Seoul's gain–loss dynamics reflect its two-pronged **strategy**: continuous inner-city redevelopment, producing cycles of temporary decline, combined with sustained suburban expansion. IBI captured the disruption of redevelopment, while BUI and NDBI provided clearer evidence of long-term growth. These patterns underscore Seoul's compact, high-density planning in the core, balanced with outward expansion into planned satellite cities.

4.3.3. Time-Series Analysis (2014--2024).

- **Built-up area trends:** Figure 6 (top-left) shows Seoul's built-up area peaking around 2017, then declining before stabilizing by 2024. IBI recorded the highest extent (~775 km^2 in 2017), which dropped to approximately 369 km^2 as redevelopment cycles cleared central districts. BUI peaked at ~509 km^2 in 2017, declining steadily to ~279 km^2. NDVI spiked to ~392 km^2 in 2017 (construction-driven vegetation loss) before falling to ~227 km^2 by 2024. NDBI peaked near ~366 km^2 in 2017, then declined to ~150 km^2 by 2024. The patterns reflect rapid mid-decade expansion followed by demolition and rebuilding, with stabilization by the early 2020s.
- **Non-built-up area trends:** Figure 6 (bottom-left) shows the inverse. IBI dropped sharply to ~816 km^2 in 2017, then rebounded to ~1,222 km^2 by 2024 as cleared sites were reclassified as non-built. NDBI showed the largest non-built extents (~1,225–1,448 km^2), indicating underestimation of built-up areas in some contexts. BUI remained relatively stable (~1,083–1,312 km^2), while NDVI fluctuated between ~1,199

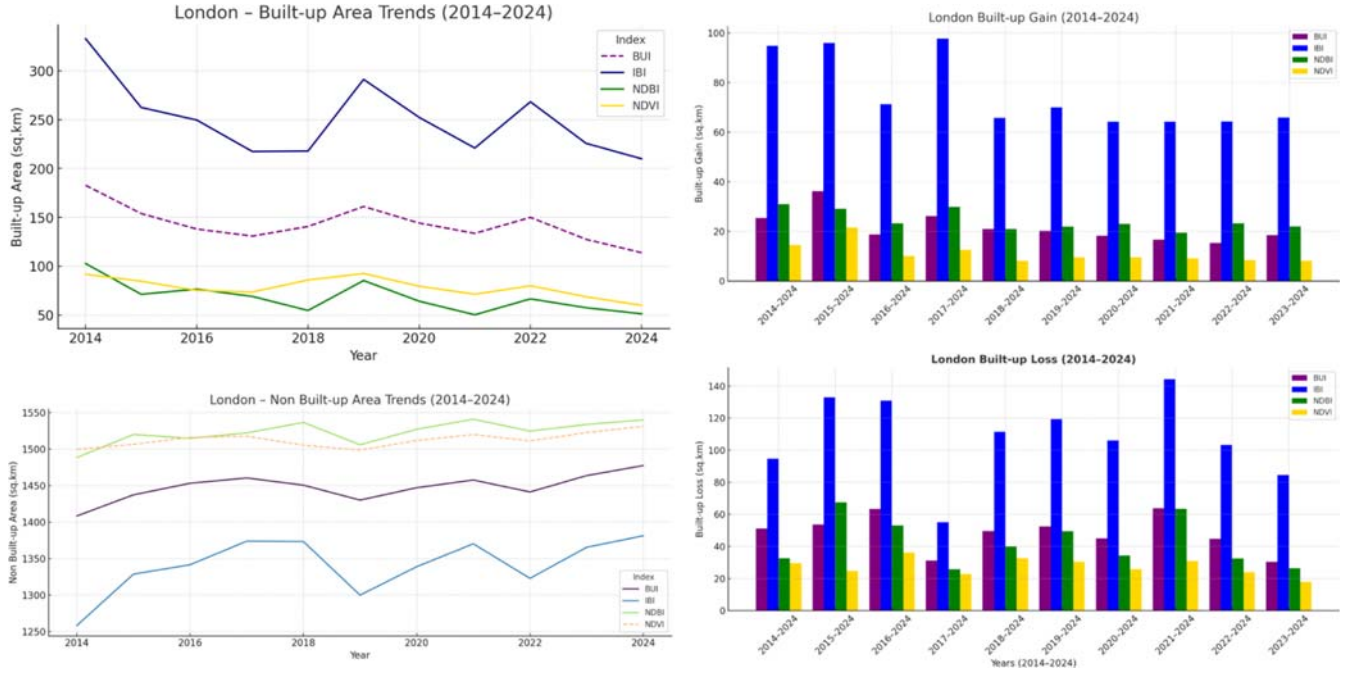


Figure 4. Time-series plots of London's urban land dynamics (2014–2024): (a) Built-up area; (b) non-built-up area; (c) annual built-up gain; (d) annual built-up loss. Results are shown according to four indices (NDVI, NDBI, BUI, IBI). London's curves demonstrate minimal net change in total built-up or open areas, with slow gains and very small losses reflecting a stable, mature city with growth occurring through gradual redevelopment rather than expansion.

and ~1,364 km^2, reflecting Seoul's resilient green cover. Overall, non-built-up land recovered after 2018, underscoring the city's integration of vegetation into its redevelopment efforts.

- **Annual built-up gain:** Figure 6 (top-right) shows IBI capturing the largest gains, peaking at ~165 km^2 in 2014–2015 and rising again to ~125 km^2 in 2023–2024. BUI showed moderate but steady gains (30–70 km^2), with peaks in 2016 and 2023. NDBI gains rose gradually to ~55 km^2 by 2024, reflecting fringe expansion. NDVI was more variable (25–70 km^2), with early spikes tied to rapid vegetation loss. The trends highlight two growth phases—early (2014–2016) and late (2023–2024)—separated by a redevelopment lull.
- **Annual built-up loss:** Figure 6 (bottom-right) reveals the disruptive impact of redevelopment. IBI showed extreme losses (~200–500 km^2) between 2017 and 2020, declining to ~150 km^2 by 2024. BUI losses peaked at ~260 km^2 in 2017–2018 before stabilizing below 100 km^2. NDBI registered mid-decade losses (~130–250 km^2), which decreased to approximately 50 km^2 by 2024. NDVI losses (100–200 km^2) peaked in 2017–2018 as cleared land temporarily re-vegetated. Overall, losses were temporary and redevelopment-driven, with balance restored by the early 2020s.

Seoul's time series highlights a cyclical pattern: rapid expansion through 2017, followed by heavy demolition and redevelopment through 2020, and then stabilization with renewed suburban growth by 2023–2024. Unlike Riyadh's steady outward growth or London's incremental densification, Seoul's urban change was

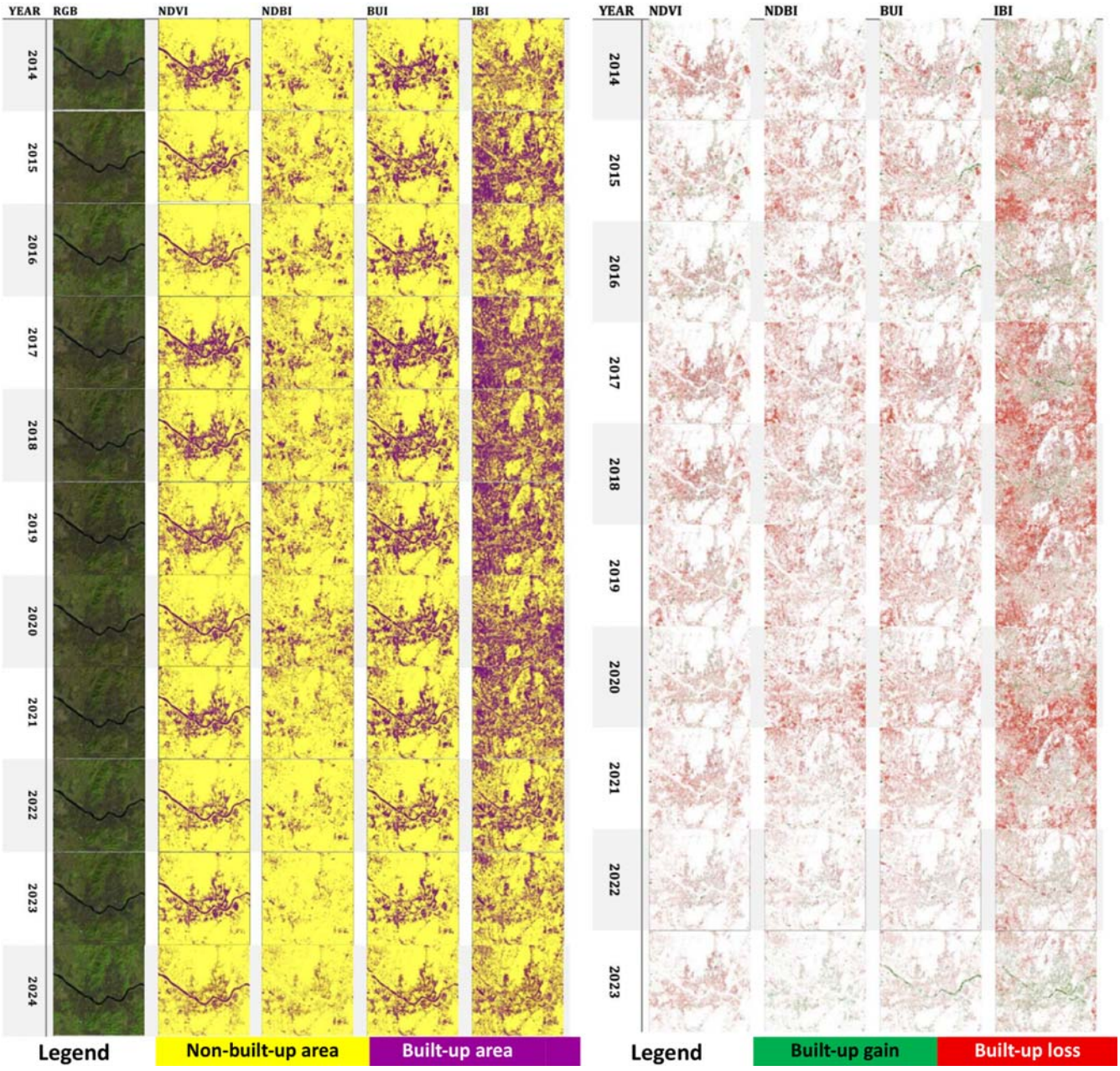


Figure 5. (a) Annual built-up vs. non-built-up classification maps and (b) built-up gain/loss maps for Seoul (2014–2024). In (a), purple indicates built-up and yellow indicates non-built-up areas for each year. In (b), green highlights areas of new urban development (gain) and red highlights areas of built-up land loss or demolition, compared to 2024.

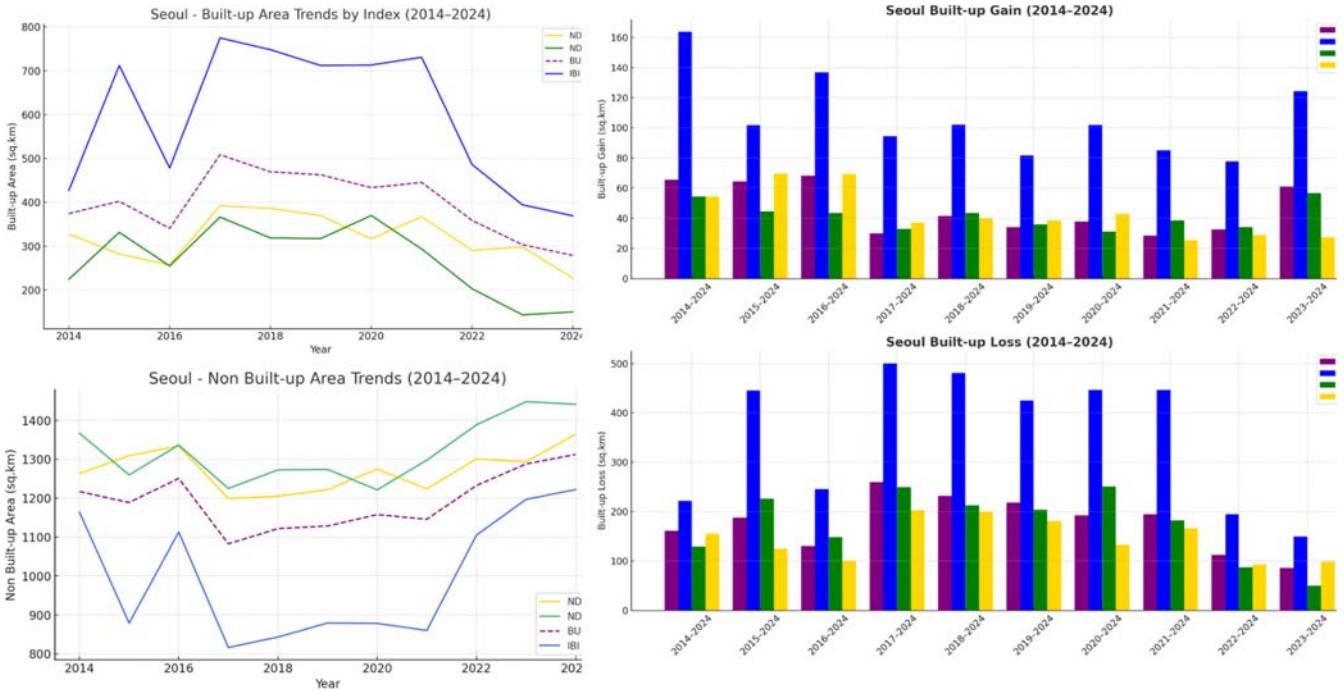


Figure 6. Time-series plots of Seoul's urban land dynamics (2014–2024): (a) Built-up area; (b) non-built-up area; (c) annual built-up gains; (d) annual built-up losses. Each panel shows results from NDVI, NDBI, BUI, and IBI. Seoul's trends are characterized by a mid-period peak in built-up area followed by stabilization, reflecting its redevelopment cycle. The substantial spikes in annual gains and losses captured by IBI highlight periods of rapid construction and demolition, respectively, whereas the other indices indicate more moderate, continuous change and recovery of open spaces following redevelopment.

highly dynamic, reflecting its dual strategy of core redevelopment and suburban expansion.

5. CONCLUSIONS

This study analyzed urban changes in Riyadh, London, and Seoul from 2014 to 2024 using multi-index remote sensing techniques in Google Earth Engine. By employing four spectral indices (NDVI, NDBI, BUI, and IBI) and a consistent threshold-based classification, the analysis captured diverse patterns of urban expansion: Riyadh demonstrated rapid outward growth of its built-up area, London exhibited minimal change constrained by stringent planning regulations, and Seoul showcased a combination of peripheral expansion and significant central redevelopment.

Several limitations were noted in the methodology, including classification errors due to index sensitivity and the challenge of vegetation masking built-up surfaces (especially in green cities like London and Seoul). In some cases, observed trends in the indices were influenced by how each index responds to mixed land cover (for example, the oversensitivity of IBI in redevelopment zones), rather than purely physical land-use change. These considerations underscore the importance of methodological transparency and selecting the optimal combination of indices for urban change detection.

Overall, the multi-index remote sensing approach proved to be a valuable tool for long-term urban monitoring. When applied with clear methodological considerations and calibrated to local conditions, it can provide city officials and planners with regular, objective insights into development patterns. For Riyadh, the findings underscore the pace of sprawl and the need for sustainable growth management; for London, they highlight the success of containment policies and the focus on regeneration; for Seoul, they reveal the intensity of redevelopment and the importance of balancing growth with green space preservation.

Moving forward, enhancements such as incorporating machine learning classification (to learn optimal thresholds or classifications automatically), using time-series analysis for continuous monitoring, and integrating socio-economic datasets (population, infrastructure, policy changes) could provide an even more comprehensive understanding of urban dynamics. Despite its challenges, remote sensing – especially with freely available satellite data and powerful cloud platforms – offers a replicable and scalable method for cities worldwide to monitor development, evaluate land-use changes, and inform more sustainable urban planning decisions.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Zahra Alhaddad – *Architecture and City Design Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia;* 0009-0002-4322-2968
Email: s202255640@kfupm.edu.sa

Corresponding Author

Muhammad Bilal – *Research Mentor, Architecture and City Design Department, College of Design and Built Environment, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia; Center for Aviation and Space Exploration;* 0000-0003-1022-3999
Email: muhammad.bilal@kfupm.edu.sa

ACKNOWLEDGMENTS

This research was conducted as part of an undergraduate program in the Department of Architecture and City Design at KFUPM. The author thanks the faculty advisors and the JURI editorial team for their guidance and support throughout the research and writing process.

The author also acknowledges the use of AI-based language tools to improve the clarity and structure of the manuscript's English, without altering the research content or originality.

REFERENCES

(1) United Nations. *World Urbanization Prospects: The 2018 Revision*. Department of Economic and Social Affairs, Population Division (2018).

(2) UN-Habitat. *World Cities Report 2020: The Value of Sustainable Urbanization*. United Nations Human Settlements Programme (2020).

(3) Seto, K. C., Güneralp, B. and Hutyra, L. R. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16083–16088 (2014).

(4) Zhou, Y., Smith, S. J., Zhao, K. and Imhoff, M. A cluster-based method to map urban area from MODIS land surface temperature and NDVI. *Remote Sens. Environ.* **152**, 1–14 (2014).

(5) Weng, Q. Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends. *Remote Sens. Environ.* **117**, 34–49 (2012).

(6) Zha, Y., Gao, J. and Ni, S. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *Int. J. Remote Sens.* **24**, 583–594 (2003).

(7) Xu, H. A new index for delineating built-up land features in satellite imagery. *Int. J. Remote Sens.* **29**, 4269–4276 (2008).

(8) Alghamdi, A., Aldomae, M. and Mubarak, F. Monitoring and modeling of urban expansion using GIS and remote sensing techniques in Riyadh, Saudi Arabia. *Egypt. J. Remote Sens. Space Sci.* **20**, 211–217 (2017).

(9) Office for National Statistics (ONS). Population estimates for the UK, England and Wales, Scotland and Northern Ireland: mid-2022. https://www.ons.gov.uk/ (2023).

(10) Gallent, N., Andersson, J. and Bianconi, M. *Planning on the Edge: The Context for Planning at the Rural-Urban Fringe*. Routledge (2006).

(11) Korea Statistical Information Service (KOSIS). https://kosis.kr (2024).

KFUPM
Journal of Undergraduate Research International

Diamond Open Access 123

Article Material Science and Engineering

Comprehensive Numerical Analysis of Highly Efficient Lead-Free $CH_3NH_3SnI_3$-Based Perovskite Solar Cell

Foyzul Karim[1], Md. Habibur Rahman Aslam[2] and Anisul Islam Suva[3]*

Cite https://doi.org/10.64589/juri/207995

Submitted: June 01, 2025 **Revised:** July 07, 2025 **Accepted:** August 19, 2025

ABSTRACT

Perovskite solar cells (PSCs) are highly efficient photovoltaic technologies, offering bandgap tunability and low production costs. However, lead toxicity, architectural limitations, and defect-induced recombination hinder their commercialization. This study investigates $CH_3NH_3SnI_3$ as a nontoxic absorber in an FTO/WSe_2/$CH_3NH_3SnI_3$/NiO/Au device architecture. Through systematic numerical simulations using Solar Cell Capacitance Simulator-1D (SCAPS-1D), the key design parameters—layer thicknesses, densities of doping, and defect concentrations—are carefully optimized. The resulting device structure exhibits a notable power conversion efficiency of 34.74%, V_{oc} of 1.1084 V, J_{sc} of 36.7788 mA/cm^2, fill factor of 85.22%, and peak quantum efficiency of 99.95% at 390 nm under AM 1.5G illumination. Sensitivity analysis reveals that both bulk ($N_t > 10^{14}$ cm^{-3}) and interface ($N_{int} > 10^{17}$ cm^{-3}) defects drastically degrade the performance, particularly at the $CH_3NH_3SnI_3$/WSe_2 interface. These findings have significant implications for highly efficient lead-free PSC designs.

Keywords: perovskite, photovoltaics, doping, defects, SCAPS-1D

1. INTRODUCTION

Solar energy has become the center of the world's search for green and sustainable energy substitutes in a bid to replace fossil fuels. Among cutting-edge photovoltaic (PV) technologies, perovskite solar cells (PSCs) have garnered considerable interest owing to their impressive power conversion efficiency (PCE), cost-effective fabrication, and amenability to scalable thin-film technologies [1,2]. Specifically, organic-inorganic halide PSCs have undergone dramatic efficiency improvements, rising from 3.8% to over 25% over the past decade, thus rivaling conventional silicon-based PV systems [3–5].

Despite these advances, the widespread adoption of PSCs is hindered by lead (Pb) toxicity in common absorbers, such as $CH_3NH_3PbI_3$, which poses major environmental and health risks [6–8]. This has stimulated research into nontoxic alternatives. The most promising lead-free alternative is $CH_3NH_3SnI_3$ (methylammonium tin iodide), a perovskite material with a 1.3 eV direct bandgap, excellent photon harvesting, and enhanced charge carrier mobility owing to its partially covalent Sn–I bond network [9–11]. These properties make $CH_3NH_3SnI_3$ a viable absorber for high-performance green PSCs.

However, $CH_3NH_3SnI_3$-based PSCs face several challenges. Sn^{2+} is easily oxidized to Sn^{4+}, reducing material stability and creating high-level defects that serve as recombination centers [12]. Additionally, bulk and interfacial defect densities continue to limit efficiency. Another key factor affecting device performance is the choice of charge transport layers (CTLs)—namely, the electron transport layer (ETL) and hole transport layer (HTL) [13]. Conventional ETLs (e.g., TiO_2) and HTLs (e.g., spiro-OMeTAD) suffer from low thermal stability, high financial cost, and degradation tendency under ambient conditions [13,14].

In this context, tungsten diselenide (WSe_2) has been presented as a high-performance ETL alternative. Owing to its favorable conduction band alignment, superior carrier mobility, good stability, and low defect density, WSe_2 can be used for electron transportation and extraction [15,16]. Coupled with thermostable NiO as the HTL, these inorganic CTL configurations contribute to enhanced performance in tin-based PSCs [17,18].

In this study, the Solar Cell Capacitance Simulator-1D (SCAPS-1D) was employed to optimize a lead-free FTO/WSe_2/$CH_3NH_3SnI_3$/NiO/Au structure. A detailed parametric study was conducted, examining the impacts of layer thickness, doping concentration, and density of defects on the PV performance metrics. Special emphasis was placed on quantifying and minimizing recombination losses, specifically at the absorber/ETL interface, to maximize the potential of lead-free PSCs.

2. METHODOLOGY

2.1. SCAPS-1D Simulation. Numerical analysis was performed with SCAPS-1D, a device modeling software created by the ELIS Department of Ghent University [19], to simulate and optimize tin-based PSC behavior. SCAPS-1D is renowned within

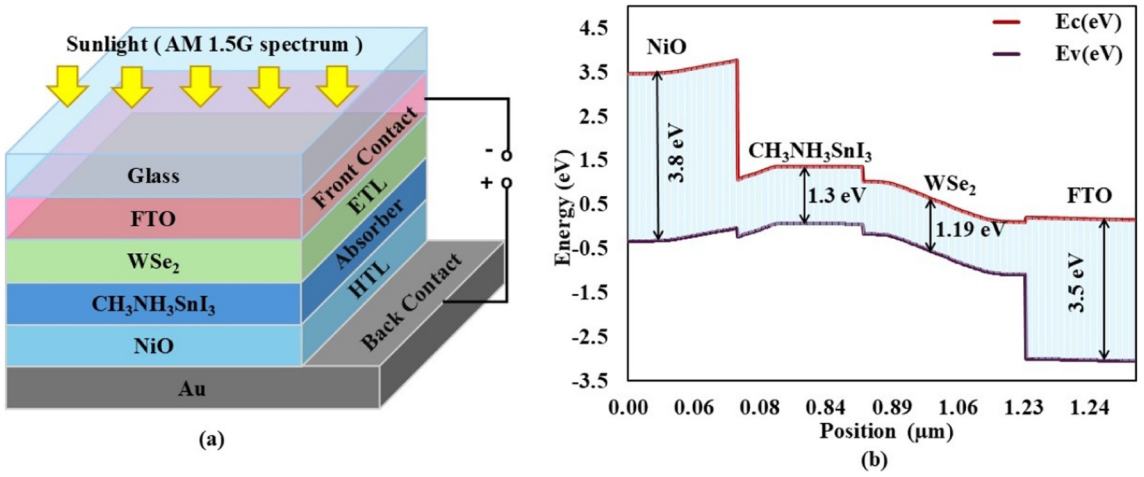


Figure 1. (a) Schematic representation and (b) energy band diagram of the designed PSC

the PV scientific community for its ability to simulate complex multilayer solar cell geometries by solving elementary semiconductor equations, that is, Poisson's equation (Eq. 1), hole continuity equation (Eq. 2), and electron continuity equation (Eq. 3), for steady-state conditions under AM 1.5G solar irradiation. Charge transport under an electric field and concentration gradients was simulated using the drift-diffusion equations (Eqs. 4 and 5). The simulator also encompasses key processes such as photon absorption, carrier generation, transport, recombination, and extraction at the interfaces. The incorporation of the Shockley–Read–Hall model enables the realistic treatment of defect-assisted recombination. These features enable the precise analysis of current–voltage (J–V) characteristics, quantum efficiency (QE), and recombination dynamics.

$$\frac{d^2\Psi(x)}{dx^2} = \frac{q}{\epsilon_o \epsilon_r}[p(x) - n(x) + N_D - N_A + \rho_p - \rho_n] \quad (1)$$

$$\frac{\delta p}{\delta t} = \frac{1}{q}\frac{\delta J_p}{\delta x} + G_p - R_p \quad (2)$$

$$\frac{\delta n}{\delta t} = \frac{1}{q}\frac{\delta J_n}{\delta x} + G_n - R_n \quad (3)$$

$$J_n(x) = q\mu_n n(x)\varepsilon(x) + qD_n\frac{dn(x)}{dx} \quad (4)$$

$$J_p(x) = q\mu_p p(x)\varepsilon(x) + qD_p\frac{dp(x)}{dx} \quad (5)$$

where

Ψ denotes the electrostatic potential

q represents the elementary charge

ε_o and ε_r are the vacuum permittivity and the material's relative permittivity, respectively

p and n indicate hole and electron concentrations

N_D and N_A are the donor and acceptor doping concentrations

ρ_p and ρ_n signify hole and electron charge concentrations

J_p and J_n refer to the current densities

G_p and G_n are the respective generation rates

R_p and R_n correspond to the recombination rates

μ_n and μ_p are the mobilities of electrons and holes

$\varepsilon(x)$ describes the electric field as a function of position

D_n and D_p are the diffusion coefficients

2.2. Basic Structure, Operation, and Input Parameters. Figure 1(a) shows the multilayer configuration of $CH_3NH_3SnI_3$-based PSC. The device comprises fluorine-doped tin oxide (FTO) as the front electrode, WSe_2 as the ETL, $CH_3NH_3SnI_3$ as the lead-free light-absorbing perovskite, NiO as the HTL, and gold (Au) as the back contact. Upon illumination, the perovskite layer generates electron–hole pairs by absorbing the incoming photons. Electrons are channeled through ETL to the FTO electrode, while holes are transferred via HTL to the Au contact owing to the photovoltaic effect. Carrier recombination, particularly at interfaces, may cause degradation; thus, optimization of the layer thickness, materials, and interface quality is required. The energy band diagram (Figure 1(b)) demonstrates well-aligned energy levels among the device layers. Conduction band alignment of WSe_2 and $CH_3NH_3SnI_3$ facilitates efficient electron extraction, whereas NiO enables efficient hole extraction and electron blocking, enhancing selectivity and reducing recombination losses.

The rear gold contact, with a work function of 5.10 eV, was suitably positioned to facilitate hole collection from the HTL. The material-dependent input parameters for all the layers, that is, FTO, ETL, absorber, HTL, and interface defects, which constitute the reference dataset, are listed in Tables 1 and 2.

3. RESULTS AND DISCUSSION

3.1. Effect of Layer Thickness on Photovoltaic Efficiency. The thickness of individual layers is highly sensitive to the performance of $CH_3NH_3SnI_3$-based PSCs. To assess their impact, SCAPS-1D simulations were conducted by adjusting one layer at a time while keeping the others fixed. Key performance indicators—open-circuit voltage (V_{oc}), short-circuit current (J_{sc}), fill factor (FF), and PCE—were investigated (Figures 2 and 3).

3.1.1. Influence of Absorber and HTL Thickness. The absorber thickness was elevated from 0.6 to 1.0 µm, keeping the remaining layers fixed at 0.05 µm. As shown in Figure 2(a), V_{oc} declines from 1.0370 to 1.0248 V because of the increased recombination in thick films. J_{sc} marginally improves (Figure 2(b)), with the maximum at 34.981 mA/cm^2 at 1.0 µm,

Table 1. *Material properties for FTO, ETL, absorber, HTL*

Material property	FTO[20]	WSe_2[21]	$CH_3NH_3SnI_3$[22]	NiO[23]
Thickness [nm]	50	150	1000	100
Bandgap, E_g [eV]	3.5	1.19	1.3	3.8
Electron affinity, X [eV]	4.40	4.50	4.17	1.46
Relative dielectric permittivity, ϵ_r	9.00	13.80	8.20	11.7
Effective density of states for conduction band N_C (cm^{-3})	2.2×10^{18}	8.3×10^{18}	1×10^{18}	2.5×10^{20}
Effective density of states for Valence band N_V (cm^{-3})	1.8×10^{19}	1.6×10^{16}	1×10^{18}	2.5×10^{20}
Thermal velocity of electron (cms^{-1})	10^7	10^7	10^7	10^7
Thermal velocity of hole (cms^{-1})	10^7	10^7	10^7	10^7
Mobility of electron, μ_n ($cm^2\ Vs^{-1}$)	20	100	1.6	2.8
Mobility of hole, μ_h ($cm^2\ Vs^{-1}$)	10	500	1.6	2.8
Donor concentration, N_D ($1\ cm^{-3}$)	10^{16}	10^{18}	-	-
Acceptor concentration, N_A ($1\ cm^{-3}$)	-	-	10^{16}	10^{16}
Density of the defect, N_t (cm^{-3})	10^{15}	10^{15}	10^{14}	10^{15}

Table 2. *Parameters for interface defect layers*

Interface	$NiO/CH_3NH_3SnI_3$	$CH_3NH_3SnI_3/WSe_2$
Defect type	Neutral	Neutral
Capture Cross Section: Electrons/holes [cm^2]	1.0×10^{-19}	1.0×10^{-19}
Energetic distribution	Single	Single
Reference for defect energy level	Above the VB maximum	Above the VB maximum
Energy level relative to reference (eV)	0.6	0.6
Total density [cm^{-2}] (integrated over all energies)	1.0×10^{10}	1.0×10^{10}

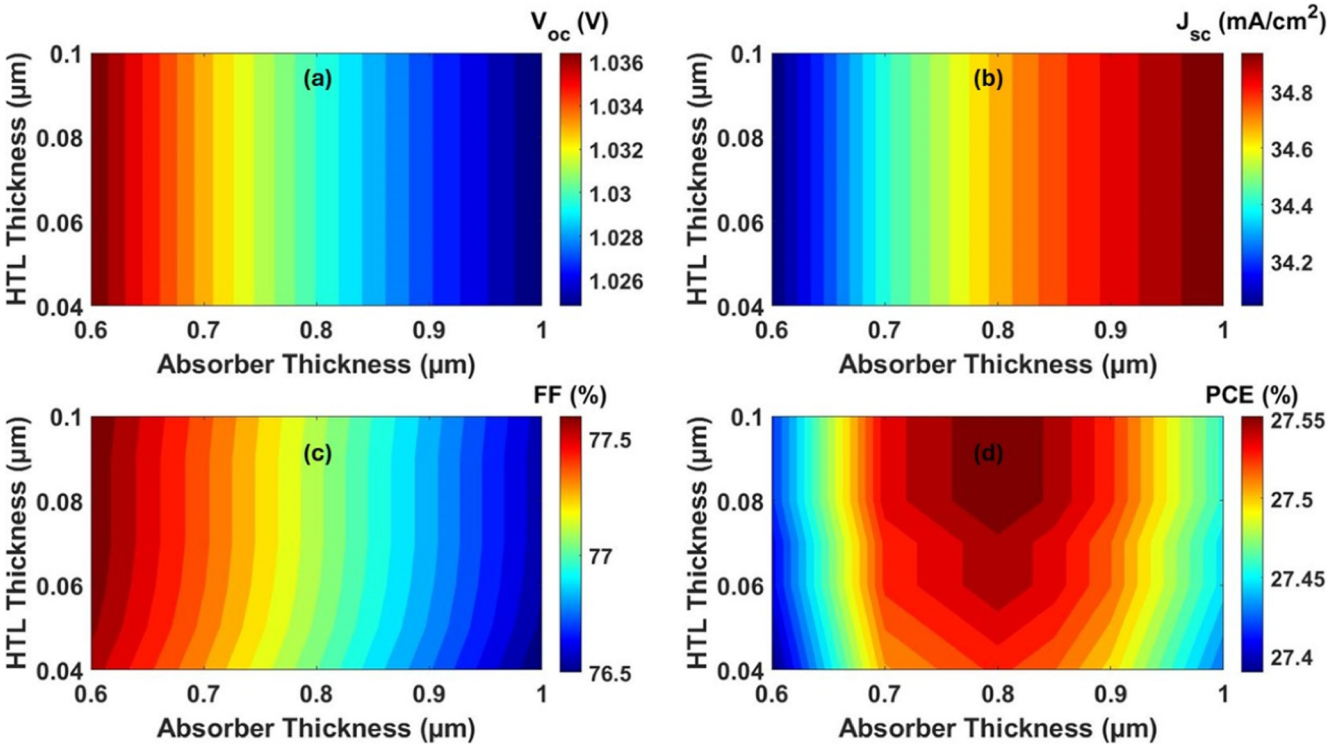


Figure 2. Relation of absorber and HTL thickness on PV performance

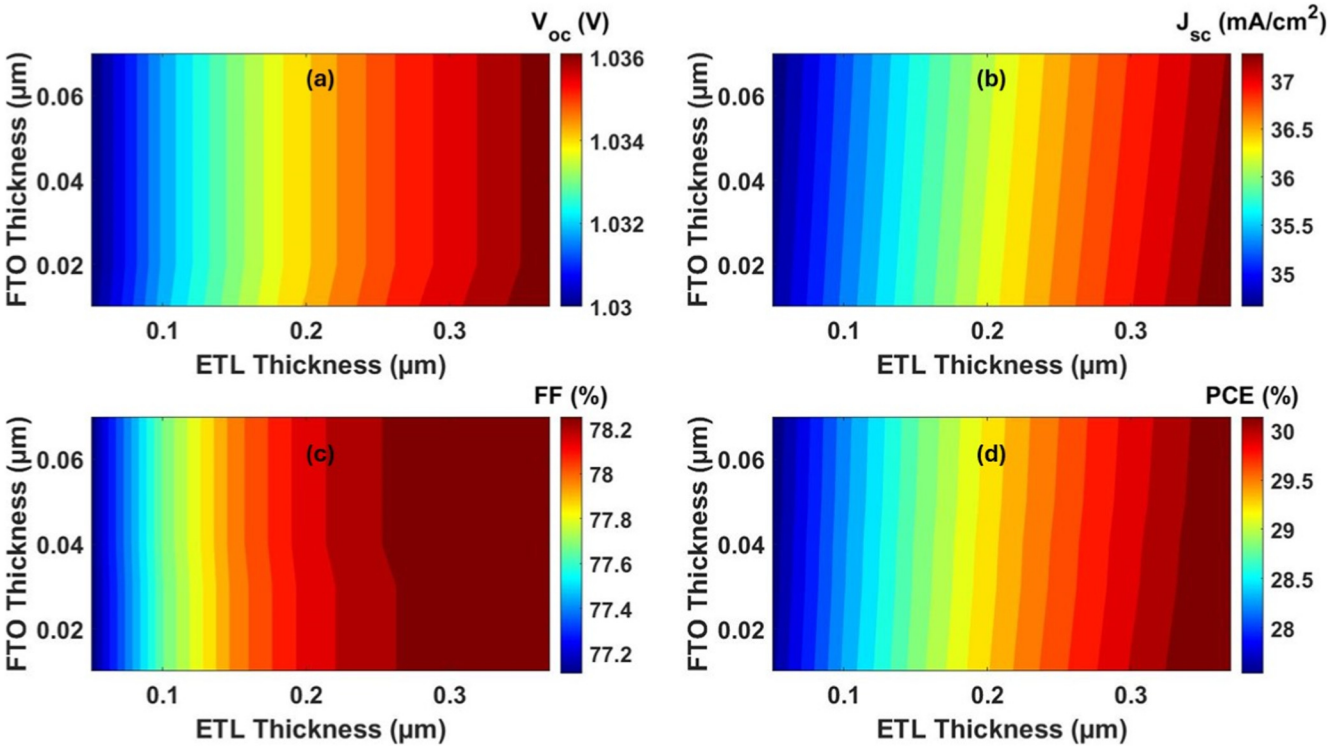


Figure 3. Relation of ETL and FTO thickness on PV performance

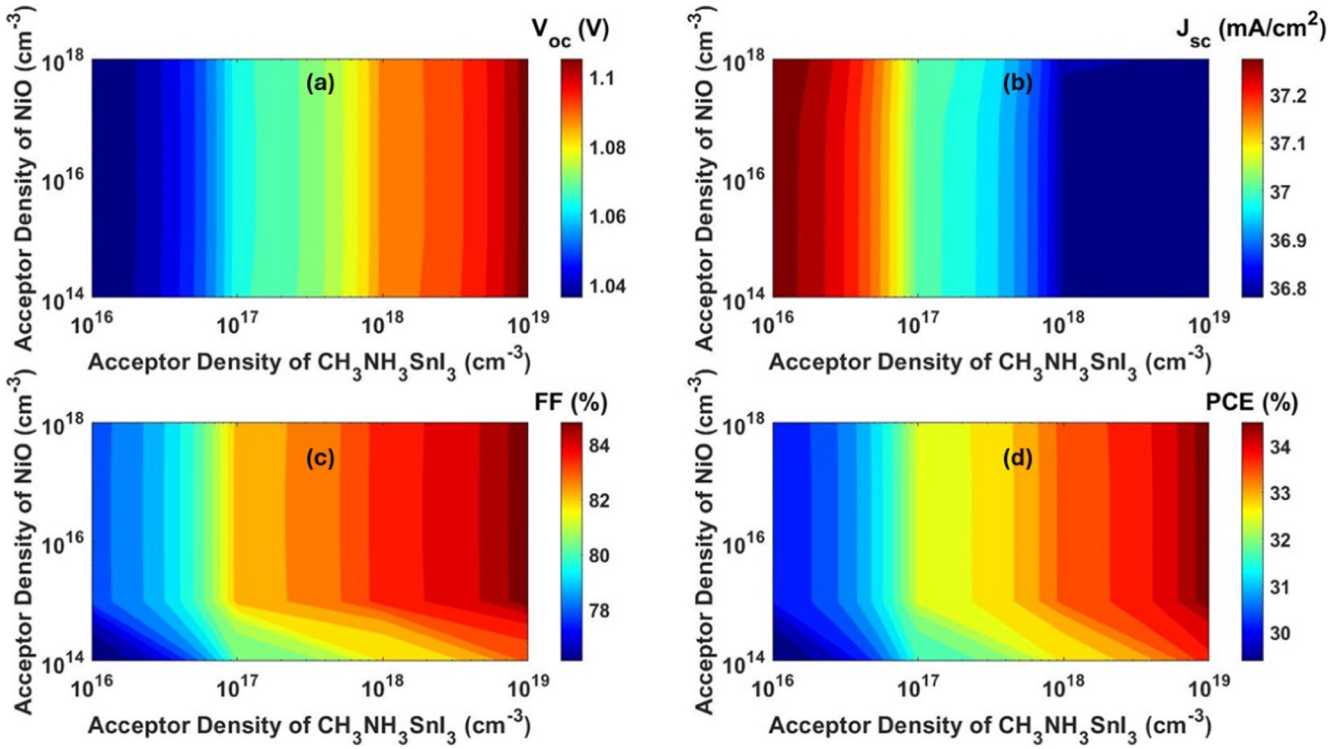


Figure 4. Relation of acceptor density of absorber and HTL on PV performance

but the improvement tails off beyond 0.8 μm. The FF decreases from 77.61% to 76.54% (Figure 2(c)) because of the series resistance effects. Accordingly, PCE achieves a top of 27.54% at 0.8 μm, which is the best balance between V_{oc}, J_{sc}, and FF (Figure 2(d)).

By immobilizing the absorber at 0.8 μm, changing the thickness of HTL from 0.04 to 0.10 μm produced small impacts on V_{oc} and J_{sc} (Figure 2(a) and 2(b)), both of which remained at constant values of ~1.030 V and 34.688 mA/cm^2. However, FF was slightly increased with a peak value of 77.13% at 0.10 μm

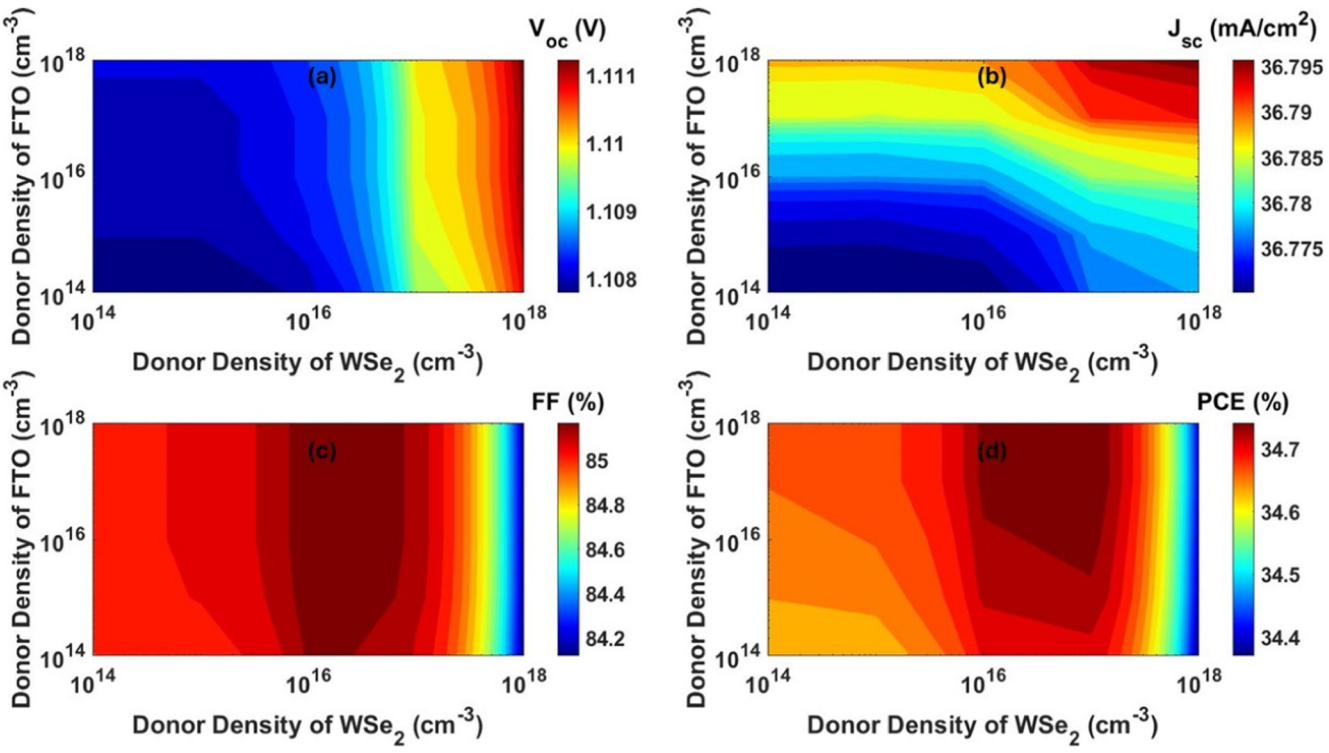


Figure 5. Relation of donor density of ETL and FTO on PV performance

(Figure 2(c)), leading to a highest PCE of 27.56% at 0.08 µm (Figure 2(d)).

3.1.2. Influence of ETL and FTO Thickness. The absorber and HTL were set at their optimal values, and the ETL thickness varied between 0.05 and 0.37 µm. V_{oc} rose from 1.0300 to 1.0361 V (Figure 3(a)), J_{sc} increased to 37.228 mA/cm^2 (Figure 3(b)), and FF increased to 78.31% (Figure 3(c)). The best PCE (30.20%) was achieved with a 0.35 µm ETL thickness (Figure 3(d)).

After optimizing all other layers, FTO thickness reduction to 0.01 µm improved J_{sc} to 37.305 mA/cm^2 (Figure 3(b)) owing to improved light transmission. V_{oc} and FF remained unchanged, resulting in a maximum PCE of 30.27% at 0.01 µm FTO (Figure 3(d)).

3.1.3. Optimized Layer Thickness Design. Simulation data reveal the optimal configuration: absorber thickness 0.8 µm, HTL (NiO) 0.08 µm, ETL (WSe_2) 0.35 µm, and FTO 0.01 µm, yielding V_{oc} = 1.0362 V, J_{sc} = 37.3054 mA/cm^2, FF = 78.30%, and PCE = 30.27%.

3.2. Impact of Doping Concentrations. The doping density is a crucial factor influencing the internal electric field, carrier behavior, and efficiency of PSCs. We investigated the impact of various doping densities on each functional layer of the designed PSC. The contour plots in Figures 4 and 5 show the key PV parameters, such as V_{oc}, J_{sc}, FF, and PCE.

3.2.1. Impact of Absorber and HTL Acceptor Doping Density. As shown in Figure 4, an increase in the absorber's acceptor doping concentration from 10^{16} to 10^{19} cm^{-3} enhances V_{oc} substantially by increasing it from 1.0362 V to 1.1084 V. This trend is attributed to the reduced nonradiative recombination and enhanced built-in potential. The J_{sc} remains relatively unchanged (37.288–36.779 mA/cm^2), indicating that charge generation and light absorption remained efficient. The FF also increases dramatically from 78.12% to 85.23%, and the total PCE increases from 30.18% to 34.74%. The optimum absorber doping concentration is therefore 10^{19} cm^{-3}.

Figure 4 also shows the impact of varying NiO doping concentrations on key performance parameters. V_{oc} and J_{sc} are considerably stable in the range 10^{14}–10^{18} cm^{-3}, but FF increases with greater doping and reaches a value of 85.23%. PCE saturates for greater doping values above 10^{15} cm^{-3}; thus, 10^{15} cm^{-3} is set as the optimum HTL doping concentration.

3.2.2. Impact of Donor Doping Density of ETL and FTO. Figure 5 shows the donor doping effect on the ETL. An increase in the density of donors from 10^{14} to 10^{16} cm^{-3} enhances V_{oc} and FF, resulting in a PCE of 34.74%. With further increased concentration of doping, e.g., 10^{17}–10^{18} cm^{-3}, FF decreases slightly despite minor increases in V_{oc} owing to increased recombination or carrier transport imbalance. Therefore, 10^{16} cm^{-3} is determined as the optimal concentration for doping.

Figure 5 further shows the performance trend with FTO doping levels (10^{14}–10^{18} cm^{-3}). V_{oc}, J_{sc}, and FF exhibited little variation. The change is minimal (< 0.04%), and the values stabilized at 34.74%. Thus, the FTO conductivity is not a limiting factor within the range considered here, and 10^{16} cm^{-3} is satisfactory for efficient performance.

3.2.3. Optimized Doping Configuration. The simulated result proposes the following optimum levels of doping for each layer: absorber at 10^{19} cm^{-3}, HTL at 10^{15} cm^{-3}, ETL at 10^{16} cm^{-3}, and FTO at 10^{16} cm^{-3}. Using this configuration, an

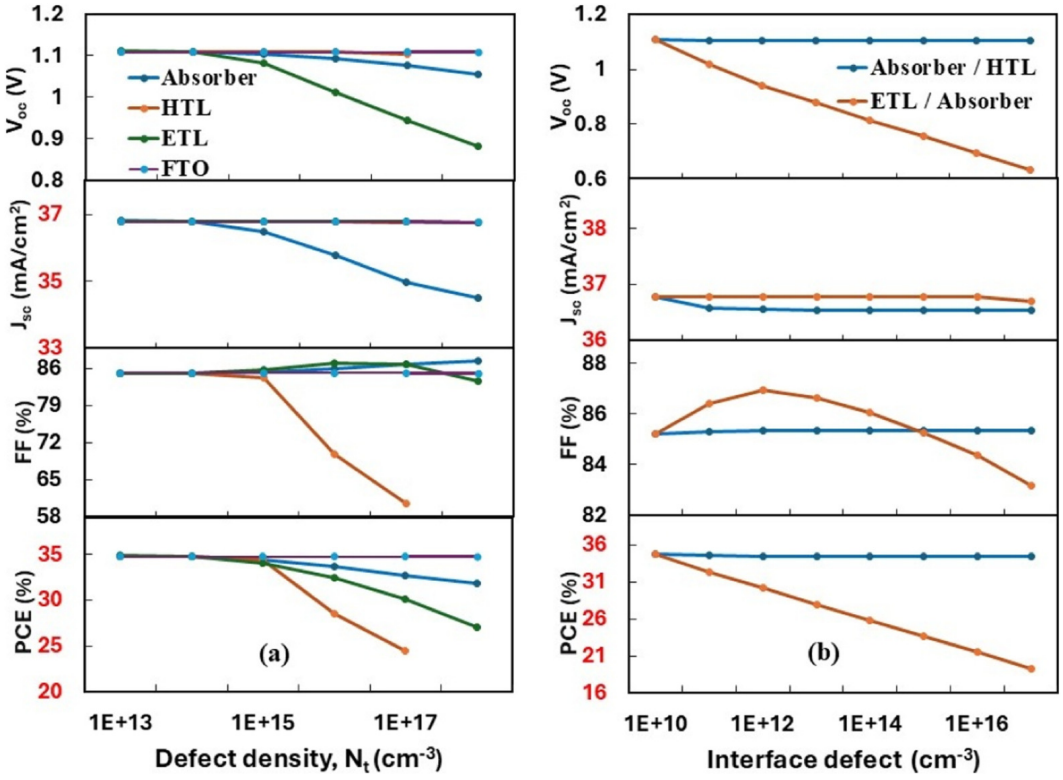


Figure 6. Effect of (a) bulk defect density (N_t) and (b) interface defect density (N_{int}) on PV performance

optimal PCE of 34.74% is achieved, demonstrating the significance of controlled doping in high-efficiency lead-free PSCs.

3.3. Effects of Defect Densities on Photovoltaic Performance. The efficiency and operational stability of $CH_3NH_3SnI_3$-based PSCs are markedly influenced by defect density within both the bulk absorber layer and its interfacial regions. The bulk (N_t) and interface (N_{int}) defect densities act as nonradiative recombination centers and affect the PV performance, as shown in Figure 6.

3.3.1. Defect Density (N_t) Effects. The absorber is highly sensitive to N_t variation between 10^{13} and 10^{18} cm^{-3}. As N_t increases, V_{oc} declines from 1.1089 V to 1.0564 V (−4.74%), and J_{sc} decreases from 36.820 to 34.508 mA/cm^2 (−6.28%). This performance degradation is primarily attributed to elevated trap-assisted recombination, which reduces the quasi-Fermi level splitting. The performance is relatively stable up to $N_t = 10^{14}$ cm^{-3}, although there is degradation that begins dominantly at $N_t \geq 10^{15}$ cm^{-3}. J_{sc} and N_t have a strong negative correlation,

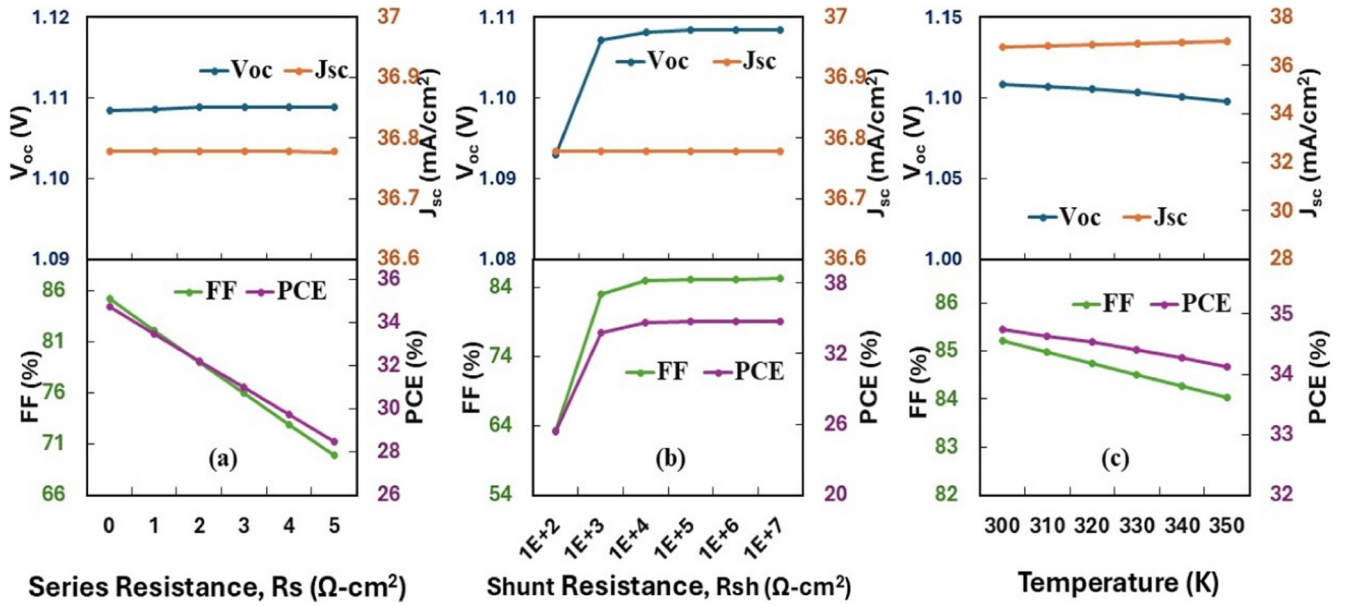


Figure 7. Impact of (a) series resistance (Rs), (b) shunt resistance (Rsh), and (c) temperature on PV performance

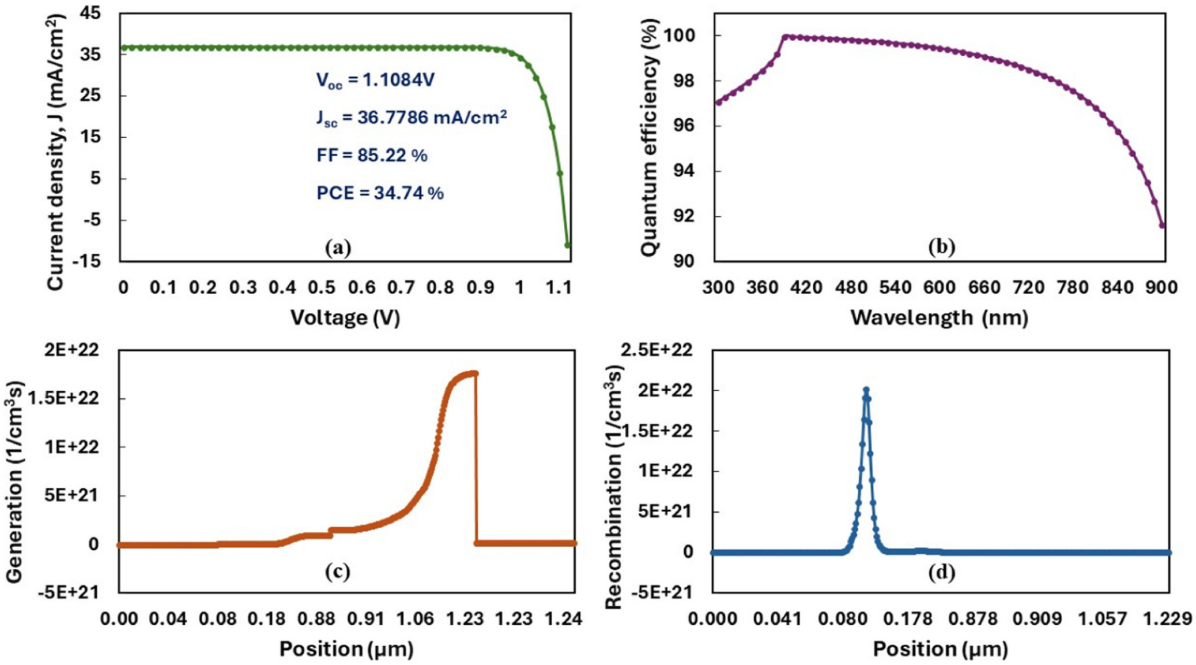


Figure 8. (a) J–V characteristics curve, (b) quantum efficiency curve, (c) generation curve, and (d) recombination curve for the optimized PSC model

Table 3. *Performance comparison of optimized PSC with existing literature*

Structure	V_{OC} (Volt)	J_{sc} (mA/cm^2)	FF (%)	PCE (%)	Ref.
ITO/ ZnO/ CdTe/ $CH_3NH_3SnI_3$/ CuSCN/ Au	0.8150	34.050	76.57	21.24	24
FTO/ NiO/ $CH_3NH_3SnI_3$/ PCBM/ Al	0.9760	33.450	70.33	22.95	25
TCO/ TiO_2/ $CH_3NH_3SnI_3$/ Spiro-OMeTAD/ BC	0.920	31.590	79.99	23.36	26
FTO/ TiO_2/ $CH_3NH_3SnI_3$/ Spiro-OMeTAD/ BC	0.890	35.320	77.45	24.36	27
FTO/ CdS/ $CH_3NH_3SnI_3$/ Cu_2O/ Pt	0.8780	33.40	85.25	25.02	28
TCO/ TiO_2/ $CH_3NH_3SnI_3$/ CuO/ Au	0.930	34.420	83.61	26.63	29
TiO_2/ $CH_3NH_3SnI_3$/ Cu_2O/ Pt	1.020	32.60	84.05	27.95	30
FTO/ TiO_2/ $CH_3NH_3SnI_3$/ Cu_2O/ Pt	0.930	40.14	75.78	28.39	31
FTO/ WSe_2/ $CH_3NH_3SnI_3$/ NiO/ Au	1.1084	36.7788	85.22	34.74	PW

**Note: PW = Present Work*

which indicates carrier collection losses owing to bulk recombination. Devices are supported with < 1% PCE loss if N_t is kept below 10^{15} cm^{-3}.

The HTL shows low sensitivity to N_t values up to 10^{15} cm^{-3}, with V_{oc} and J_{sc} remaining nearly invariant. However, for values greater than this, the FF falls precipitously from 84.35% to 69.96%, with PCE falling from 34.74% to 28.52%. The results show that for V_{oc} and J_{sc} under stable conditions, defect-induced recombination reduces the efficiency of charge transport.

ETL demonstrates significant V_{oc} degradation from 1.1124 V to 0.8815 V (−20.76%) at $N_t = 10^{18}$ cm^{-3}, while J_{sc} is not significantly affected. This implies that defects degrade voltage generation without significantly affecting photon absorption or carrier generation. Optimal N_t must be below 10^{14} cm^{-3} for ETL. The FTO contact was highly stable, with negligible variation in all performance parameters across six orders of magnitude in N_t.

3.3.2. Interface Defect Density (N_{int}) Effects. The NiO/$CH_3NH_3SnI_3$ interface is highly N_{int} resistant up to 10^{17} cm^{-3}, with a PCE loss of less than 0.7%, and V_{oc} and J_{sc} remaining below 2%. Contrastingly, the $CH_3NH_3SnI_3$/WSe_2 interface is sensitive; V_{oc} decreases from 1.1084 V to 0.6351 V (−42.7%) as N_{int} increases from 10^{10} to 10^{17} cm^{-3}, with a PCE loss of 44.2%. This highlights the immediate need for interface defect passivation on the ETL side.

3.4. Influence of Series Resistance, Shunt Resistance, and Temperature on PV Parameters. Figure 7 shows the parasitic resistances and temperature sensitivities of important PV parameters. As shown in Figure 7(a), increasing the series resistance (Rs) induces a sharp drop in both FF and PCE, whereas V_{oc} and J_{sc} are relatively stable. In contrast, a higher shunt resistance (Rsh) significantly enhances the performance by curbing leakage losses, as shown in Figure 7(b). Figure 7(c) indicates that temperature variation optimizes performance under moderate-level conditions and deteriorates under elevated temperatures because of increased recombination.

3.5. Enhanced PV Parameters and Carrier Dynamics Analysis of Optimized PSC. The optimized PSC was investigated in depth using the J–V characteristics, QE, carrier generation, and recombination profiles (Figure 8). The J–V curve (Figure 8(a)) under AM 1.5G illumination demonstrates high device performance, achieving a V_{oc} of 1.1084 V, J_{sc} of 36.7788 mA/cm^2, FF of 85.22%, and PCE of 34.74%, owing to effective charge extraction and good band alignment. The QE spectrum (Figure 8(b)) peaks at 99.95% at a wavelength of 390 nm and gradually decreases to 91.63% at 900 nm, exhibiting strong absorption across the visible range. Carrier generation is confined to the absorber layer (Figure 8(c)), and the minimum generation occurs in the transport layers. Figure 8(d) shows a distinct recombination peak at the absorber near the HTL/absorber interface, suggesting the requirement of passivating defects to maximize efficiency.

The current FTO/WSe_2/$CH_3NH_3SnI_3$/NiO/Au structure performs better than most previously reported $CH_3NH_3SnI_3$-based PSCs because of its extremely high PCE under AM 1.5G illumination (Table 3).

4. CONCLUSIONS

This study presents an in-depth numerical optimization of lead-free $CH_3NH_3SnI_3$-based PSC using SCAPS-1D simulations, focusing on structural, electrical, and defect-related parameters. The optimized structure shows an impressive PCE of 34.74% under AM 1.5G illumination conditions, with a V_{oc} of 1.1084 V, J_{sc} of 36.7788 mA/cm^2, FF of 85.22%, and peak quantum efficiency of 99.95% at 390 nm, surpassing the previously reported values of $CH_3NH_3SnI_3$-based PSCs.

The optimized parameters are as follows: the thickness of the absorber is 0.8 μm, NiO is 0.08 μm, WSe_2 is 0.35 μm, and FTO is 0.01 μm, with corresponding doping concentrations of 10^{19} cm^{-3} (absorber), 10^{15} cm^{-3} (HTL), 10^{16} cm^{-3} (ETL), and 10^{16} cm^{-3} (FTO). These conditions enhance carrier mobility, minimize nonradiative recombination, and promote efficient charge extraction.

Defect analysis showed the necessity of defect passivation because bulk defect densities above 10^{14} cm^{-3} at the absorber and interface defect densities above 10^{17} cm^{-3}—particularly at the $CH_3NH_3SnI_3$/WSe_2 interface—led to a 44.2% loss in PCE. In contrast, the HTL and FTO layers were defect-tolerant, demonstrating their robustness in device operation.

These findings provide a promising design framework for high-efficiency, environmentally benign, and lead-free PSCs, emphasizing the importance of optimized geometries, doping profiles, and defect control in advancing practical perovskite photovoltaic technologies.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Foyzul Karim – *Electrical and Electronic Engineering, Chittagong University of Engineering & Technology (CUET), 4349, Chittagong, Bangladesh;* 0009-0007-8318-0519
Email: u1902169@student.cuet.ac.bd

Md. Habibur Rahman Aslam – *Electrical and Electronic Engineering, Chittagong University of Engineering & Technology (CUET), Bangladesh;* 0009-0008-0751-9645
Email: u1902135@student.cuet.ac.bd

Corresponding Author

Anisul Islam Suva – *Research Mentor, Research Assistant Professor, Institute of Energy Technology, Chittagong University of Engineering & Technology (CUET), Bangladesh;* 0009-0000-7967-7548
Email: anisulislam.me@cuet.ac.bd

ACKNOWLEDGEMENTS

The authors acknowledge the use of the SCAPS-1D simulation software in this study and sincerely thank Dr. Marc Burgelman of Ghent University, Belgium, for his generous provision of access to this valuable tool for the broader scientific community.

CONFLICTS OF INTEREST AND FINANCIAL DISCLOSURE

The authors declare no conflict of interest. This study was conducted independently and did not receive any external funding.

REFERENCES

(1) Olabi, A. G., et al. "Renewable energy systems: Comparisons, challenges and barriers, sustainability indicators, and the contribution to UN sustainable development goals." *International Journal of Thermofluids* 20 (2023): 100498.

(2) Hu, Zhelu, et al. "The current status and development trend of perovskite solar cells." *Engineering* 21 (2023): 15–19.

(3) Kojima, Akihiro, et al. "Organometal halide perovskites as visible-light sensitizers for photovoltaic cells." Journal of the american chemical society 131.17 (2009): 6050–6051.

(4) Kim, Gi-Hwan, and Dong Suk Kim. "Development of perovskite solar cells with >25% conversion efficiency." *Joule* 5.5 (2021): 1033–1035.

(5) Machín, Abniel, and Francisco Márquez. "Advancements in photovoltaic cell materials: silicon, organic, and perovskite solar cells." *Materials* 17.5 (2024): 1165.

(6) Mandadapu, Usha, et al. "Design and simulation of high efficiency tin halide perovskite solar cell." *Int. J. Renew. Energy Res* 7.4 (2017): 1603–1612.

(7) Hima, Abdelkader, and Nacereddine Lakhdar. "Enhancement of efficiency and stability of CH3NH3GeI3 solar cells with CuSbS2." *Optical Materials* 99 (2020): 109607.

(8) Dutta, Sriman, and Saurabh Kumar Pandey. "Device Engineering and Materials Perspective Analysis of Highly Efficient Lead-Free $MASnI_3$ Perovskite Solar Cell." *IEEE Journal of Selected Topics in Quantum Electronics* 30.3: Flexible Optoelectronics (2023): 1–6.

(9) Jayan, K. Deepthi, and Varkey Sebastian. "Comprehensive device modelling and performance analysis of $MASnI_3$ based perovskite solar cells with diverse ETM, HTM and back metal contacts." *Solar Energy* 217 (2021): 40–48.

(10) Ke, Weijun, and Mercouri G. Kanatzidis. "Prospects for low-toxicity lead-free perovskite solar cells." *Nature communications* 10.1 (2019): 965.

(11) Alipour, Hossein, and Abbas Ghadimi. "Optimization of lead-free perovskite solar cells in normal-structure with WO_3 and water-free PEDOT: PSS composite for hole transport layer by SCAPS-1D simulation." *Optical Materials* 120 (2021): 111432.

(12) Mahmoudi, Tahmineh, et al. "Suppression of Sn_{2+}/Sn_{4+} oxidation in tin-based perovskite solar cells with graphene-tin quantum dots composites in active layer." *Nano Energy* 90 (2021): 106495.

(13) Hossain, M. Khalid, et al. "An extensive study on multiple ETL and HTL layers to design and simulation of high-performance lead-free $CsSnCl_3$-based perovskite solar cells." *Scientific Reports* 13.1 (2023): 2521.

(14) Fatima, Qawareer, et al. "A critical review on advancement and challenges in using TiO_2 as electron transport layer for perovskite solar cell." *Materials Today Sustainability* (2024): 100857.

(15) Paul, Indrojit, Abu Rayhan, and M. A. Khan. "Design and Performance Analysis of Photovoltaic Solar Cells Using WSe_2 as an Absorber Layer with SnS_2 Electron Transport Layer." *New Energy Exploitation and Application* 4.1 (2025): 83–101.

(16) Morshed, Md Saklain, Kanak Kanti Bhowmik, and Md Shafiqul Islam. "Investigation of MoS_2/Perovskite/WSe_2 on Si Tandem Structure of Solar Cell." *2022 12th International Conference on Electrical and Computer Engineering (ICECE)*. IEEE, 2022.

(17) Danjumma, Sani Garba, Yakubu Abubakar, and Sahabi Suleiman. "Nickel oxide (NiO) devices and applications: a review." *J. Eng. Res. Technol* 8 (2019): 12–21.

(18) Mitoff, S. P. "Electrical conductivity and thermodynamic equilibrium in nickel oxide." The Journal of Chemical Physics 35.3 (1961): 882-889.

(19) Niemegeers, A.; Burgelman, M.; Decock, K.; Verschraegen, J.; Degrave, S. SCAPS Manual. Ph.D. Thesis, University of Gent, Gent, Belgium, 2014. https://scaps.elis.ugent.be/

(20) Alam, Intekhab, Rahat Mollick, and Md Ali Ashraf. "Numerical simulation of $Cs_2AgBiBr_6$-based perovskite solar cell with ZnO nanorod and P3HT as the charge transport layers." *Physica B: Condensed Matter* 618 (2021): 413187.

(21) Gautam, Sakshi, et al. "Performance analysis of WSe_2 solar cell with Cu_2O hole transport layer by optimization of electrical and optical properties." *Journal of Computational Electronics* 21.6 (2022): 1373–1385.

(22) Mehrabian, Masood, Elham Norouzi Afshar, and Omid Akhavan. "TiO_2 and C_{60} transport nanolayers in optimized Pb-free $CH_3NH_3SnI_3$-based perovskite solar cells." *Materials Science and Engineering: B* 287 (2023): 116146.

(23) Ahmmed, Shamim, et al. "Enhancing the open circuit voltage of the SnS based heterojunction solar cell using NiO HTL." *Solar Energy* 207 (2020): 693–702.

(24) Qasim, Irfan, et al. "Design and numerical investigations of eco-friendly, non-toxic (Au/CuSCN/$CH_3NH_3SnI_3$/CdTe/ZnO/ITO) perovskite solar cell and module." *Solar Energy* 237 (2022): 52–61.

(25) Shamna, M. S., K. S. Nithya, and K. S. Sudheer. "Simulation and optimization of $CH_3NH_3SnI_3$ based inverted perovskite solar cell with NiO as Hole transport material." *Materials Today: Proceedings* 33 (2020): 1246–1251.

(26) Du, Hui-Jing, Wei-Chao Wang, and Jian-Zhuo Zhu. "Device simulation of lead-free $CH_3NH_3SnI_3$ perovskite solar cells with high efficiency." *Chinese Physics B* 25.10 (2016): 108802.

(27) Sahoo, Arpita, Ipsita Mohanty, and Sutanu Mangal. "Effect of acceptor density, thickness and temperature on device performance for tin-based perovskite solar cell." *Materials Today: Proceedings* 62 (2022): 6210–6215.

(28) Arif, Fozia, et al. "Simulation and numerical modeling of high performance $CH_3NH_3SnI_3$ solar cell with cadmium sulfide as electron transport layer by SCAPS-1D." *Results in Optics* 14 (2024): 100595.

(29) Imani, Shayesteh, et al. "Simulation and characterization of $CH_3NH_3SnI_3$-based perovskite solar cells with different Cu-based hole transporting layers." *Applied Physics A* 129.2 (2023): 143.

(30) Omarova, Zhansaya, et al. "In silico investigation of the impact of hole-transport layers on the performance of $CH_3NH_3SnI_3$ perovskite photovoltaic cells." *Crystals* 12.5 (2022): 699.

(31) Patel, Piyush K. "Device simulation of highly efficient eco-friendly $CH_3NH_3SnI_3$ perovskite solar cell." *Scientific reports* 11.1 (2021): 3082.

KFUPM
Journal of Undergraduate Research International