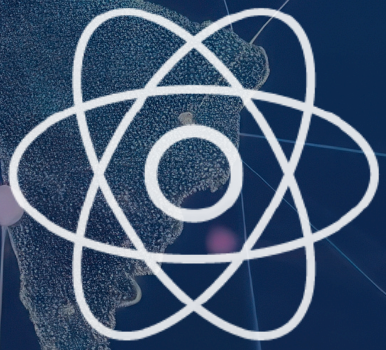
JOURNAL OF ENGINEERING RESEARCH & SCIENCES

JENRS



www.jenrs.com
ISSN: 2831-4085

Volume 5 Issue 1
January 2026

EDITORIAL BOARD

Editor-in-Chief

Editorial Board Members

Editorial

The *Journal of Engineering Research and Sciences (JENRS)* continues its commitment to disseminating high-quality research that addresses contemporary scientific, technological, and societal challenges. The papers featured in this issue represent a diverse collection of studies spanning higher education assessment, sustainable digital infrastructure, advanced optical communications, data-driven business intelligence, ethical implications of artificial intelligence, and machine learning applications in computer vision. Collectively, these contributions demonstrate the growing importance of interdisciplinary research in advancing knowledge, innovation, and sustainable development across multiple domains.

The first paper examines the predictors of academic success among undergraduate students in Mongolia by analyzing data from more than 21,000 graduates across major universities. The study evaluates the relative influence of university admission test scores and prior academic achievement on undergraduate performance. The findings reveal that high school academic performance serves as a stronger predictor of university success than admission test scores, while a combined model incorporating prior academic indicators and third-year performance provides the highest predictive accuracy. The study offers valuable insights for policymakers and educational institutions seeking to improve admission practices and student outcomes through evidence-based decision-making [1].

The second paper addresses the critical challenge of energy efficiency in modern data centers through computational fluid dynamics (CFD)-based analysis of cooling system performance. By investigating dynamically controlled air-cooling units operating under both normal and failure conditions, the research evaluates the effects of control strategies, airflow leakages, and hot-air recirculation on thermal management. The results demonstrate notable energy savings under normal operating conditions and highlight the significant performance degradation caused by leakages during failure scenarios. The study contributes practical design recommendations for enhancing data center sustainability and operational reliability [2].

The third contribution explores the versatility of Nested Antiresonant Nodeless Fiber (NANF) technology for future optical communication networks. Through detailed numerical analysis, the study demonstrates how adjustments to a single structural parameter can transform NANF performance from quasi-single-mode transmission to quasi-two-mode operation. The findings underscore the adaptability of NANF as a platform capable of supporting both conventional communication systems and emerging space-division multiplexing applications. This work represents an important advancement in the development of low-loss, high-capacity optical transmission technologies [3].

The fourth paper presents a cloud-native data architecture designed to address the challenges of fragmented data ecosystems within retail and consumer packaged goods industries. By integrating multiple business data streams into a unified and governed platform, the proposed framework significantly reduces decision-making time, lowers computational costs, and improves operational efficiency. Furthermore, the architecture incorporates FinOps and GreenOps principles, demonstrating measurable environmental and economic benefits while maintaining robust analytical performance. The study provides a scalable and sustainable blueprint for modern enterprise data management [4].

The fifth paper investigates demographic biases embedded within large language models by examining how Meta LLaMA-3.1-8B-Instruct attributes personality and Dark Triad traits to synthetic demographic personas. Through extensive psychometric analysis and statistical evaluation, the research identifies systematic variations in personality trait assignments across gender, race, religion, and regional categories. The findings reveal the presence of latent

psychometric biases within model representations and raise important ethical considerations regarding the deployment of artificial intelligence systems in decision-making environments. The study contributes to the growing discourse on fairness, transparency, and accountability in AI technologies [5].

The final paper provides a comparative evaluation of deep learning and traditional machine learning techniques for binary face classification. By assessing end-to-end convolutional neural networks, hybrid CNN-MLP architectures, and transfer learning approaches based on ResNet50 feature extraction, the study offers a comprehensive examination of performance trade-offs between accuracy and computational complexity. The results demonstrate the effectiveness of both optimized deep learning models and feature-based classical classifiers, providing valuable guidance for researchers and practitioners selecting appropriate methodologies for computer vision applications [6].

In conclusion, the papers published in this issue highlight the breadth and depth of contemporary research addressing challenges in education, information technology, artificial intelligence, communications engineering, and data analytics. The diverse methodologies and innovative findings presented herein contribute meaningful knowledge to their respective fields while emphasizing the importance of sustainable, ethical, and data-driven approaches to future technological development. It is hoped that these studies will inspire further research, foster interdisciplinary collaboration, and support the advancement of science and engineering for the benefit of society.

References:

[1] A. Jargalsaikhan, A. Amartuvshin, "Predicting University Success in Mongolia: The Roles of Admission Tests and Prior Academic Achievement," *Journal of Engineering Research and Sciences*, vol. 5, no. 1, pp. 1, 2026, doi:10.55708/js0501001.

[2] S.A. Surwase, S. Badde, R. Balakrishnan, "CFD Analysis of Data Center Hall Cooling Performance under Normal and Failure Modes with Control Strategies and Airflow Leakages," *Journal of Engineering Research and Sciences*, vol. 5, no. 1, pp. 9, 2026, doi:10.55708/js0501002.

[3] S. Ota, H. Kubota, "Cross-Sectional Structure of Nested Antiresonant Nodeless Fiber for Single-Mode and Few-Mode Transmission," *Journal of Engineering Research and Sciences*, vol. 5, no. 1, pp. 29, 2026, doi:10.55708/js0501003.

[4] P. Chowdhury, "A Cloud-Native Decision Intelligence Architecture for Sustainable CPG Supply Chain Networks," *Journal of Engineering Research and Sciences*, vol. 5, no. 1, pp. 35, 2026, doi:10.55708/js0501004.

[5] N.V. Oikonomou, I. Palaiokrassas, D.V. Oikonomou, ofia P. Chaliasou, N. Rigas, "Demographic Stereotype Elicitation in LLMs through Personality and Dark Triad Trait Attribution," *Journal of Engineering Research and Sciences*, vol. 5, no. 1, pp. 46, 2026, doi:10.55708/js0501005.

[6] N.V. Oikonomou, D.V. Oikonomou, S.P. Chaliasou, N. Rigas, "Binary Image Classification with CNNs, Transfer Learning and Classical Models," *Journal of Engineering Research and Sciences*, vol. 5, no. 1, pp. 66–75, 2026, doi:10.55708/js0501006.

Editor-in-chief

Dr. Jinhua Xiao

JOURNAL OF ENGINEERING RESEARCH AND SCIENCES

Volume 5 Issue 1 **January 2026**

CONTENTS

Received: 30 October 2025, Revised: 21 December 2025, Accepted: 23 December 2025, Online: 12 January 2026

DOI: https://doi.org/10.55708/js0501001

Predicting University Success in Mongolia: The Roles of Admission Tests and Prior Academic Achievement

Ankhbayar Jargalsaikhan [1,2], **Amarzaya Amartuvshin** [*3]

[1]Department of Education study, National University of Mongolia, 14200, Mongolia
[2]Department of Physics and Mathematics Mongolian University of Life Sciences, Ulaanbaatar, 17029, Mongolia
[3]Department of Mathematics, National University of Mongolia, Ulaanbaatar, 14200, Mongolia
Email(s): ankhbayar@muls.edu.mn (A. Jargalsaikhan), amarzaya@smcs.num.edu.mn (A. Amartuvshin)
*Corresponding author: Amarzaya Amartuvshin, Department of Mathematics, National University of Mongolia, 14200, Mongolia, amarzaya@smcs.num.edu.mn

ABSTRACT: This research investigated the factors predicting academic success in Mongolian universities, focusing on university admission test scores and prior academic achievement (high school grade point average). Using data from 21,186 undergraduate students who graduated from major Mongolian universities between 2014 and 2024, the study examined how these factors relate to undergraduate grade point average. Results indicate that admission test scores show a statistically significant, albeit weak, association with undergraduate performance, whereas high school certificate scores demonstrate a stronger predictive effect. A model that includes high school certificate score, admission test score, and third-year grade point average demonstrates the strongest predictive power for final undergraduate grade point average. These findings suggest the need to re-evaluate admission criteria, placing greater emphasis on high school academic performance and reassessing the predictive validity of the national university admission examination. The results highlight the importance of strengthening pre-university education and creating supportive learning environments to enhance students' academic success.

KEYWORDS: Academic preparedness, Academic performance, Predictive validity

1. Introduction

Developing countries, including Mongolia, require a highly qualified workforce, making the quality of higher education crucial for national development. The basis for gaining good quality education at the undergraduate stage depends on the quality of high school level. Knowledge and skills that acquired at the high school level and the earlier levels of education plays important role for the higher involvement and achievement in the undergraduate level of education. This paper is an extended version of the work originally presented at the International Symposium on Computer Science and Educational Technology, ISCSET 2024 [1]. It extended in the sense that the authors added data of students of National University of Mongolia graduated between 2022-2024 and conducted extended analysis using predictor variables.

Higher education enrollment in Mongolia has been increasing steadily since the 1990s, aligning with global trends [2]. However, despite the increasing enrollment rates, the employment rates of graduates have declined, leading to criticism over the high unemployment rate among graduates in the country. Contributing factors include low socio-economic development and limited job opportunities in the labor market. A country's socio-economic development has a significant impact on students' academic achievement. Furthermore, the quality of graduates plays a crucial role in determining the employment rate, which subsequently has a substantial role on the economic development of the country. Higher levels of education among citizens tend to contribute to greater socio-economic development [2,3].

A high school graduate or someone from a higher educational institution who has passed the General University Admission Examination (GUAE) is eligible to apply to Mongolian higher education institutions (HEIs). The GUAE includes a mandatory Mongolian language exam and additional subject-specific tests, selected by the student based on the requirements of the intended university major. In this research, the authors considered GUAE scores and high school grade point average (HGPA) as quantitative measures of student academic preparedness, while the undergraduate grade point average (UGPA) reflected academic performance or achievement at the university level.

The overarching goal of the research is to examine the relationship between the academic achievements of undergraduate students in Mongolian HEIs and their prior educational performance. Specifically, it focused to analyze the relation between students' undergraduate grade point averages (GPA) with their scores on the general university admission examinations, high school graduation certificate scores and other possible scores. To achieve this objective, the authors conducted correlation and regression analyses to explore the relationships among these variables across different student groups. It examined the relationships between undergraduate GPA, entrance examination scores, high school achievement scores, and first-year GPA. Additionally, the study aimed to develop a simple predictive model to estimate students' undergraduate GPA based on these factors.

The specific research objectives of the paper are:

- To assess the correlation between undergraduate GPA and entrance exam scores, high school certificate scores, and GPAs during the periods of undergraduate study.
- To develop a model that accurately predicts undergraduate GPA using the aforementioned variables.

The authors used data from undergraduate students at the National University of Mongolia (NUM), Mongolian University of Life Sciences (MULS), University of Finance and Economics (UFE), and Mongolian State University of Education (MSUE), who graduated between 2014 and 2024. A total of 21,186 students participated in this study. The University of Finance and Economics is a leading private university in the country, while the remaining institutions are public universities.

2. Review of Literature

Defining academic performance or achievement at any level of education and accurately measuring it remain challenging issues that continue to be central focus areas for educational researchers. According to [4], academic achievement is defined as the performance outcomes in intellectual areas studied at educational institutions such as universities. It is a fundamental indicator of intellectual development and is regarded as a critical determinant of individual and societal progress.

Several researchers primarily conceptualize academic achievement as a student's ability to complete specific academic tasks [5,6]. It is commonly evaluated through Grade Point Average (GPA) or other officially documented academic records [7,8]. In this research we use UGPA as a main estimate of academic performance of undergraduate students.

In some cases, scholars have also attempted to assess academic achievement using non-academic outcomes [9]. While both approaches encompass essential dimensions of academic success, they are not entirely interchangeable [6].

Academic preparedness is a pivotal factor in students' academic success. In the context of Australian universities, authors in [10] demonstrated that students with low academic preparedness face greater difficulties in their studies.

Another critical aspect of academic preparedness that directly influences students' academic achievement is their high school internal assessment scores. In [11], it analyzed data from first-year students in New Zealand and concluded that, for social science and humanities subjects, school-based assessments are better predictors of academic achievement at the university level. Conversely, external assessment or entrance examination scores more effectively forecast university performance in disciplines of natural sciences. Similarly, in [12], the authors studied the relationship between secondary education outcomes and academic achievement for educational science students case in Finland. It has shown that, the overall entrance examination results explained 15% of the variance in study success of Finnish Educational Science students.

A study in [13], it also showed the importance of high school-based grades of major subjects for their future study at the university. They used a sample of 113 students graduated from international Baccalaureate (IB) high school and 314 ordinary high school leavers of Holland, determined a predictive validity of grades of high school major subjects for university academic achievements [13]. They targeted to predict academic performance of these students in the first and fourth years of study at the university based on the results of three major subject's assessment results in the last year of the high school using the t-test and multiple correlation analysis. As a result, the GPAs of the first and fourth year of undergraduate study of the students was more relevant to the mean of the scores of three main subjects with highest value, than to the student's high school GPAs. Besides, for alumni who graduated from the IB, the GPA of the beginning year of the undergraduate study and the GPA of the high school had the highest influence on the GPA at the undergraduate graduation.

Using regression analysis in 1998, in [14] it identified a positive but weak correlation between undergraduate students' SAT scores and their academic rankings within the classroom. Similarly, in [15], authors investigated the potential of predicting undergraduate academic success through SAT scores, finding a weak correlation between admission test scores and academic performance in both studies. Notably, the latter study employed multidimensional correlation analysis.

The assessment of entrance examination scores' predictive validity for academic achievement extends beyond the undergraduate level. Numerous studies focus on determining whether scores from globally recognized exams, such as the GRE, can forecast students' academic success at the graduate level.

A meta-analysis in [16], utilized a sample of 1,753 academic records from 85,000 graduate students to explore whether academic achievements are influenced by GRE scores and UGPA scores. As a result, they concluded that these scores are valid predictors of graduate GPA. Further research in [17], as well as in [18], authors examined the relationship between GRE scores and academic performance among master's and doctoral students across various departments. All these studies consistently revealed a weak correlation between GRE scores and graduate academic success.

3. Research Methods and Research Results

3.1 Research methods

The research analyzed data collected from graduates of NUM, MULS, UFE, and MSUE, covering the period from 2018 to 2024. The dataset included academic records of 12,030 students from NUM, 3,015 students from five different schools and faculties within MULS, 853 students from UFE, and 5,288 students from MSUE, making a total of 21,186 undergraduate graduates. During the study, the relationships between known and unknown variables were systematically examined, the form of their correlations was identified, and the expected values of the dependent variables were estimated.

The researchers employed the GUAE score, the average high school certificate score, the first-year GPA (FYGPA) of students, and a moderator variable as predictor variables, with the undergraduate GPA (UGPA) of graduates serving as the dependent variable. During the analysis of the relationships, the scope of the outcome variables was adapted in various ways depending on the specific context. Regression analyses were performed individually for each case, field of study, and university. Data processing was conducted using SPSS version 29 and Microsoft Excel 2019.

Moderating effects are commonly conceptualized as interaction effects, where a moderator variable alters the strength or direction of the relationship between an independent variable and a dependent variable. This interaction may strengthen, weaken, or even reverse the relationship. In regression analysis, moderating effects are typically assessed by incorporating an interaction term—defined as the product of the independent variable and the moderator variable—into the regression model. A statistically significant interaction term indicates the presence of a moderating effect.

Our moderator variable, denoted as 't' in the models, is a composite three-way interaction term. It was constructed by multiplying the standardized z-scores of these three predictor variables (GUAE, HGPA, and FYGPA).

The inclusion of this specific interaction term as a moderator was driven by the theoretical premise that the combined influence of these foundational academic indicators (pre-university preparedness and early university performance) might not be simply additive, but rather interactive. We hypothesized that the predictive utility of one factor (e.g., GUAE scores) for overall university success might depend on the levels of other factors (HGPA and FYGPA). For instance, a student with a lower GUAE score might compensate through strong HGPA and FYGPA, or conversely, the benefits of a high GUAE score might be amplified or diminished depending on subsequent academic performance. This complex interplay aims to capture a more nuanced and holistic understanding of academic success predictors than individual variables alone.

Preliminary analyses revealed normality assumption for UGPA and GUAE results was failed, as indicated by the Kolmogorov-Smirnov test, which produced a significance level of less than 0.001, below the accepted threshold of 0.05. To compare UGPA and GUAE scores across different universities and fields of study, the Kruskal-Wallis test was applied, revealing statistically significant differences between groups. Specifically, UGPA scores among graduates varied significantly across universities ($\chi^2 = 483.1$, $p < 0.05$), while GUAE results also showed significant variation among universities ($\chi^2 = 5380.6$, $p < 0.05$). When the authors analyzed the differences in UGPA and UGEA scores across different graduation years, the results confirmed their statistical significance, with $\chi^2 = 260.6$, $p < 0.05$ for UGPA, and $\chi^2 = 915.9$, $p < 0.05$ for GUAE. Accordingly, suitable regression models were selected to analyze these relationships, and their statistical significance was rigorously assessed. The following section summarizes the models employed in this study.

The study employed several statistical models, notably multiple regression analysis and analysis of variance (ANOVA), to examine the impact of predictor variables such as HGPA, UGEA, and additional moderating factors on UGPA across various contexts.

Student majors were categorized into six broad fields of study: Natural Sciences (NS), Social Sciences and Education (SSE), Humanities (H), Business Studies (BS), Engineering and Technology (ET), and Legal Studies (LS). This categorization was based on the order approved by the Minister of Education regarding the approval of the names of professional fields/programs. For instance, the Natural Sciences (NS) group includes majors such as Physics, Chemistry, Biology, and Mathematics. The Social

Sciences and Education (SS) group comprises disciplines like Sociology, Psychology, Economics, Teaching and Education. Humanities (H) includes fields such as History, Philosophy, and Literature. Business Studies (BS) covers subjects like Accounting, Finance, and Marketing. Engineering and Technology (ET) incorporates Computer Science, Civil Engineering, and Electrical Engineering. Lastly, Legal Studies (LS) includes Law and Criminology.

A graduate here is understood as graduates of undergraduate study. The correlation between the UGPA and GUAE results was determined, and in order to predict the UGPA of the students based on the GUAE scores the authors developed following statistical models as shown in table 1.

Table 1: Models Used in the Study

Models	Dependent variable	Independent variable	Sample
Model 1	UGPA	GUAE score	20868
Model 2	UGPA	HGPA	7229
Model 3	UGPA	HGPA and GUAE	7229
Model 4	UGPA	GPA of years of study	
Model 5	UGPA	HGPA, GPA of 3rd year of study	4825
Model 6	UGPA	GUAE, GPA of 3rd year of study	5678
Model 7	UGPA	HGPA, GUAE, GPA of 3rd year of study	4825
Model 8	UGPA	HGPA, GUAE, GPA of 1st year of study	1667

3.2 Results

We present the overall statistics of the graduates' GPA and their entrance exam scores in the table 2.

Table 2: Descriptive statistics for UGPA and GUAE

Variable	n	Aver-age	Median	mod	s.dev	Variance
UGPA	21186	3.01	3.09	3.1	0.56	0.312
GUAE	21186	612.9	620.2	800	79.74	6358.5

Variable	Skew-ness	Kurtosis	Range	min	max
UGPA	-0.658	0.513	3.16	1	4
GUAE	-0.358	0.016	564	236	800

The correlation coefficient between the GUAE score and graduates' GPA was 0.256, indicating a weak but positive relationship as shown in table 3. Additionally, a significance level with $p<0.05$ for all universities confirms statistical significance of the relationship. The R^2 value of 0.066 suggests that GUAE scores account for 6.6% of the variance in future UGPA. According to the analysis of variance, each regression model predicts graduate GPA based on GUAE scores with statistical significance, and all regression coefficients are significant.

Table 3: Correlation Between UGPA and GUAE Scores, by Academic Fields of Study

Fields	N	R	R^2	b_0 P	b_1 P
NS	5371	0.289	0.083	<0.001	<0.001
	Regression: y=1.743+0.002x				
SS	4481	0.347	0.121	<0.001	<0.001
	Regression: y=1.97+0.002x				
H	3967	0.232	0.054	<0.001	<0.001
	Regression: y=2.14+0.001x				
BS	3363	0.216	0.047	<0.001	<0.001
	Regression: y=1.946+0.002x				
LS	551	0.119	0.014	<0.001	0.005
	Regression: y=2.592+0.001x				
ET	3135	0.215	0.046	<0.001	<0.001
	Regression: y=1.304+0.002x				
Total	20868	0.256	0.066	<0.001	<0.001
	Regression: y=1.895+0.002x				

While the correlation between GUAE scores and UGPA ($r = 0.25$) was statistically significant ($p < 0.05$), likely due to the large sample size, it suggests only a weak practical relationship. This indicates that GUAE scores explain a relatively small proportion of the variance in undergraduate GPA.

The similar picture can be seen with the relationship between graduate's UGPA with HGPA. The UGPA depends on high school grade point average weakly but this relation is statistically significant.

The correlation coefficient between the HGPA score and graduates' GPA of all students is 0.378, indicating a weak but positive relationship as shown table 4. Additionally, a significance level of $p<0.05$ for all fields of studies confirms statistical significance. The R^2 value of 0.143 suggests that GUAE scores account for 14.3% of the variance in future GPA. According to the analysis of variance, each regression model predicts graduate GPA based on HGPA scores with statistical significance, and all regression coefficients are significant.

The next analysis is the correlation of UGPA with HGPA and GUAE by student's academic field of study as shown in table 5.

Table 4: Correlation Between UGPA and HGPA Scores

Fields	N	R	R^2	b_0	b_1
				P	P
NS	2830	0.422	0.178	0.557	<0.001
	Regression: y=0.067+0.032x				
SS	2543	0.336	0.113	<0.001	<0.001
	Regression: y=0.838+0.026x				
H	1413	0.315	0.099	0.001	<0.001
	Regression: y=0.638+0.027x				
BS	387	0.482	0.232	0.302	<0.001
	Regression: y=-0.318+0.037x				
ET	56	0.412	0.17	0.693	0.002
	Regression: y=-0.371+0.036x				
Total	7229	0.378	0.143	<0.001	<0.001
	Regression: y= 0.36+0.03x				

Table 5: Correlation of UGPA With HGPA and GUAE, by Academic Fields of Study

y=UGPA, x = HGPA, z = GUAE score, t = moderator						
Fields	N	R	R2	beta		
				x	z	t
NS	2830	0.491	0.241	0.393	0.256	0.092
	Regression: y = -0.633 + 0.03x + 0.002z + 0.03t					
SS	2543	0.44	0.193	0.284	0.284	0.048
	Regression: y = 0.328+0.022x+0.001z+0.012t					
H	1413	0.387	0.149	0.283	0.172	0.121
	Regression: y = 0.342 + 0.024x + 0.001z + 0.051t					
BS	387	0.502	0.252	0.411	0.147	-0.045
	Regression: y =-0.523+0.031x+0.001z-0.025t					
ET	56	0.428	0.184	0.32	-0.027	-0.166
	Regression: y = 0.507 + 0.028x + 0.001z -0.121t					
Total	7229	0.455	0.207	0.318	0.258	0.052
	Regression: y=-0.064+0.025x+0.001z+0.019t					

Correlation coefficient of UGPA with HGPA and GUAE of all students is 0.455 indicating positive but weaker relations. However, this relation is statistically significant. For students of Business study, The UGPA depends on HGPA and GUAE moderately, while for students of other subjects this relation is weak.

Based on the results presented in Tables 3-5, the authors conclude that GUAE and HGPA scores are not strong predictors of students' UGPA as shown in table 6. In search of other factors that may contribute to a more accurate model to predict UGPA in conjunction with HGPA and GUAE scores, the authors checked the correlations of UGPA with student's yearly GPAs.

Table 6: Correlation UGPA with GPA Scores of Years of Study

Years of study	N	R	R^2	ANOVA	b_0	b_1
				P	P	P
1	14659	0.528	0.279	<0.001	<0.001	<0.001
	Y=1.623+0.492x					
2	8550	0.847	0.718	<0.001	<0.001	<0.001
	Y=1.092+0.657x					
3	10185	0.852	0.726	<0.001	<0.001	<0.001
	Y=1.032+0.673x					
4	10223	0.821	0.674	<0.001	<0.001	<0.001
	Y=1.022+0.654x					

Surprisingly, the first-year GPA was the weakest predictor of graduation GPA, while the second and third-year GPAs proved to be stronger indicators. This contradicts with findings in [1], where the first-year GPA was the most significant predictor of graduation success. The expanded dataset from NUM appears to have influenced these correlations. Consequently, it is important to analyze the correlations among HGPA, GUAE scores, and the GPAs of the first and third years of study to better understand their respective influences on UGPA. This examination can provide insights into how early academic performance and entrance exam results relate to overall university success.

To identify the most effective models for predicting graduate GPA, the authors analyzed the relationship of UGPA with various combinations of HGPA, GUAE and student's first- and third-year's GPAs as shown in table 7, 8 and 9.

Table 7: Correlation of UGPA with HGPA and 3rd Year GPA, by Academic Fields of Study

x = HGPA, z = 3rd year GPA, t = moderator						
Fields	N	R	R2	beta		
				x	z	t
NS	965	0.875	0.765	0.15	0.808	0.019
	Regression: y = -0.2+0.013x+0.694z+0.009t					
SS	2487	0.833	0.693	0.089	0.803	0.045
	Regression: y = 0.558+0.007x+0.647z+0.02t					
H	1373	0.893	0.798	0.077	0.862	0.034
	Regression: y = 0.399+0.007x+0.699z+0.016t					
Total	4825	0.859	0.738	0.090	0.825	0.038
	Regression: y=0.409+0.007x+0.677z+0.017t					

Table 8: Correlation of UGPA with GUAE, 3rd Year GPA, by Academic Fields of Study

x = GUAE score, z = 3rd year GPA, t = moderator						
Fields	N	R	R2	beta		
				x	z	t
NS	965	0.872	0.761	0.135	0.800	0.007
	Regression: y = 0.484+0.001x+0.687z+0.003t					
SS	2487	0.834	0.695	0.103	0.806	0.023
	Regression: y = 0.842+0.001x+0.65z+0.009t					

H	1373	0.893	0.798	0.078	0.872	0.009
	Regression: y = 0.733+0.001x+0.707z+0.003t					
BS	853	0.887	0.787	0.209	0.793	0.033
	Regression: y =0.001+0.002x+0.652z+0.018t					
Total	5678	0.854	0.729	0.078	0.831	0.019
	Regression: y=0.775+0.001x+0.681z+0.008t					

Table 9: Correlation of UGPA with HGPA and GUAE Scores, 3rd-year GPA by Fields of Study

x = HGPA, z = GUAE score, k=3rd year GPA, t = moderator							
Fiel-ds	N	R	R2	beta			
				x	z	k	t
NS	965	0.878	0.771	0.19	0.1	0.79	-0.005
	Regression: y = -0.229+0.01x+0.001z+0.668k-0.002t						
SS	2487	0.836	0.699	0.07	0.09	0.78	-0.017
	Regression: y = 0.458+0.005x+0.001z+0.628k-0.005t						
H	1373	0.895	0.800	0.05	0.06	0.86	0.01
	Regression: y = 0.431+0.004x+0.001z+0.695k+0.004t						
Total	4825	0.861	0.741	0.06	0.08	0.81	-0.003
	Regression: y=0.39+0.005x+0.001z+0.666k-0.001t						

The p value of the ANOVA is less than 0.001 for all cases, which shows the statistical significance of this Model.

The findings indicate that a model combining a student's HGPA, GUAE scores, and 3rd-year GPA is a better predictor of UGPA than other combinations of these factors. It's important to note that all these relationships are strongly positive. This is because academic performance in a student's penultimate year (3rd-year GPA) inherently reflects a more stable and mature pattern of academic engagement and accumulated knowledge. It is temporally closer to the final graduation GPA, thereby capturing current academic aptitude and effort more accurately than earlier indicators such as admission test scores or even first-year GPA, which may reflect initial adjustment phases rather than sustained performance.

From the viewpoint of the practicality, the combination of HGPA, GUAE, and 1st-year GPA also provides a reasonably accurate prediction of student UGPA as shown in table 10.

Table 10: Correlation of UGPA with HGPA, GUAE and First Year GPA, by Academic Fields of Study

x = HGPA, z = GUAE score, k=first-year GPA, t = moderator							
Fields	N	R	R2	beta			
				x	z	k	t
NS	3555	0.858	0.737	0.006	0.197	0.739	0.018
	Regression: y=0.391+0.001x+0.001z+0.605k+0.006t						
SS	8255	0.805	0.648	0.041	0.037	0.778	-0.001
	Regression: y=0.776+0.003x+0.001z+0.657k-0.001t						
H	4887	0.759	0.577	0.075	0.051	0.728	-0.02
	Regression: y=0.35+0.007x+0.001z+0.616k-0.009t						
Total	16667	0.788	0.620	0.039	0.042	0.76	0.001
	Regression: y=0.706+0.003x+0.001z+0.642k-0.008t						

The p value of the ANOVA is less than 0.001 for all cases, which shows the statistical significance of this Model.

The results of the multiple regression analysis demonstrate a strong positive relationship between HGPA, GUAE, first-year GPA, and UGPA for students in the Natural Sciences and Social Sciences. A positive association is also observed for students in the Humanities, although to a lesser extent.

4. Conclusions and Discussions

4.1 Discussions

Although the regression models 1-8 were statistically significant, the observed R^2 values, lower than 12% (Table 3 and 4), indicate that the independent variables explain only a small fraction of the variance in UGPA. This suggests that while these models identify statistically significant relationships, their practical utility for accurately predicting individual student performance remains limited. This underscores the importance of considering the factors identified by these models in shaping student academic performance.

For other Models, the findings are particularly relevant for education policymakers, agencies within the Ministry of Education, and university admissions officers. The analysis reveals that high school certificate scores (HGPA) demonstrate a stronger influence on graduates' GPA compared to GUAE scores. Consequently, a re-evaluation of admissions criteria, with increased emphasis on HGPA, may be warranted.

Our finding that GUAE has a weak predictive validity aligns with the Finnish case in [12].

Strongest relations of graduate's UGPA with GUAE and HGPA of students from the fields Social Sciences. Which doesn't follow the findings in [11].

While Model 7, which incorporates 3rd-year GPA, demonstrated higher predictive power for graduation GPA due to its temporal proximity to the outcome, Model 8, utilizing first-year GPA alongside HGPA and GUAE scores, offers distinct practical advantages. Its strength lies in its early detection value for identifying students at potential academic risk much earlier in their university careers. By providing predictive insights after the first year, Model 8 enables timely and proactive interventions, such as targeted academic advising, tutoring, and support programs. This allows institutions to address emerging academic challenges before they escalate, thereby maximizing the window of opportunity for student support and potentially improving overall retention rates. Furthermore, the availability of first-year GPA data also enhances administrative convenience, facilitating more efficient resource allocation and informed policy decisions regarding student success initiatives. Thus, despite a potentially slightly lower raw predictive accuracy compared to Model 7, Model 8's utility in fostering a proactive and responsive educational environment makes it a highly valuable tool for practical application.

4.2 Conclusions

Based on the findings highlighting the limited predictive power of GUAE scores and the more significant influence of high school academic achievement (HGPA) on undergraduate academic performance, we propose the following recommendations aimed at enhancing student success and educational quality in Mongolia:

I. Reforming University Admissions and Assessment Policiy:

- Revise the content and structure of the GUAE to move beyond mere factual knowledge assessment towards evaluating critical thinking and problem-solving skills.
- Increase the weight placed on high school based assessments (HGPA) and incorporate other supplementary criteria (e.g., portfolios, essays, interviews) into the university admissions process.

II. Strengthening Pre-University Education:

- Promote continuous professional development programs for high school teachers to enhance teaching quality.
- Update pre-university level curricula to ensure better alignment with university needs and requirements, fostering a seamless transition for students.
- Emphasize the development of students' learning strategies and critical thinking skills at the pre-university level.
- Foster greater collaboration and communication between high schools and universities to align expectations and curricula.

III. Adopting International Best Practices:

- Conduct further studies on international best practices in university admissions and pre-university education, adapting relevant strategies to the unique Mongolian context.

While this study provides valuable insights into factors predicting academic success in Mongolian universities, it is important to acknowledge certain limitations that warrant consideration and highlight avenues for future research. Firstly, our analysis was primarily limited to academic variables such as admission test scores and prior academic achievement. We did not incorporate crucial non-academic factors like psychological variables (e.g., motivation, self-efficacy, learning strategies) or socio-economic background (e.g., family income, parental education), which are known to significantly influence student success and could offer a more comprehensive understanding. Secondly, although our study included a large and diverse student population across multiple universities and majors, the findings may still exhibit possible differences across majors and universities depending on specific institutional policies, pedagogical approaches, or disciplinary characteristics that were not disaggregated in this analysis. Future research could explore these variations in greater detail. Finally, due to the correlational nature of our research design, we are unable to infer direct causal relationships between the identified predictors and academic outcomes. Our findings indicate associations and predictive power, but they do not definitively establish that these factors cause subsequent university performance. These limitations, however, open important avenues for more nuanced and experimental future investigations.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A. Amarzaya, J. Ankhbayar, and M. Narantuya, "A study of the predictive validity of Mongolian university admission tests," in *Proceedings of the International Symposium on Computer Science and Educational Technology*, Laubusch, Germany, 2024.

[2] R. A. N. Al-Tameemi, C. Johnson, R. Gitay, A.-S. G. Abdel-Salam, K. Al Hazaa, A. BenSaid, and M. H. Romanowski, "Determinants of poor academic performance among undergraduate students: A systematic literature review," *International Journal of Educational Research Open*, vol. 4, Art. no. 100232, 2023, doi: 10.1016/j.ijedro.2023.100232.

[3] K. Hayat, K. Yaqub, M. A. Aslam, and M. S. Shabbir, "Impact of societal and economic development on academic performance: A literature review," *IRASD Journal of Economics*, vol. 4, no. 1, 2022, doi: 10.52131/joe.2022.0401.0064.

[4] B. Spinath, "Academic achievement," in *International Encyclopedia of the Social and Behavioral Sciences*, Elsevier, 2012, pp. 1–8, doi: 10.1016/B978-0-12-375000-6.00001-X.

[5] M. Maqableh, M. Jaradat, and A. Azzam, "Exploring the

determinants of students' academic performance at university level: The mediating role of internet usage continuance intention," *Education and Information Technologies*, vol. 26, no. 4, pp. 4003–4025, 2021, doi: 10.1007/s10639-021-10453-y.

[6] L. Caixia, Z. A. Bakar, and X. Qianqian, "Self-regulated learning and academic achievement in higher education: A decade systematic review," *International Journal of Research and Innovation in Social Science*, vol. 9, no. 3, pp. 4488–4504, 2025, doi: 10.47772/IJRISS.2025.90300358.

[7] H. Jossberger, S. Brand-Gruwel, M. W. J. van de Wiel, and H. P. A. Boshuizen, "Exploring students' self-regulated learning in vocational education and training," *Vocations and Learning*, vol. 13, no. 1, pp. 131–158, 2020, doi: 10.1007/s12186-019-09232-1.

[8] D. J. Madigan and T. Curran, "Does burnout affect academic achievement? A meta-analysis of over 100,000 students," *Educational Psychology Review*, vol. 33, no. 2, pp. 387–405, 2021, doi: 10.1007/s10648-020-09533-1.

[9] G. Yaxin and Z. M. Noordin, "Study on the effect of peer relationships on academic achievement among college students," *International Journal of Academic Research in Progressive Education and Development*, vol. 13, no. 1, 2024, doi: 10.6007/IJARPED/v13-i1/20780.

[10] C. Baik, R. Naylor, S. Arkoudis, and A. Dabrowski, "Examining the experiences of first-year students with low tertiary admission scores in Australian universities," *Studies in Higher Education*, vol. 44, no. 3, pp. 526–538, 2019, doi: 10.1080/03075079.2017.1383376.

[11] M. Johnston, B. E. Wood, S. Cherrington, S. Boniface, and A. Mortlock, "Representations of disciplinary knowledge in assessment: Associations between high school and university assessments in science, mathematics and the humanities and predictors of success," *Educational Assessment*, vol. 27, no. 4, pp. 301–321, 2022, doi: 10.1080/10627197.2022.2088495.

[12] J. Vulperhorst, C. Lutz, R. de Kleijn, and J. van Tartwijk, "Disentangling the predictive validity of high school grades for academic success in university," *Assessment and Evaluation in Higher Education*, vol. 43, no. 3, pp. 399–414, 2018, doi: 10.1080/02602938.2017.1353586.

[13] J. Kunnari, J. Pursiainen, and H. Muukkonen, "The relationship between secondary education outcomes and academic achievement: A study of Finnish educational sciences students," *Journal of Further and Higher Education*, vol. 47, no. 9, pp. 1155–1168, 2023, doi: 10.1080/0309877X.2023.2222263.

[14] W. G. Bowen and D. Bok, *The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions*, 20th Anniversary ed. Princeton, NJ, USA: Princeton University Press, 1998.

[15] B. Bridgeman, J. Pollack, and N. Burton, "Predicting grades in college courses: A comparison of multiple regression and percent succeeding approaches," *Journal of College Admission*, 2008.

[16] N. R. Kuncel, S. A. Hezlett, and D. S. Ones, "A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance," *Psychological Bulletin*, vol. 127, no. 1, pp. 162–181, 2001, doi: 10.1037/0033-2909.127.1.162.

[17] N. W. Burton and M. Wang, "Predicting long-term success in graduate school: A collaborative validity study," *ETS Research Report Series*, vol. 2005, no. 1, pp. i–61, 2005, doi: 10.1002/j.2333-8504.2005.tb01980.x.

[18] B. Bridgeman, N. Burton, and F. Cline, "Understanding what the numbers mean: A straightforward approach to GRE predictive validity," *ETS Research Report Series*, vol. 2008, no. 2, 2008, doi: 10.1002/j.2333-8504.2008.tb02132.x.

ANKHBAYAR JARGALSAIKHAN is a senior lecturer, Department of Physics and Mathematics, School of Applied Sciences, Mongolian University of Life Science. He did his M.S at the National University of Mongolia in 2011. He is a PhD student at the National University of Mongolia. His research focuses on Mathematical Modeling and Education study.

AMARZAYA AMARTUVSHIN is a Ph.D. and Professor, Department of Mathematics, School of Arts and Sciences, National University of Mongolia. He did his PhD at the Tokyo Metropolitan University in 2003. His research focuses on Differential geometry, Surface theory, Bonnet surfaces, and Mathematics education.

Received: 17 September 2025, Revised: 21 December 2025, Accepted: 23 December 2025, Online: 13 January 2026

DOI: https://doi.org/10.55708/js0501002

CFD Analysis of Data Center Hall Cooling Performance under Normal and Failure Modes with Control Strategies and Airflow Leakages

Sushil Ashok Surwase*, Suribabu Badde, R. Balakrishnan

Engineering Design & Research Centre, L&T Construction, Larsen & Toubro Limited, Chennai, India – 600 089

Email(s): sushil.surwase@lntecc.com (S. A. Surwase), badde.babu@lntecc.com (S. Badde), rbk@lntecc.com (R. Balakrishnan)

*Corresponding author: Sushil Ashok Surwase, India, sushil.surwase@lntecc.com

ABSTRACT: Data centers have become the backbone of an increasingly digitized world, supporting the rapid growth of cloud computing, big data, IoT, 5G, and other emerging IT technologies, with rising demand and innovations in AI and ML reinforcing their significance. Data centers are energy intensive, with data processing and storage accounting for 3 to 4% of global energy consumption, which continues to grow annually. Improving their efficiency is therefore a major industrial challenge, offering substantial cost savings. The modern data center involves an intricate interaction between various mechanical, electrical and control systems. The many possible operating configurations and non-linear interdependencies make it challenging to understand and optimize energy efficiency. In the present study, computational fluid dynamics (CFD) analysis is used to assess the cooling performance of a dynamically controlled data center hall with non-raised floor configuration and hot aisle containment (HAC) strategy. The operation of air-cooling units (ACUs) is dynamically regulated in response to the data hall IT load through an integrated network of sensors and controllers. These controllers modulate ACU fan speed and chilled water flow rates to maintain the IT cabinet inlet air temperature within each ACU's zone of influence and below the specified threshold. This control strategy, informed by real-time temperature and pressure sensor data, ensures desired thermal conditions within the data hall while optimizing overall cooling power consumption. This study focuses on two modes of operation for the purpose of design analysis, i.e., normal mode (NM) and failure mode (FM). Based on CFD simulation results, the present paper highlights the effects of control strategy used for ACUs, cooling airflow leakage, recirculation of hot air on the performance of the data hall cooling design. Different simulation scenarios, which accommodate all possible combinations during NM and FM of operation i.e., with & without control and with & without leakages are evaluated to understand the significance of various design parameters, leading towards the right design. Results show that the control strategy delivers approximately 9.89% energy savings in normal mode, while leakages significantly degrade performance during failure mode.

KEYWORDS: CFD, Control strategy, Data center, Leakage

1. Introduction

The information technology (IT) sector and related technologies are changing at an exponential rate. Data centers have become a key infrastructure to support the rapid development of cloud computing, big data, internet of things (IoT), 5G, Metaverse etc. [1]; thus, data centers serve as the backbone of information in an increasingly digitized world [2]. The demand for data center services has gone up rapidly [3]. The advancements in technologies such as artificial intelligence (AI), machine learning (ML) leading to development of smart appliances, digitalization of transport, buildings and various industries simply reinforce its relevance [4]. Data centers are energy intensive buildings whose size and number have increased in response to the growing demands of a digital economy [5]. Data processing and storage represent 3 to 4% of global energy consumption, and this consumption is significantly growing year on year [6], [7], [8].

Data centers consume a lot of energy. Due to the large energy demands, data center generates a large amount of heat. Thus, cooling is a major aspect of data center design. Enhancing the efficiency of data centers is a significant challenge in the industry, as it can result in significant cost savings [9]. The modern data center involves an intricate interaction between various mechanical, electrical and control systems. The many possible operating configurations and non-linear interdependencies make it challenging to optimize energy efficiency [10].

1.1. Literature Review

The concept of hot and cold aisles was first introduced and formalized in [11], where it was demonstrated that an alternating arrangement of cold and hot aisles significantly improves data center cooling efficiency compared to earlier layouts that lacked floor planning and led to serious reliability problems. One of the earliest applications of CFD to data center cooling was presented in [12], at a time when such studies were scarce. The study proposed an alternative cooling arrangement with ceiling mounted heat exchangers, offering space saving advantages over conventional modular air conditioning unit designs. Using experimentally validated CFD simulations, the authors demonstrated the ability to predict system inlet temperatures and identify hot spots, highlighting the importance of CFD based design for reliable cooling in future high power and high density data center. Their methodology combined with experimental validation, became a benchmark for subsequent academic and industrial CFD studies. The paper [13] on airflow and cooling within data center provides one of the earliest comprehensive formulations of airflow management from a fluid mechanics perspective. The study evaluated how raised floor height, CRAC unit, tile layout and open area, and underfloor obstructions influence plenum airflow, noting that deliberate obstructions such as inclined solid or perforated partitions can beneficially redirect flow. The work also addressed above floor management strategies to prevent hot air recirculation into rack inlets, including sufficient cold air supply, air curtains, partitions, drop ceilings, and ducted racks. The work established the theoretical foundation for many subsequent CFD studies.

In [14], a literature review was conducted to examine corridor isolation and the integration of Building Information Modelling (BIM) with Computational Fluid Dynamics (CFD) in data centers. The authors identified a research gap in studies combining BIM and CFD for corridor isolation. Their findings revealed that hot aisle containment (HAC) provides greater cooling efficiency, lower power usage effectiveness (PUE), and improved working conditions compared to cold aisle containment (CAC). CFD simulations showed that leakage size and position, significantly influence airflow patterns and cooling capacity, while increased supply airflow does not mitigate leakage losses. Cold corridor isolation was found suitable for low-load data centers (up to 5 kW per cabinet) but can reduce personnel comfort, whereas hot aisle isolation is more efficient and preferred in high-load environments (up to 10 kW per cabinet). The study concluded that integrating BIM and CFD offers a reliable approach for designing and optimizing thermal management in data centers.

A comparative CFD analysis of three airflow-organization strategies: underfloor precision supply, inter-column supply, and rack backplane cooling was carried out in [15]. The investigation introduced thermal performance indices such as ASE, ARE, MCRI, RTI, SHI and RHI to evaluate system effectiveness. Results showed that adopting either CAC or HAC increased ASE and reduced SHI values, while the backplane configuration eliminated hot spots without requiring full aisle containment. Optimizing airflow organization scheme, significantly enhances cooling efficiency and energy utilization while minimizing hot spots. The HAC scheme showed the best thermal and energy performance, offering valuable insights into selecting efficient cooling strategies. A similar numerical and experimental study [16] compared HAC and CAC in legacy data centers, focusing on thermal performance and air leakage. The results showed that HAC outperformed CAC at a 15% leakage rate, delivering a 24.9% thermal performance improvement and allowing the supply air temperature in the HAC system to be raised by 2°C. The authors also noted that accurately measuring and validating leakage is challenging and therefore used the IT supply temperature range as a practical indicator of relative leakage effects. Furthermore, [17] conducted a CFD based comparison of raised floor and hard floor configurations with HAC in high density data centers, demonstrating that the raised floor HAC system delivers superior thermal performance over hard floor HAC system. The results showed that adopting a raised floor improves air distribution efficiency by 28% and reduces recirculation ratio by around 40%.

In [18], a combined cooling system that integrates heat storage, waste-heat recovery and different renewable energy sources with conventional air conditioning was modeled. The proposed system reported approximately 16% annual energy savings, an increase in system COP from 3.9 to 4.6, and a reduction of PUE from 1.36 to 1.30. The paper [19] investigated two improvement methods to achieve a uniform temperature distribution in data centers using CFD: (i) installing adjustable underfloor deflectors beneath perforated tiles with varied opening ratios to balance cold-air distribution, and (ii) replacing standard floor grilles near cooling units with fan-floor modules to enhance airflow delivery. Simulation results showed that the deflector method increased airflow to front end cabinets by 18.1% and reduced rear end airflow by 5.1%,

while the fan-floor approach achieved a 4.9% increase and 3.8% reduction, respectively. Both methods improved thermal uniformity and showed that airflow is a key factor that influences cabinet temperature, reducing cabinet maximum outlet temperatures by up to 2.81°C.

The Kao Data case study [20] demonstrated the use of CFD based digital twin modelling (via Future Facilities' 6SigmaDCX) to validate and optimize the indirect evaporative cooling (IEC) design of a high-density, 100% free-cooled sustainable colocation data center. The study conducted both internal and external airflow analyses: internally, the data hall whitespace was evaluated under normal operation and failure mode scenarios to verify cooling efficiency and uniform airflow distribution; externally, a range of wind speeds and directions were simulated to assess the risk of recirculation. Simulation analyses confirmed that the IEC system could maintain target temperatures without mechanical refrigeration, achieving a PUE of approximately 1.2. The study highlighted the value of CFD in enabling design optimization and refining the decision-making process. AKCP [21] illustrates the broader value of CFD to optimize data center airflow and thermal performance. The study emphasizes four key analyses: design airflow analysis to identify hotspots and uneven distribution, "Day One" analysis for early operational optimization, equipment switchover simulation to ensure resilience during cooling unit failures, and leakage analysis to reduce bypass losses and notes that simulation-driven optimization can lower operational costs and carbon footprint.

A new type of ducted HAC system for data center rack cooling was proposed and experimentally evaluated in [22]. The authors studied the effects of different hot duct containment configurations, door states, diffuser types, blanking panel percentages, and airflow volume scenarios on air distribution and cooling performance. They proposed average inlet rack temperature, standard deviation of temperature and temperature difference across rack as practical metrics instead of percentage leakage. Results showed that ducted containment offered performance close to that of full airtight containment but at a lower cost. The paper [23] combined experimental testing with physics based modelling to quantify cold air bypass and determine the optimal DP across aisle containment in data center. The results showed that even with containment, substantial bypass can occur through the rack itself, with bypass airflow reaching up to 20% of the ACU supply. The paper demonstrated that practical mitigation measures such as improved rack design and blocking leakage paths reduced power consumption by up to 8.8%, while optimizing the DP across the cold and hot aisles delivered up to a 16% reduction in power consumption. The authors of [24] conducted a CFD study of a data center with cold aisle containment (CAC), validated by experiments, to assess the impact of leakage. They argued for including realistic fan curves (both server and CRAC fans) in models, noting that fixed flow boundary conditions are a poor approximation in CAC systems. Their findings showed that rack level leakage can cause an inlet temperature rise of about 4°C, and identified a critical leakage threshold of approximately 15%, above which the containment allows so much hot air to recirculate that the benefits of containment are completely lost. In [25], validated CFD modelling was used to assess airflow improvements in a raised floor data center, testing blanking panels, vertical partitions and partial cold aisle enclosure. Partial cold aisle enclosure produced the greatest benefit, allowing a 3°C increase in supply air temperature while maintaining acceptable rack inlet conditions, thereby improving energy efficiency. The study also noted that RTI can be unreliable for identifying bypass or recirculation in complex airflow scenarios. In [26], the effect of CRAC unit placement by comparing units placed in line with the rack rows to units placed perpendicular to the rack rows was investigated. Using RTI, SHI, and RHI as performance indicators, it was found that the perpendicular layout improves airflow uniformity from perforated tiles, reduces hot air recirculation and cold air bypass, and significantly enhances overall cooling performance.

Advanced cooling control strategies for data centers with raised floors and HAC, proposing a decentralized MPC controller design to improve thermal management was examined in [27]. The approach used a dynamic thermal model and zone based control structure to regulate CRAC blower speeds and supply air temperatures. The decentralized control system structure lowers the risk of failure associated with centralized controllers and maintains acceptable rack inlet temperatures while reducing cooling power consumption. In [28], the concept of a smart cooled data center with variable capacity cooling system to allocate cooling dynamically where and when required was proposed. The cooling system consists of adjustable vents, sensors for real time temperature and pressure monitoring and CRAC units with VFD for fans speed and three way valves for chilled water control. Later, [29] implemented and experimentally tested this distributed sensor network coupled with CRAC control in a raised floor data center, reporting a 50% reduction in cooling power consumption and a 25% cost reduction in space and power.

The thermal performance of air cooled data centers under raised floor and non-raised floor configurations was numerically evaluated in [30]. They found that a non-raised floor design with overhead supply and overhead return strategy gives the best thermal performance. They also recommend using overhead supply and return even in raised floor setups, because obstructions (such as pipes and cables) in the underfloor plenum (should be used for only housing pipes and cable) significantly affect air flow

distribution. Importantly, their results showed that using a ceiling return is better than a room level return for both raised floor and non-raised floor design.

The effect of air flow leakage from HAC system on their cooling performance was analyzed by the author of [31]. He evaluated the influence of leakage area, supply air ratio and rack cooling load on the performance of HAC system and found that leakage areas have the largest impact on the performance. An increase in leakage area raises the rate of air leakage, while the nature and location of the leakage paths alter airflow patterns, both of which negatively impact the cooling performance. He also finds that simply increasing supply airflow only reduces temperature of hot air exiting and does not mitigate leakage, and that varying rack cooling loads has little impact on leakage rates. The authors of [32] investigated airflow leakage in CAC and HAC systems. They introduced a Leakage Impact Factor (LIF) to quantify and rank leakage paths such as gaps beneath racks, above racks, and around containment doors. They assumed no leakage through the racks to isolate the effect of containment leakage. Their results showed that leakage beneath racks is the largest contributor to unwanted heat transfer into cold spaces, and they concluded that slight over provisioning of pressure differential is required to mitigate leakage effects. The authors of [33] motivated by experimental data showing air recirculation from the hot aisle to the cold aisle through the gap beneath server cabinets, investigated how tile perforation area, CRAC provisioning, leakage pressure gradients, and CAC affect cooling performance. Results indicate that even small under-cabinet leakage can reduce cooling effectiveness, with the effect being particularly sensitive to under provisioned conditions.

In [34], the authors demonstrated that properly sealed cold aisle containment (CAC) supports higher server heat loads (25.2 kW/cabinet) compared with standard hot aisle/cold aisle layouts (14.6 kW/cabinet). Their research also highlights the critical role of sealing accessories such as grommets and blanking panels, and unused U-slot closures being crucial for improving containment performance. In [35], the authors evaluated the effect of partial aisle containment in both hard floor and raised floor data center layouts under two supply flow rates, 100% and 50%. Their results showed that at a 100% flow rate, the top or side cover fully prevented recirculation in the raised floor configuration, while only reducing it in the hard floor configuration. However, at 50% flow, hard floor setup developed hotspots at the row ends: The side cover improved performance for hard floor layouts and the top cover worsened recirculation. In raised floor configurations partial containment remained beneficial over an open aisle under reduced airflow, with the side cover offering the best results and the top cover providing little improvement.

A containerized data center using CAC with an airside heat exchanger and waterside evaporative water chiller to improve performance in tropical and subtropical regions was demonstrated in [36]. CFD simulations evaluated temperature distribution and thermal performance under varying inlet air temperatures and velocities. Results showed that supply air temperature had minor impact, while inlet air velocity strongly influenced air distribution and thermal management. Overall, the overhead downward flow system with CAC significantly enhanced air distribution and thermal performance in large scale data centers. A comprehensive CFD based analysis of a real data center comprising 208 racks was conducted by authors of [37] to assess how airflow and thermal performance change under varying thermal loads and air supply velocities. They simulated four distinct case studies: two with spatially varying heat loads and two under uniform load, each tested with both maximum and minimum air velocity conditions. Their results showed that while operating CRAC units at maximum airflow can successfully cool the room, it does so at a high energy cost. Consequently, the authors argue that instead of costly CRAC upgrades, sustainability can be improved by optimizing rack layout such as removing selected end of row racks and thereby eliminating hot spots by improving airflow.

According to the authors of [38], the standard k–ε turbulence model is particularly well suited for turbulent flows due to its approach for calculating turbulent viscosity and conductivity. It is also the most extensively validated and commonly implemented model in commercial CFD codes. Furthermore, [39] report that previous studies have demonstrated the k–ε turbulence model outperforms the SST, k–ω, RSM, and RNG k–ε models. The paper [40] focused on improving the accuracy of CFD simulations for data center airflow by comparing different turbulence models, including the widely used k–ε model, Reynolds Stress Model (RSM), and Detached Eddy Simulation (DES). Using a full-scale data center test facility, the CFD results were validated against the experimental measurements. The study found that while the k–ε model captures general flow patterns, it fails to predict low velocity zones present above server racks. The differences in flow fields predicted by the different turbulence models are mostly observed in areas far from the main components of the data center. RSM and DES produced very similar results, with RSM being more computationally efficient and thus recommended for data center airflow modeling. A CFD based study to enhance the design of water cooled data centers using a rear door air to liquid heat exchanger for a 40 kW server rack was conducted in [41]. The simulation, performed with ANSYS and the RNG k-ε turbulence model, showed that inlet air temperature strongly affects rack thermal performance. The rear door liquid cooling system effectively reduced

outlet air from about 40°C to near room temperature of 24°C, efficiently handling the full heat load without additional room cooling.

Finally, [42] states that thermal airflow within data centers exhibits inherently complex behavior with recirculating flow. Considering an inlet velocity of 1 m/s at the supply vents and a rack height of 2.4 m, the resulting Reynolds number is approximately 10^5, indicating turbulent flow.

1.2. Role of CFD in Data Center Design

The CFD simulation plays an important role in data center design [43]:

- *Virtual Design and 3D Analysis:* Minimize rework by testing the design or design changes prior to implementation. Helps to validate and analyze design effectiveness through detailed 3D analysis of air flow and heat transfer in a data center.
- *Performance-Based Analysis:* Identifies issues with data center performance, such as improper air flow (excess or insufficient supply of cold air, bypass, recirculation of hot air, mixing of cold and hot air) during the design phase.
- *What-if Scenarios:* Using predictive results provided by CFD simulation, design and what-if scenarios can be evaluated, minimizing risk of failure such as server overheating and helps in identification of potential failures, which leads to an accurate design.

1.3. Raised Floor Versus Non-Raised Floor Data Hall

Data centers should be designed to operate at an optimal temperature for the highest efficiency of equipment. There are various cooling design approaches such as uncontained room cooling, CAC, HAC, in-row cooling, direct to chip cooling, immersion cooling each having advantages and disadvantages over one another. Irrespective of the cooling approach used, a data hall can be either raised-floor or slab floor (non-raised floor). Researchers continue to debate whether raised floor or non-raised floor configurations provide a better supply air path, with no clear conclusion yet. The thermal performance depends on the cooling conditions and IT environment, and although both approaches reduce loss of cooled air, they differ in practical implementation and operation [17]. The topic of raised floor versus slab floor construction is a topic that often sparks heated discussions in the data center industry as both having advantages and disadvantages over one another. Earlier, almost all the data centers used raised floor. In recent years, non-raised floor data center have gained popularity. The decision to go with raised floor or non-raised floor data centers is now driven by operational objectives, business objectives, business needs and market demands [44].

1.4. Scope of Study

In the present study, CFD analysis is used to assess the cooling performance of a dynamically controlled data hall with practical leakages, non-raised floor configuration and HAC strategy. The operation of air cooling units (ACUs) is dynamically regulated in response to the data hall IT load through an integrated network of sensors and controllers. These controllers modulate ACU fan speed and chilled water flow rates to maintain the IT cabinet inlet air temperature within each ACU's zone of influence and below the specified threshold. This control strategy, informed by real-time temperature and pressure sensor data, ensures stable thermal conditions within the data hall while optimizing overall cooling power consumption.

This study focuses on two modes of operation for the purpose of design analysis, i.e., normal mode (NM) and failure mode (FM). In NM steady state operation, all the ACUs are functional. During FM of operation, a designated number of cooling equipment are offline in the worst-case scenario. Both modes operate with 100% IT loads with uniform distribution of load in data hall.

Based on CFD simulation results, the present paper highlights the effects of control strategy used for ACUs, cooling airflow leakages and recirculation of hot air on the performance of the data hall cooling design. Results from different simulation scenarios, which accommodate all possible combinations during NM and FM of operation i.e., with & without control and with & without leakages are evaluated to understand the significance of various design parameters, leading towards right design. Data center metrics such as ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers) and SLA (Service Level Agreement) compliance are plotted for all simulation scenarios.

1.5. Novelty and Contribution

While numerous prior studies have examined thermal behavior and airflow management in data centers, most rely on highly simplified models of data halls (e.g. limited number of IT Cabinets, idealized geometric layouts or reduced scale representations), overlooking the complexity of real operational environments. Existing literatures predominantly focuses on raised floor configuration and CAC. Only a few investigations addresses non-raised floor facilities with HAC designs, that gained popularity and are increasingly adopted in modern data centers but remain underrepresented and less thoroughly investigated. Moreover, most prior studies typically assume fixed ACU fan speeds and constant chilled water flow rates, failing to explore the dynamic optimization crucial for energy efficiency. To the author's knowledge, studies that do incorporate control often omit details of the control strategy, leaving its impact largely unexplored.

Addressing these gaps, the novelty of the present study lies in its comprehensive, holistic CFD simulation of a dynamically controlled, existing full-scale data hall comprising 308 IT cabinets configured with a non-raised floor and HAC design, thereby offering a level of practical complexity rarely addressed in previous works. A dedicated control strategy for optimizing ACU fan speed and chilled water flow rate is developed, described in detail, and its impact on overall energy consumption is quantified. Finally, unlike prior studies, which typically examine leakage effects, equipment failures, or normal steady state operation in isolation, this research provides a holistic evaluation of data hall performance under both NM and FM, including the practical leakages and active control strategy. By integrating real scale, dynamic control and multi-scenario operation, these contributions advance the state of knowledge by offering practical insights into the design, operational control strategy, and optimization of large-scale modern data centers.

2. Methodology

The air flow and temperature distribution within the data center are governed by the fundamental principles of conservation of mass, momentum, and energy. The full mathematical formulation and derivation of the Navier-Stokes equations (momentum conservation) and the energy conservation equation are omitted here, as these fundamental equations are well established and comprehensively documented in standard CFD books and literature [45], [46], [47], [48]. However, the underlying physics, key assumptions, the selection of the turbulence model, and the details of the computational approach including discretization, solving procedures, and convergence criteria critical to this simulation are discussed in detail in the following sections.

2.1. Governing Equations

The computational model is based on conservation laws, specifically the continuity, momentum, and energy equations, supplemented by the ideal gas equation of state. These equations collectively describe the steady state motion of an incompressible Newtonian fluid (air) and the associated heat transfer by the active information technology (IT) equipment, along with significant auxiliary sources such as uninterruptible power supplies (UPS) and lighting systems. The analysis considers a three-dimensional domain. Key assumptions include modeling the working fluid air as an ideal gas, treating the fluid flow as incompressible and turbulent, and assuming steady-state heat transfer process.

The mass balance (or continuity equation) ensures that mass is conserved within the fluid domain . It dictates that for any control volume within the simulation, the rate at which mass enters must equal the rate at which it leaves, plus any change in mass stored inside. This balance is fundamental for calculating the pressure field in incompressible flows and the density changes in compressible flows, ensuring a physically realistic flow pattern. The momentum balance applies Newton's second law to fluid motion, stating that the net force on a fluid element equals its rate of momentum change. These forces include surface forces like pressure gradients, viscous stresses (internal friction), and body forces like gravity. By solving this balance, CFD determines the fluid's velocity field which along with the pressure field describes the flow dynamics. The energy balance ensures that total energy is conserved by accounting for all energy transfers based on first law of thermodynamics. It relates changes in internal energy to heat transfer (conduction and convection) and the work done by pressure and viscous forces. This equation is solved to determine the temperature distribution throughout the fluid domain, making it essential for simulations involving heat transfer, and fluid property variations caused by temperature changes.

Modeling air as an ideal gas allows the simulation to account for the buoyancy effect by providing the necessary density variation caused by changes in temperature and pressure within the flow field. The equation of state is crucial for solving the system of governing equations, as it provides a way to calculate the density required in the continuity and momentum equations, based on the pressure and temperature calculated by the momentum and energy equations.

A steady state analysis is performed by setting all the time derivatives to zero ($\partial/\partial t = 0$). This choice is justified because the primary objective is to predict the long term, time averaged thermal equilibrium and characteristic mean operating temperatures of the data hall, providing a computationally efficient approach.

As air velocities in the data hall are typically low (with Mach number, Ma < 0.3), incompressible flow assumption is applied. This neglects density variations due to pressure changes, which is a significant simplification used in the continuity and momentum equations.

2.2. Turbulence Modelling

The air flow patterns in data centers are highly complex and recirculating. Based on typical operating conditions such as an air inlet velocity of 1 m/s at supply vent and IT cabinet height of 2.4 m, the Reynolds number is approximately 10^5, clearly indicating a turbulent flow regime [42].

To computationally model the inherently turbulent flow characteristic of large indoor spaces such as data hall, these fundamental principles are typically expressed in their Reynolds-Averaged Navier–Stokes (RANS) form. RANS models decompose each instantaneous flow variable (e.g., velocity, pressure, temperature etc.) into a

time-averaged mean component and a fluctuating component. This time-averaging process introduces the Reynolds stress tensor ($-\rho\overline{u_i'u_j'}$) into the momentum equations, which represents the effective momentum transfer due to turbulent fluctuations.

Since the Reynolds stress terms are unclosed (i.e., they introduce more unknowns than available equations), a turbulence model is required. Specifically, the Reynolds stress tensor, is a symmetric second-order tensor and thus introduces six independent unknown components into the three RANS momentum equations. These six unknowns cannot be determined solely by the existing four RANS equations (continuity and three momentum equations).

A common and robust approach is to employ the Boussinesq turbulence hypothesis, which postulates that the Reynolds stresses are directly proportional to the mean rate of strain tensor, analogous to the relationship between viscous stress and strain for a laminar flow. This hypothesis effectively replaces these six unknowns with a single scalar quantity, the eddy viscosity (μ_t). The major drawback of this hypothesis is that it assumes the turbulent flow is isotropic (the same in all directions), which is often not true for complex engineering flows and cannot accurately predict stresses in highly anisotropic flows where turbulent stress and mean strain are misaligned (e.g., highly swirling or separating flows). Despite this simplification, it works remarkably well for a vast range of engineering applications, including the data center flows.

$$-\rho\,\overline{u_i'u_j'} = 2\mu_t\,\bar{S}_{ij} - \frac{2}{3}\rho\,k\,\delta_{ij} \quad (1)$$

Where, $\bar{S}_{ij}$ is the mean rate of strain tensor, ρ is the fluid density, k is the turbulent kinetic energy , δ_{ij} is the Kronecker delta. With the six Reynolds stress components now expressed in terms of $\bar{S}_{ij}$, and the new variable μ_t (which itself depends on k), the closure problem is reduced from six unknowns to one primary unknown, the eddy viscosity μ_t.

Unlike molecular viscosity (μ), eddy viscosity is a flow property, not a fluid property, which varies throughout the flow field and is computed from averaged flow variable [49], necessitating the use of two-equation models for closure. For instance, the widely-adopted Standard $k-\varepsilon$model solves two auxiliary RANS transport equations, one for the turbulent kinetic energy (k) and another for turbulent kinetic energy dissipation rate (ε) [50]. These two variables are then used to calculate the turbulent viscosity, μ_t , thus achieving closure for the RANS equations.

$$\mu_t = C_\mu * \rho * \frac{k^2}{\varepsilon} \quad (2)$$

While the Standard $k-\varepsilon$ model is utilized for its robustness and wide applicability, it is essential to acknowledge its inherent limitations. The model is known to perform less accurately for flow with strong adverse pressure gradients, substantial boundary layer separation, rotating fluid flows or flow over curved surfaces. This model also assumes a fully turbulent flow regime, an assumption that may not hold across all regions of the airflow within the data center.

The standard $k-\varepsilon$ turbulence model remains the most commonly used approach for CFD simulations of data centers despite its well-known limitations because it offers a uniquely advantageous combination of numerical robustness, computational efficiency, and extensive historical validation. Its exceptional stability makes it unlikely to diverge or crash even on complex or coarse meshes, an attribute that is particularly valuable in data center design where many preliminary configurations must be evaluated rapidly and stability is prioritized over marginal increases in accuracy. The model is also computationally inexpensive, adding only two additional transport equations to the RANS formulation, whereas more advanced models such as the Reynolds stress model (RSM) require solving seven additional equations (six for the Reynolds stress tensor components plus one for epsilon), significantly increasing memory requirements and runtime for the large computational domains typical of data halls. Furthermore, the $k-\varepsilon$ model's empirical constants have been calibrated over decades against a wide range of turbulent flows, and commercial CFD packages have optimized their implementations extensively, reinforcing its position as an industry-standard model [51]. Although it fails to perfectly capture the physics of small, highly anisotropic eddies with high fidelity, it typically provides sufficiently accurate predictions of mean air flow and mean temperature distributions to identify hot spots, characterize recirculation, and support overall decision making. In practice, higher-fidelity alternatives such as the realizable or RNG $k-\varepsilon$, the k-ω SST model, or the RSM are employed when detailed accuracy in near-wall behavior, swirl, turbulence anisotropy or modeling of flow inside a server rack is required, but for large-scale airflow in typical data halls, the standard $k-\varepsilon$ model continues to offer the most effective balance between stability, computational cost, and engineering reliability and practicality.

2.3. Near-Wall Treatment and Wall Functions

The simulation employs the Standard $k-\varepsilon$ (SKE) turbulence model coupled with the wall function approach for near-wall modeling. This coupling is necessary because resolving the steep velocity profiles within the thin viscous sublayer of a turbulent boundary layer requires an extremely fine mesh (viscous sublayer resolving approach, $y^+ \approx 1$), leading to prohibitively high computational cost. Moreover, the SKE model is not

formulated for low Reynolds number wall treatment, its core assumptions specifically the local isotropy of turbulence (turbulence is highly anisotropic near wall) and the validity of the ε - transport equation (The ε - transport equation is unsuitable near the wall because it is derived under the assumption of high local Reynolds numbers. But near the wall, viscous effects dominate, leading to low local Reynolds numbers) break down in the viscous sublayer and buffer region. To achieve computational feasibility, wall functions are used. These are semi-empirical formulas based on the universal law of the wall, effectively bypassing the need to resolve the viscous sublayer with the mesh. The universal law of the wall describes the mean velocity profile of turbulent flow close to the wall, stating that when the mean flow velocity ($\bar{u}$) and distance from the wall (y) are scaled using friction velocity and kinematic viscosity to yield the dimensionless velocity (u^+) and dimensionless distance (y^+), the resulting relationship becomes universally constant and independent of the overall Reynolds number. This universal velocity profile is characterized by a linear relationship in the viscous sublayer , transitioning through a buffer region and a logarithmic relationship in the log layer.

This approach requires the first grid cell center to be located within the turbulent logarithmic region of boundary layer, satisfying the meshing guideline of $30 < y^+ < 300$. This compromise is acceptable for data hall flows, which are generally high Reynolds number and attached.

2.4. Thermal Modelling and Buoyancy

Although a highly accurate thermal model could include the heat source at individual servers, a black box approach was utilized for each IT cabinet in this study, as the inclusion of server level heat details only offered a marginal contribution to overall data center thermal accuracy [52].

Furthermore, buoyancy effects, which are highly significant in thermally stratified air cooled data halls, are incorporated using the Boussinesq approximation. This approximation simplifies the equations by treating the fluid density (ρ) as constant in all equations, except within the gravity (buoyancy) term of the momentum equation. In this term, density is assumed to vary linearly with temperature.

$$\rho = \rho_{ref}\left[1 - \beta\left(T - T_{ref}\right)\right] \quad (3)$$

where ρ_{ref}, β and T_{ref} are the reference density, thermal expansion coefficient and reference temperature.

This simplification is valid provided air properties are constant, the flow is incompressible and exhibits small temperature-induced density variations resulting from a small temperature difference (ΔT). The use of this approximation is strongly justified for this study because data hall operates with a maximum design ΔT of 12 ± 1°C between the supply and return air. This value is well within the widely accepted limit (typically < 20 K) for air [53]. This assumption enables the accurate prediction of temperature and buoyancy driven airflow patterns such as the thermal plume rising from IT equipment without solving the full compressible Navier - Stokes equations, thereby reducing computational costs.

2.5. Numerical Methodology and Convergence

Together, the RANS formulation, the Boussinesq turbulence hypothesis, the Standard $k-\varepsilon$ model, wall function, and the Boussinesq approximation establish a practical and robust framework for predicting the steady state distribution of air velocity, temperature, and pressure within data center spaces.

The set of coupled, non-linear partial differential equations (PDEs) comprising the RANS momentum, mass, energy, and turbulence equations cannot be solved analytically. They are solved using a computational approach. A computational approach based on the Finite Volume Method (FVM) was employed to discretize and solve the equations numerically. In FVM, the physical domain of the data hall is first divided into a finite number of non-overlapping continuous sub-regions, known as control volumes (or the computational mesh). In FVM, the governing PDEs are integrated over each control volume. This integration converts the differential equations into a system of linear algebraic equations that link the value of a variable (e.g., velocity or temperature) at the center of one control volume to the values in its neighboring control volumes. A primary challenge in solving the RANS equations is the inherent coupling between the pressure field and the velocity field, as the mass conservation (continuity) equation does not explicitly contain a term for pressure. This requires a specialized iterative algorithm for solution. For this steady state simulation, the SIMPLE (Semi-Implicit Method for Pressure-Linked Equations) algorithm, or a similar segregated scheme, was utilized. This algorithm iteratively adjusts the pressure and velocity fields until the mass and momentum equations are simultaneously satisfied throughout the entire domain. The system of algebraic equations is solved iteratively until a converged steady state solution is achieved. Convergence is confirmed when the residuals which represent the imbalance in the conservation equations for each control volume have reduced to a specified level. Specifically, the simulation was considered converged when the residuals for pressure, velocity, temperature and turbulence parameters (k and ε) tended to 1. Furthermore, monitoring key performance metrics, such as the return air temperature to cooling unit, ensured that these values stabilized and ceased to change significantly between iterations.

2.6. Validity of CFD Simulation

'DataCenterDesignPro' which is an industry standard data center specific CFD software (Previously recognized as '6SigmaRoom' Release 16.3, which is the latest version at the time of analysis) is used for modeling and CFD simulation. The software is a physics based simulation tool for data center design that utilizes digital twin models. It enables the rapid and accurate creation of digital twins, which serve as virtual representations of existing or planned data centers. These models allow the exploration of multiple design configurations and failure scenarios, supporting the optimization of new data center designs as well as the reevaluation of legacy facilities. By leveraging CFD simulations, the software facilitates the entire design process, from conceptual prototyping to detailed engineering [54].

It includes a comprehensive database of IT equipment and cooling systems, with information collected and verified directly by the respective manufacturers. This capability enables accurate modeling of real data center. The software incorporates the latest cooling technologies and offers greater flexibility in addressing a wide range of design challenges compared to other commercial CFD tools and emerges as the most extensive and feature rich. It is widely recognized as the most accurate tool in the industry for data center design. It is used by several global researchers and data center designers, including Facebook, Microsoft, and IBM, in their data center projects [55], [56], [57], [15], [16], [17], [19], [20] demonstrating the reliability of its CFD results.

2.7. Mesh Independence study

A mesh independence study is a fundamental requirement in all CFD simulations. It is conducted to confirm that the numerical solution is insensitive to further mesh refinement indicating that the discretization error due to the mesh size has been minimized and therefore represents the correct underlying physical behavior. The regions with high gradients are typically assigned a finer mesh. While a finer mesh enhances solution accuracy, it leads to a substantial increase in computational time. By progressively refining the computational grid and comparing key solution variables, it is possible to determine the point at which additional refinement produces negligible changes in the results.

The automatic grid generation feature of the software was used to generate the unstructured hexahedral mesh. Five different meshes were generated, and variables such as ACU return air temperature, cabinet inlet and outlet temperatures, and room temperature were monitored for each mesh size. The mesh containing 7,288,454 cells was found to be optimal, as further refinement caused negligible variation in the selected variable values, and was thus chosen for analysis. This method ensures the accuracy and reliability of the simulation results while avoiding unnecessary computational cost.

The essential components of the data hall (such as ACUs, PDUs, IT Cabinets etc.) can be added either from the software's built-in library or through a neutral data format. This approach ensures the accuracy of simulation, as the mesh independence study of these components is already validated. Achieving grid independence alone does not guarantee simulation accuracy. The correctness of boundary conditions and the choice of turbulence model also significantly affect the simulation results.

3. Data Center Hall

A data hall layout with non-raised floor configuration and HAC strategy as shown in Figure 1, 2 and 3 is used for the CFD analysis. The data hall measures approximately 41 m in length and 20 m in width, with a total floor area of 810 m². The floor to ceiling height is 5.7 m. The number of IT Cabinet are 308 and the corresponding total IT load is 1920 kW (6.23 kW per cabinet). Each cabinet has a height of 2.4 m and a footprint of 0.6 m×1.2 m.

The cabinets are organized into 14 rows in face-to-face and back-to-back configurations, with the back sides of the cabinets facing each other to form HAC of varying sizes. Data hall layout is with caging. The cages are used to create enclosed areas within the data hall which provides an additional layer of security for IT cabinets in a colocation data center. The doors are provided in the containment to allow access to the rear of the cabinets. There are 11 ACUs separated by partition walls and 28 power distribution units (PDUs) which are located inside the data hall.

3.1. CFD Model of the Data Hall with IT Cabinets

The 3D view of the data hall CFD model is shown in Figure 1. The data hall consists of ACUs, PDUs, IT cabinets, cages, hot aisle enclosure, power cables, data cables, lighting, structural beam, structural column, partition walls, walls, temperature and pressure sensors etc. as shown in Figure 2 and 3.

The CFD model of the data hall was fully constructed using the tools and options available in the software 'DataCenterDesignPro', with the CAD layout used as a reference. This CAD layout of the data hall was imported into the software to guide the modeling process, ensuring that the virtual representation accurately reflected the physical layout and arrangement of the data hall.

The ACU is custom built based on the manufacturer's specifications and all other important elements such as PDUs, IT Cabinets etc. are imported from software's built-in library. The other details such as power cables and data cables are included in the model as flow obstructions.

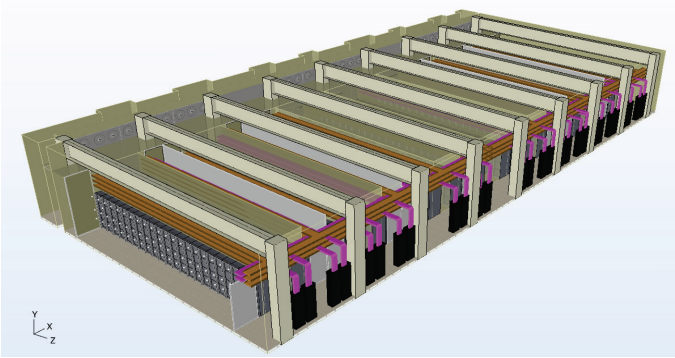


Figure 1: 3D view of data hall CFD model

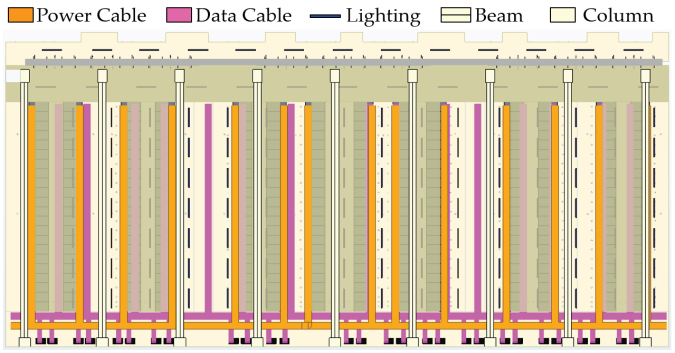


Figure 2: Plan view of data hall CFD model

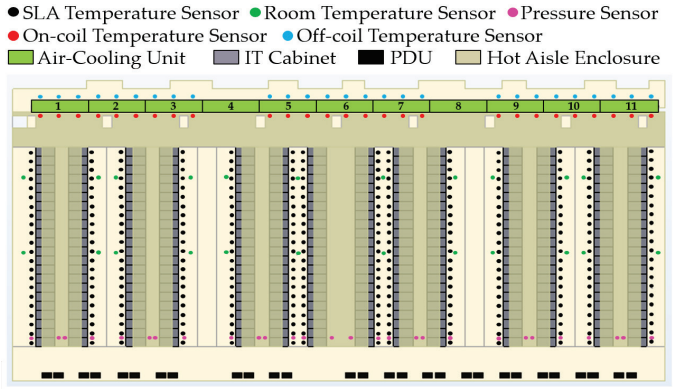


Figure 3: Plan view of data hall CFD model with sensor locations

3.1.1. *Materials*

In data hall CFD simulations, accurately defining material properties is essential because these parameters directly influence heat transfer, and overall thermal performance. Materials with different thermal conductivities, densities, and heat capacities respond differently to temperature loads, affecting how heat is absorbed, stored, and dissipated within the space. Since data halls contain diverse architectural and equipment surfaces that interact with cooling systems, neglecting realistic material properties can lead to significant deviations between simulated and actual thermal conditions. Therefore, incorporating correct material characteristics ensures more reliable predictions of temperature distribution, airflow patterns, and cooling efficiency, ultimately supporting effective thermal management and design optimization.

The walls of the data hall are constructed from cement mortar. The HAC, partition wall and panels are fabricated from polycarbonate. The column, floor and ceiling are composed of concrete. The IT cabinets, ACUs and I-beam are made of mild steel. The key material properties include density, thermal conductivity and specific heat capacity and are listed in Table 1.

Table 1: Material Properties

Material	Density kg/m^3	Thermal Conductivity W/(m.K)	Specific Heat J/(kg.K)
Air	1.19	0.026	1005
Water	997	0.612	4186
Cement Mortar	1200	0.5	850
Polycarbonate	1200	0.19	1300
Concrete	2100	1.37	1000
Mild Steel	7860	63	420

3.1.2. *Sensors*

The temperature and pressure sensors are used to monitor and control the conditions of the data hall. The aim is to ensure not only the sufficient air flow is provided for each IT cabinet but also efficiently cooling them without wastage of energy. The control strategy is developed in such a way that an efficient operation of data hall is achieved while ensuring all SLA temperature requirement are also being met.

The SLA sensors are placed at 0.9m & 1.5m off floor and 0.3m away from IT cabinet air intake side. The top-level sensors are placed at 2.4m off floor level (IT cabinet top level) and 0.3m away from IT cabinet air intake side. The pressure sensors are placed at the far end of each cabinet row. One pressure sensor is placed in the room while another is placed in hot aisle to measure the pressure differential across the IT cabinet. The room temperature sensors are placed in cages to measure room temperature.

One on-coil temperature sensor is placed in front and one off-coil temperature sensor is placed behind each heat exchanger of an ACU. In this case, on-coil temperature is defined as the temperature of hot return air from the conditioned space of data hall (after passing over IT cabinet and removing heat thereby cooling it) and entering the heat exchanger (cooling coil) of an ACU. The off-coil temperature is defined as the temperature of air leaving the heat exchanger of an ACU after getting cooled to the design value by exchanging heat with chilled water supplied by the chillers.

3.2. *Working Principle*

The ACUs supply cold air at the design supply temperature and it fills the data hall room. The cold air then passes through the IT cabinets and takes away heat generated by them. The hot air then gets collected in hot aisle enclosure. The hot return air from IT cabinets then

passes through heat exchangers of ACUs and gets cooled to the design supply temperature and the cycle repeats.

As cold air fills the data hall and hot air is contained in an enclosure, the design approach is called hot aisle containment (HAC) design. The heat exchanger of an ACU is liquid-air type heat exchanger in which heat transfer takes place between hot return air from IT cabinet and chilled water supplied by chillers. The ACU supply air temperature will be higher than off-coil temperature as it includes heat dissipation from fan. The room air temperature will be higher than supply air temperature because it includes heat dissipation from lighting, PDUs and heat gained from unconditioned wall etc. to it.

3.3. Data Center Modes of Operations

Two modes of operation are considered for data center design analysis as follows:

- *Normal Mode (NM)* steady state operation is with all ACUs functional.
- *Failure Mode (FM)* operation where the designated number of ACUs are offline in the worst-case scenario.

4. Air Cooling Unit (ACU)

Data centers require precise thermal management to ensure the reliability and efficiency of IT equipment. The core device responsible for this management is ACU. To efficiently match the dynamic heat load generated by servers, these units rely on sophisticated control strategies. The two primary control mechanisms involve modulating the fan speed and regulating the chilled water flow rate. Fan speed varies using a Variable Frequency Drive (VFD). Simultaneously, the chilled water flow rate through the heat exchangers is regulated by a two-way Pressure Independent Control Valve (PICV).

4.1. ACU Construction

Each ACU has four fans and three heat exchangers (Cooling coils) as shown in Figure 4. The design supply air temperature is 25°C and return air temperature is 37°C. The design ΔT of 12 ± 1°C is to be maintained between supply and return of ACU.

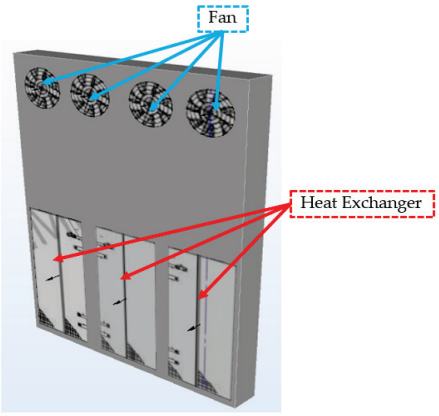


Figure 4: ACU CFD model

4.2. ACU position for NM and FM

Total nine ACUs are available during NM of operation as shown in Figure 5 with green color. The ACU 4 and 8 are for future expansion. The future expansion ACU is replaced by solid obstruction in the model which does not allow the flow through it. The ACUs 6 and 7 are considered offline during FM.

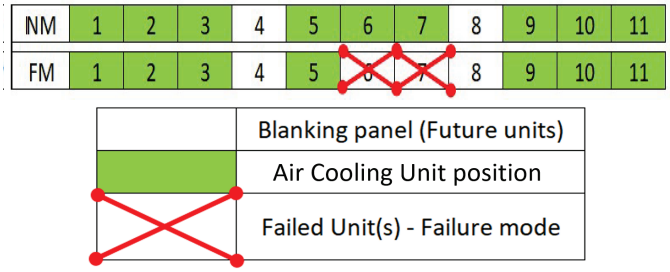


Figure 5: ACU position for NM and FM

4.3. ACU Control Strategy

4.3.1. Fan Control Logic

All the ACU fans should operate at the same speed. Therefore, there is group control associated with fans and a single controller is used to control the fan speed. The fan speed control is a combination of temperature differential (DT) and pressure differential (DP) control. In this case DT is defined as the difference between on-coil temperature and room air temperature. The DP is pressure difference measured at end of the IT cabinet row across IT cabinet.

The DT between average on-coil temperature sensor readings and average room temperature sensor (corresponding cabinet row) readings are calculated. The maximum measured DT will be used for DT control. The fan speed will increase proportionally when DT increases and vice versa. The DT increases when hot return air temperature increases. When return air temperature increases the fan speed increases which supply more cooled air flow which bring down the temperature to the set point of 37°C and vice versa. The minimum fan speed is set at 40% and maximum is set at 100%.

DP at the far end of each cabinet row is measured by DP sensors. When DP is lower than set point of 3 Pa, fan speed will increase based on a reversed proportional curve. When measured DP is lower than 3 Pa (e.g. at fan initial start-up, or failure scenario), DP control will be dominant and fan speed will ramp up to ensure DP set point is reached and when measured DP is higher than 3 Pa, fan control logic will be DT control. The minimum measured DP will be used for DP control.

The DP control ensures that IT cabinets are provided with enough airflow and hot air from hot aisle is not recirculating back to the inlet of IT cabinet by maintaining sufficient DP across it. As DP across cabinet drops below 3 Pa, fan speed increases to increase the ACU air flow thereby increasing pressure on IT cabinet intake side. The

greater the drop below 3 Pa, more will be increase in fan speed.

The control switches between DT control and DP control until steady state is reached which gives highest cooling efficiency without recirculation of hot air. During operation, control will be predominantly DT controlled.

4.3.2. *Cooling Coil Control Logic*

The cooling coil control is set based on off-coil temperature sensor readings. There are three off-coil temperature sensors per ACU. The off-coil temperature is controlled at 25°C. The maximum recorded off-coil temperature out of temperature recorded by three off-coil sensors is used to control coolant (chilled water) flow through the heat exchangers of ACU.

As the off-coil temperature increases coolant flow through the heat exchanger increases to bring down the temperature to the set point of 25°C and vice versa. The chilled water valve minimum opening is set at 10% and maximum opening is set at 100% of design maximum flow rate. The cooling coil control is on individual ACU, there is no group control associated.

5. CFD Simulation

5.1. *Design Data for CFD Simulation*

- The heat dissipation from each fan of an ACU is 2.28 kW. The total fan heat dissipation is 9.10 kW for each ACU.
- The total PDU heat dissipation is 6.72 kW. The heat dissipation from lighting and small power is 9.65 kW.
- Room total heat load is 2018.27 kW during NM and 2000.07 kW during FM, which includes IT load, lighting, small power, PDU, and ACU fan heat dissipation.
- All nine ACUs are modeled with each having 274.20 kW total cooling capacity. Total available cooling capacity is 2467.80 kW during NM and 1919.40 kW during FM.
- There is sufficient cooling capacity available during NM. but the cooling capacity is slightly less than the room total heat load during the FM.

5.2. *Design Considerations*

The following design considerations were made during the NM and FM of operation of data hall:

- Data hall operates with 100% IT load with uniform distribution of load throughout the data hall.
- The radiant heat transfer is negligible compared to the dominant conduction and convection in the data hall.
- Moreover, the room is located within the interior of the building and the influence of solar radiation on its thermal environment is minimal. Therefore, the effects of solar and thermal radiation are neglected.
- The thermal conductivity and specific heat capacity of the fluid were assumed to remain constant, as their variations with temperature and pressure are relatively small.
- The heat dissipation from lighting, occupants, and other minor power sources is included and amounts to 9.65 kW.
- In FM of operation, two ACUs are taken offline namely ACU 4 and ACU 8 as shown in Figure 5.
- Typical small gaps (5%) considered for cabinet leakages.
- A fixed temperature boundary condition is provided for data hall walls.
- For hot aisle enclosure leakage, specified gap size of 0.561 mm with 100% open area is considered.
- The coolant used is chilled water and chillers supply chilled water to ACUs at 22°C.
- The chiller COP is 3.71.
- The rated speed of ACU fan is 1530 rpm.
- The rated fan air flow rate is maximum 4.88 m^3/s. The maximum total air flow rate is 19.5 m^3/s per ACU.
- The ACU heat exchanger effectiveness is 0.80 and cooling capacity is 91.40 kW.
- The chilled water flow rate to each heat exchanger of an ACU can be maximum 3.63 l/s. The maximum chilled water flow rate is 10.90 l/s per ACU.

5.3. *Simulation Scenarios*

For NM and FM of operation, 4 simulations are carried out each as listed in Table 2.

Table 2: NM and FM simulation scenarios

Normal Mode			Failure Mode		
Case No.	Control	Leakage	Case No.	Control	Leakage
1	✓	✓	1	✓	✓
2	✓	×	2	✓	×
3	×	✓	3	×	✓
4	×	×	4	×	×

"✓" *implies* "with", "×" *implies* "without"

6. CFD Simulation Results

After simulation is set up, run and converged, results can be visualized. CFD simulation provides visualization of performance characteristics such as temperature, velocity and pressure that are difficult to capture in the real world. Key results from CFD for evaluating design are ASHRAE compliance plot, top level and SLA sensor analysis, mean inlet temperature of the cabinets, effect of leakages, pressure distribution in space, percentage cooling capacity used, ACU supply and return air

temperature, chilled water temperature curve, fan speed and cooling power curve.

The thermal requirements of IT equipment are typically defined in terms of inlet air temperature of the equipment. According to ASHRAE guidelines [58], the allowable inlet air temperature range is 15 to 32°C, while the recommended operating range is 18 to 27°C.

6.1. Normal Mode (NM)

6.1.1. NM Simulation Results for Case 1

The CFD simulation results of NM Case 1 which is with control and with leakages is discussed in detail:

Figures 6, 7 and 8 shows the temperature distribution in space at height of 0.9m, 1.5m and 2.4m off floor respectively. The SLA sensor readings at 0.9 m and 1.5m off floor level in front of each cabinet is in the range of 25.91 to 26.51°C and 25.88 to 26.52°C respectively. The SLA sensor reading meets the design requirement. The top-level sensor reading at 2.4 m off floor level in front of each cabinet is in the range of 25.84 to 26.72°C.

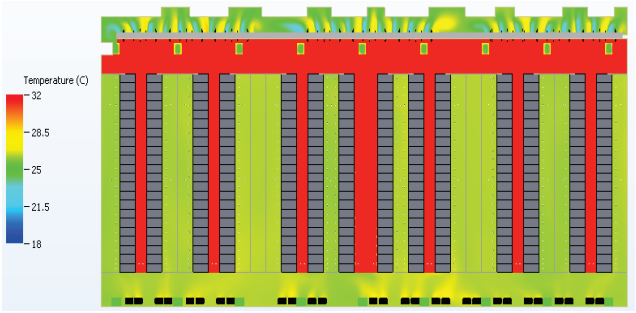


Figure 6: Temperature distribution in space at height of 0.9m off floor

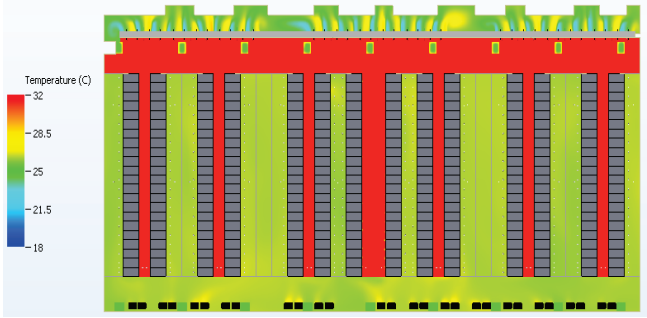


Figure 7: Temperature distribution in space at height of 1.5m off floor

Figure 10 shows the mean inlet air temperature of the IT cabinet. The mean air temperature measured at IT cabinet intake side is in the range 25.91 to 27.08°C which is slightly above the ASHRAE recommended temperature range (18 to 27°C). Figure 11 shows the temperature plot across the IT cabinets (IT cabinets hidden). The temperature plot across cabinets indicates leakage of hot air from hot aisle into the cabinet inlet through cabinet's typical small gaps, resulting in increased peak inlet air temperature for five IT cabinets. Therefore, in case higher peak inlet air temperature is recorded due to leakages, SLA sensor readings will take precedence over the ASHRAE readings to check for compliance.

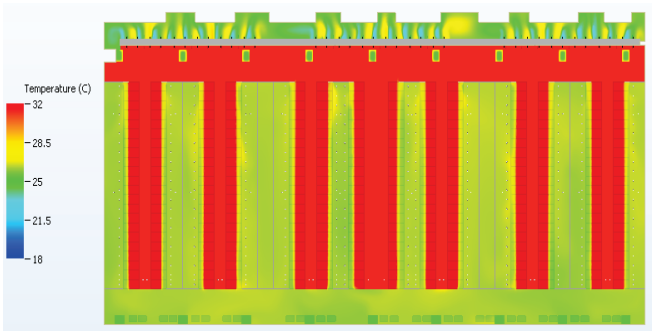


Figure 8: Temperature distribution in space at height of 2.4m off floor

Figure 9 shows the ASHRAE temperature compliance plot for NM Case 1. The ASHRAE compliance temperature is the peak inlet air temperature of the IT cabinet. The peak air temperature measured at IT cabinet intake side is in the range of 25.97 to 30.11°C. All the IT cabinets except five are having peak inlet air temperature in the range of 18 to 27°C, which comply with ASHRAE recommended temperature range.

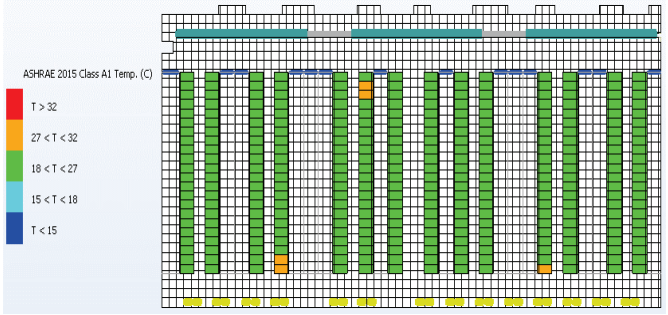


Figure 9: ASHRAE temperature compliance plot for NM Case 1

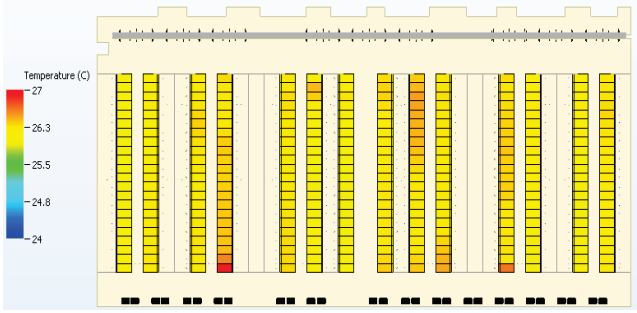


Figure 10: Mean inlet air temperature of the IT cabinet

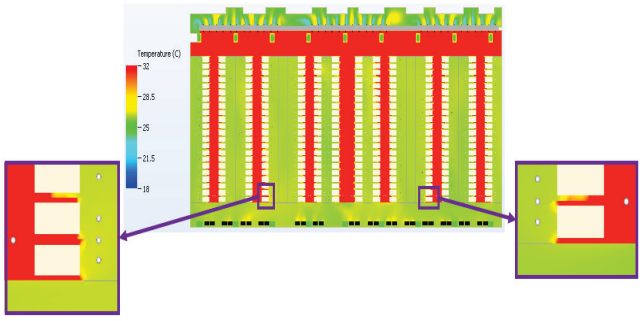


Figure 11: Temperature plot across the IT cabinets (IT cabinets hidden)

Figure 12 shows pressure distribution in space at 2.4m off floor (IT cabinet height level) with differential pressure across the IT cabinet at the far end of each cabinet row. Figure 12 also shows the fan speed and corresponding air flow rate from an ACU. During normal steady state

operation of the data hall, fans operate at 1395 rpm with air flow rate of 17.78 m³/s. All fans operate at the same speed and air flow rate, as there is a group control associated with ACU fans and a single controller is used to control the speed. The DP varies from 2.94 to 24.02 Pa. Figure 13 shows the cooling capacity utilization of an ACU. The cooling capacity of ACUs varies from 203.11 to 236.63 kW. Only 74.07 to 86.30% of the total available cooling capacity of 274.2 kW is utilized by the ACUs.

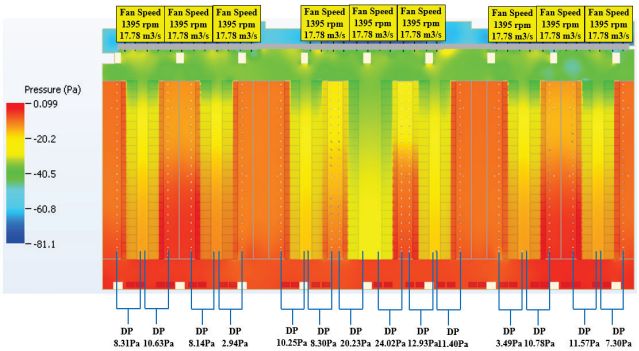


Figure 12: Pressure distribution in space at 2.4m off floor

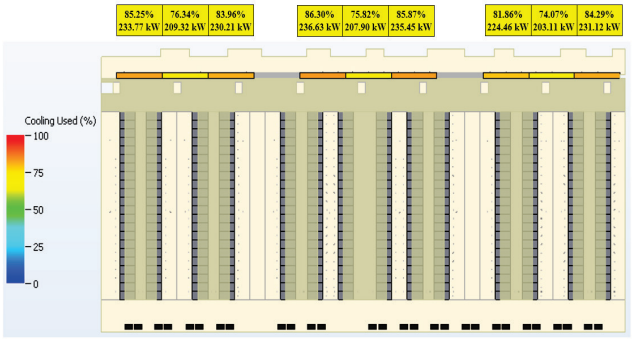


Figure 13: ACU cooling load distribution

Figure 14 shows the average ACU on-coil and off-coil temperature. The ACU on-coil temperature is the average of the temperature recorded by on-coil sensor of heat exchangers of an ACU. Similarly, the ACU off-coil temperature is the average of the temperature recorded by off-coil sensor of heat exchangers of an ACU. The simulation result showed an average of 25.60°C supply and 35.92°C return which is close to the design values. This indicates that the control logic is functioning well.

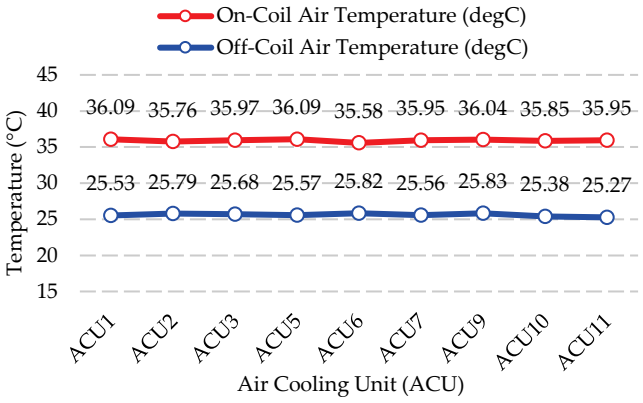


Figure 14: ACU average on-coil and off-coil temperature

6.1.2. *NM Simulation Results for Case 2, 3 and 4*

Figure 15 shows the ASHRAE temperature compliance plot for NM Case 2. The peak air temperature measured at IT cabinet intake side is in the range of 26.10 to 27.05°C. All the IT cabinets except one are having peak inlet air temperature in the range of 18 to 27°C, which comply with ASHRAE recommended temperature range.

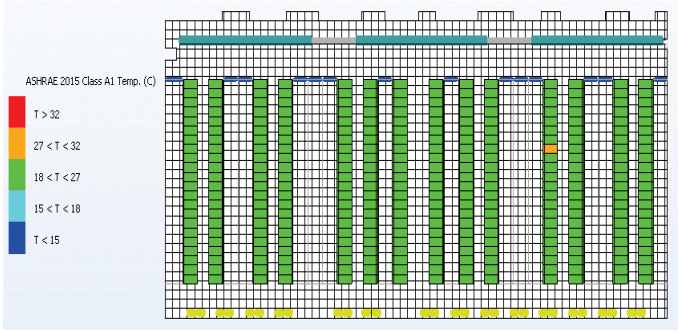


Figure 15: ASHRAE temperature compliance plot for NM Case 2

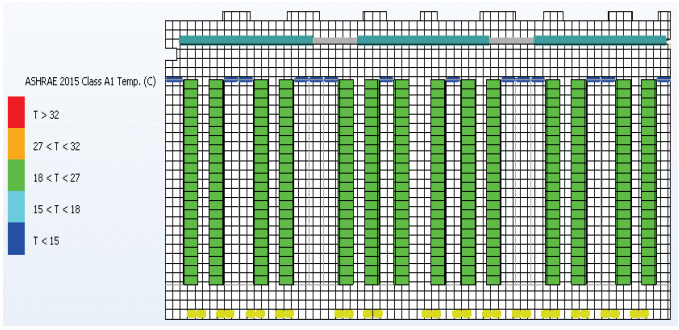


Figure 16: ASHRAE temperature compliance plot for NM Case 3

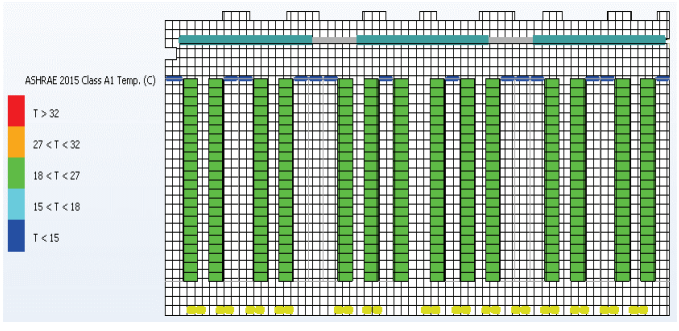


Figure 17: ASHRAE temperature compliance plot for NM Case 4

Figure 16 and 17 shows the ASHRAE temperature compliance plot for NM Case 3 and 4 respectively. The peak air temperature measured at IT cabinet intake side is in the range of 25.10 to 26.36°C and 25.12 to 25.96°C for Case 3 and 4 respectively. All the IT cabinets are having peak inlet air temperature in the range of 18 to 27°C for both Case 3 and 4, which comply with ASHRAE recommended temperature range.

6.1.3. *NM Simulation Results Summary*

Table 3 summarizes the NM simulation results. From CFD simulation result it is observed that:

Table 3: Normal Mode simulation results summary

Normal Mode				
Case	1	2	3	4
Control	✓	✓	✗	✗
Leakages	✓	✗	✓	✗
Simulation Results				
Fan Speed Controller Temp. Input	9.99°C	11.00°C	-	-
Fan Speed Group Controller Output	85.30%	69.80%	-	-
Fan Speed	1395 rpm	1253 rpm	1530 rpm	1530 rpm
Chilled Water Controller Output	62.39 to 85.47%	60.24 to 81.01%	-	-
Pressure Difference	2.94 to 24.02 Pa	11.39 to 30.83 Pa	12.33 to 38.97 Pa	69.00 to 96.98 Pa
Maximum On-Coil Temperature	36.79°C	37.48°C	35.50°C	34.54°C
Minimum On-Coil Temperature	33.94°C	36.70°C	33.19°C	33.97°C
Maximum Off-Coil Temperature	26.02°C	26.02°C	25.49°C	24.22°C
Minimum Off-Coil Temperature	24.68°C	24.76°C	24.13°C	25.31°C
Total Cooling Power (kW)	203.11 to 236.63	206.64 to 232.21	202.78 to 232.03	206.40 to 230.50
Coolant Temperature Out (Average)	27.20 to 29.78°C	27.41 to 29.90°C	25.99 to 27.86°C	26.07 to 27.64°C
Cabinet Maximum Temperature In	25.97 to 30.11°C	26.10 to 27.05°C	25.10 to 26.36°C	25.12 to 25.96°C
Cabinet Mean Temperature In	25.91 to 27.08°C	25.99 to 26.62°C	25.05 to 25.68°C	25.05 to 25.64°C
Room Temperature	25.94 to 26.49°C	25.98 to 26.71°C	24.98 to 25.73°C	24.99 to 25.66°C
Top-level Temperature Sensor at 2.4m	25.84 to 26.72°C	25.90 to 26.95°C	25.07 to 26.03°C	25.06 to 26.01°C
SLA Temperature Sensor at 1.5m	25.88 to 26.52°C	25.93 to 26.60°C	25.04 to 25.67°C	25.03 to 25.62°C
SLA Temperature Sensor at 0.9m	25.91 to 26.51°C	25.98 to 26.59°C	25.01 to 25.66°C	25.01 to 25.59°C

The leakage causes recirculation of hot air from hot aisle back into the cabinet inlet through cabinets typical small gaps, resulting in increased peak inlet air temperature for cabinets. For case without control logic (Case 3&4), all IT cabinets are having peak inlet air temperature in the range of 18 to 27°C, and SLA sensor readings at 0.9 m and 1.5 m off floor in front of each cabinet are less than 27°C, which met the design requirement.

For cases with control logic (Case1&2), ASHRAE compliance is not met due to recirculation of hot air because of leakages. But SLA sensor readings at 0.9 m and 1.5 m off floor in front of each cabinet is less than 27°C, which met the design requirement.

6.2. Failure Mode (FM)

6.2.1. FM Simulation Results for Case 1 to 4

Figure 18 shows the ASHRAE temperature compliance plot for FM Case 1. The peak air temperature measured at IT cabinet intake side is in the range of 25.84 to 34.82°C. The simulation result showed that many IT cabinets are having peak inlet air temperature between 27 to 32°C and greater than 32°C, which does not comply with ASHRAE recommended temperature range.

Figure 19 shows the ASHRAE temperature compliance plot for FM Case 2. The peak air temperature measured at IT cabinet intake side is in the range of 25.81 to 27.27°C. All the IT cabinets except five are having peak inlet air temperature in the range of 18 to 27°C, which comply with ASHRAE recommended temperature range.

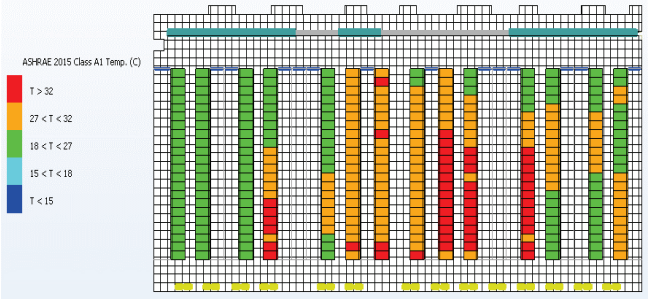


Figure 18: ASHRAE temperature compliance plot for FM Case 1

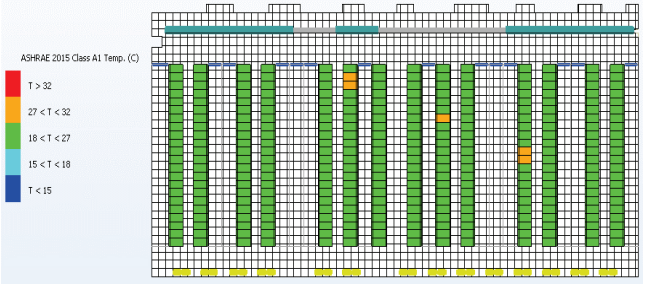


Figure 19: ASHRAE temperature compliance plot for FM Case 2

Figure 20 shows the ASHRAE temperature compliance plot for FM Case 3. The peak air temperature measured at IT cabinet intake side is in the range of 25.43 to 34.66°C. The simulation result showed that many IT cabinets are having peak inlet air temperature between 27 to 32°C and greater than 32°C, which does not comply with ASHRAE recommended temperature range.

Figure 21 shows the ASHRAE temperature compliance plot for FM Case 4. The peak air temperature measured at IT cabinet intake side is in the range of 25.58 to 27.15°C. All the IT cabinets except two are having peak inlet air temperature in the range of 18 to 27°C, which comply with ASHRAE recommended temperature range.

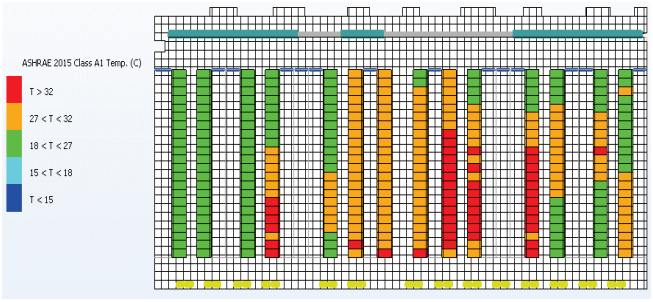


Figure 20: ASHRAE temperature compliance plot for FM Case 3

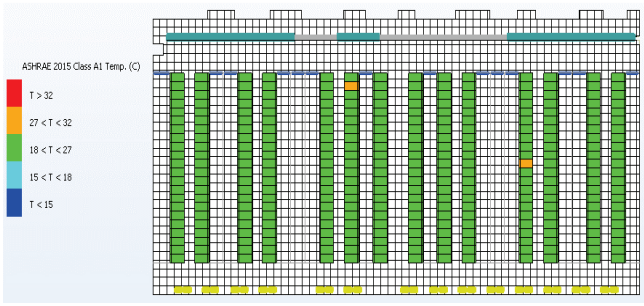


Figure 21: ASHRAE temperature compliance plot for FM Case 4

6.2.2. *FM Simulation Results Summary*

Table 4 summarizes the FM simulation results. From CFD simulation result it is observed that:

During FM of operation, no significant difference observed when simulation is run with or without control (Comparing Case 1&2 with Case 3&4). This is because during FM fan speed is ramped up to 100% and chilled water controller output also is almost 100%. The negative DP is observed across cabinets in the area served by offline cooling units. The area served by offline cooling units also experiences recirculation of hot air from the hot aisle into the cold aisle.

The results showed significant effect of leakages on the performance especially during FM. When comparing Case 1&3 with Case 2&4, with leakages heavy recirculation of air is observed, which causes leakage of hot air from hot aisle into the cold aisle. The recirculated hot air enters cabinet inlet resulting in increased inlet temperature of cabinet, in turn return air temperature increases and supply air temperature increases. Thus, increasing the room temperature. Because of which SLA temperatures recorded are higher and design requirement is not met. Therefore, it is important to minimize leakages.

Table 4: FM simulation results summary

Failure Mode (FM)				
Case	1	2	3	4
Control	✓	✓	✖	✖
Leakages	✓	✖	✓	✖
Simulation Results				
Fan Speed Controller Temp. Input	12.20°C	11.80°C	-	-
Fan Speed Group Controller Output	100.00%	100.00%	-	-
Fan Speed	1530 rpm	1530 rpm	1530 rpm	1530 rpm
Chilled Water Controller Output	81.02 to 100%	84.53 to 100%	-	-
Pressure Difference	-6.40 to 7.33 Pa	-6.35 to 15.77 Pa	-6.39 to 7.47 Pa	-6.38 to 15.86 Pa
Maximum On-Coil Temperature	39.17°C	38.37°C	38.98°C	38.24°C
Minimum On-Coil Temperature	35.61°C	36.56°C	35.32°C	36.36°C
Maximum Off-Coil Temperature	26.72°C	26.47°C	26.60°C	26.41°C
Minimum Off-Coil Temperature	24.96°C	24.97°C	24.74°C	24.85°C
Total Cooling Power (kW)	248.4 to 353.5	261.3 to 340.7	252.3 to 349.1	264.5 to 337.3
Coolant Temperature Out (Average)	27.16 to 29.84°C	27.22 to 29.55°C	26.96 to 29.73°C	27.15 to 29.49°C
Cabinet Maximum Temperature In	25.84 to 34.82°C	25.81 to 27.27°C	25.43 to 34.66°C	25.58 to 27.15°C
Cabinet Mean Temperature In	25.80 to 28.46°C	25.76 to 26.64°C	25.40 to 28.37°C	25.49 to 26.52°C
Room Temperature	25.82 to 27.16°C	25.79 to 26.93°C	25.41 to 27.08°C	25.46 to 26.75°C
Top-level Temperature Sensor at 2.4m	25.84 to 28.41°C	25.78 to 27.20°C	25.41 to 28.31°C	25.49 to 27.16°C
SLA Temperature Sensor at 1.5m	25.81 to 27.88°C	25.76 to 26.78°C	25.40 to 27.78°C	25.47 to 26.65°C
SLA Temperature Sensor at 0.9m	25.79 to 27.54°C	25.74 to 26.67°C	25.38 to 27.42°C	25.46 to 26.55°C

7. Operational Impact of Control Strategy & Leakages

7.1. Normal Mode Steady State Operation

The leakage causes recirculation of hot air from hot aisle back into the cabinet inlet through cabinets typical small gaps, resulting in increased peak inlet air temperature for only a few cabinets which is not a concern. The SLA sensor readings at 0.9 m and 1.5 m off floor in front of each cabinet are less than 27°C for all the cases, which met the design requirement.

When Case 1 and Case 3, both of which incorporate practical leakage conditions, are compared, the influence of the control strategy on the cooling performance of the data hall becomes evident. Under the control strategy, the ACUs operate at an optimized fan speed and chilled water flow rates, resulting in reduced overall power consumption. Specifically, the fan speed decreases from 1530 rpm to 1395 rpm, lowering the fan power demand. Similarly, the total chilled water flow rate across all ACUs decreases from 98.1 l/s to 77 l/s, which reduces pump power consumption due to the presence of a variable frequency drive (VFD).

The reduction in ACU fan speed also decreases the heat dissipation by fans into the data hall space, thereby lowering the cooling load on the chiller. Furthermore, the decrease in chilled water flow rate increases the chilled water return temperature for a fixed supply water temperature. As a result, the chiller evaporator operates at a higher temperature, increasing the evaporator saturation pressure. Because the condenser pressure remains unchanged (ambient conditions are constant), the compressor lift is reduced, leading to lower chiller compressor power consumption.

Overall, the implementation of the control strategy yields an approximate 9.89% reduction in cooling power consumption.

7.2. Failure Mode Operation

The results with control and without control logic are almost similar. Significant effect of leakages is observed on the performance especially during FM. Only in an ideal case with no leakages, SLA sensor readings at 0.9 m and 1.5 m off floor in front of each cabinet are than 27°C, which met the design requirement. But in a practical case with leakages, SLA sensor readings at 0.9 m and 1.5 m off floor in front of each cabinet are more than 27°C, which does not meet the design requirement, necessitating revision of the design cooling capacity.

In the FM with the control strategy active, the ACU fans continue to operate at full speed and the chilled water flow rate remains close to the rated value. Because the data hall heat load is nearly equal to the available cooling capacity, no meaningful optimization is possible under this condition. As a result, the control strategy provides no substantial reduction in power consumption.

8. Conclusions

A CFD based approach was adopted in this paper to assess the cooling performance of a dynamically controlled, non-raised floor data hall with a HAC configuration. The control strategy, which adjusts ACU fan speed and chilled-water flow rates using real-time temperature and pressure feedback, effectively maintains cabinet inlet temperatures within allowable limits while reducing overall cooling energy consumption. Under normal mode operation, control strategy reduces fan, pump, and chiller compressor power consumption and lowers the chiller cooling load, resulting in approximately 9.89% overall energy savings. The leakages become critical during failure mode, as only the idealized no leakage scenario satisfies the SLA temperature requirement, whereas practical leakage results in non-compliance with the design criteria.

The key insights from the CFD analysis are highlighted as follows:

- The accurate prediction of data center cooling performance is made possible by the use of CFD technology.
- By using performance-based analysis, issues were identified and addressed at the design stage itself, which minimized rework by testing the design or design changes prior to implementation.
- By analyzing multiple simulation scenarios, potential failures are identified, which minimizes the risk of failures and leads to an accurate design for the data center.
- The design requirements are met while the efficient operation of the data center is achieved through the control strategy used.
- The proper design studies and predictions from the CFD simulation provide assurance that the data center will perform reliably and efficiently under normal and failure mode of operation.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to acknowledge the support and resources provided by Buildings & Factories (B&F) IC, L&T Construction, under whose auspices this project was undertaken for the client. We also extend our sincere gratitude to the design wing of B&F IC, L&T Construction,

which carries out performance-based designs through its CFD department, for their collaboration and insightful feedback, which greatly contributed to the successful outcome of this work.

References

[1] Y. Zhang, J. Liu, "Prediction of Overall Energy Consumption of Data Centers in Different Locations," *Sensors*, vol. 22, no. 10, pp. 3704, 2022, doi:10.3390/s22103704.

[2] E. Masanet, A. Shehabi, N. Lei, S. Smith, J. Koomey, "Recalibrating global data center energy-use estimates," *Science*, vol. 367, no. 6481, pp. 984–986, 2020, doi:10.1126/science.aba3758.

[3] CISCO, Cisco Global Cloud Index: Forecast and Methodology, 2016–2021, 2018.

[4] International Energy Agency, Digitalization & Energy, 2017.

[5] A. Shehabi, S.J. Smith, E. Masanet, J. Koomey, "Data center growth in the United States: Decoupling the demand for services from electricity use," *Environmental Research Letters*, vol. 13, no. 12, 2018, doi:10.1088/1748-9326/aaec9c.

[6] ABB, Data centers energy efficiency and management, 2023.

[7] M. Law, "Energy efficiency predictions for data centres in 2023," 2022.

[8] Y. Liu, X. Wei, J. Xiao, Z. Liu, Y. Xu, Y. Tian, "Energy consumption and emission mitigation prediction based on data center traffic and PUE for global data centers," *Global Energy Interconnection*, vol. 3, no. 3, pp. 272–282, 2020, doi:10.1016/j.gloei.2020.07.008.

[9] P. Sharma, P. Pegus II, D. Irwin, P. Shenoy, J. Goodhue, J. Culbert, "Design and Operational Analysis of a Green Data Center," *IEEE Internet Computing*, vol. 21, no. 4, pp. 16–24, 2017, doi:10.1109/MIC.2017.2911421.

[10] J. Gao, "Machine Learning Applications for Data Center Optimization," 2014.

[11] Sullivan R, Alternating Cold and Hot Aisles Provides More Reliable Cooling for Server Farms, 2000.

[12] C.D. Patel, C.E. Bash, L. Stahl, D. Sullivan, "Computational Fluid Dynamics Modeling of High Compute Density Data Centers to Assure System Inlet Air Specifications," in *IPACK*, ASME, 2001.

[13] S. Patankar, "Airflow and Cooling in a Data Center," *ASME Journal of Heat Transfer*, vol. 132, no. 7, 2010, doi:10.1115/1.4000703.

[14] S. Pogorelskiy, I. Kocsis, "BIM and Computational Fluid Dynamics Analysis for Thermal Management Improvement in Data Centres," *Buildings*, vol. 13, no. 10, 2023, doi:10.3390/buildings13102636.

[15] D. Jiang, "Effects and optimization of airflow on the thermal environment in a data center," *Frontiers in Built Environment*, vol. 10, , 2024, doi:10.3389/fbuil.2024.1362861.

[16] J. Cho, C. Park, W. Choi, "Numerical and experimental study of air containment systems in legacy data centers focusing on thermal performance and air leakage," *Case Studies in Thermal Engineering*, vol. 26, , 2021, doi:10.1016/j.csite.2021.101084.

[17] J. Cho, J. Woo, B. Park, T. Lim, "A comparative CFD study of two air distribution systems with hot aisle containment in high-density data centers," *Energies*, vol. 13, no. 22, 2020, doi:10.3390/en13226147.

[18] C. Zhou, Y. Hu, R. Liu, Y. Liu, M. Wang, H. Luo, Z. Tian, "Energy Performance Study of a Data Center Combined Cooling System Integrated with Heat Storage and Waste Heat Recovery System," *Buildings*, vol. 15, no. 3, 2025, doi:10.3390/buildings15030326.

[19] Y. Guo, C. Zhao, H. Gao, C. Shen, X. Fu, "Improving Thermal Performance in Data Centers Based on Numerical Simulations," *Buildings*, vol. 14, no. 5, 2024, doi:10.3390/buildings14051416.

[20] Kao Data, Using Simulation to Validate Cooling Design, 2021.

[21] AKCP, Computational Fluid Dynamics to Improve the Performance of Data Centers, 2021.

[22] B. Zhan, S. Shao, M. Lin, H. Zhang, C. Tian, Y. Zhou, "Experimental investigation on ducted hot aisle containment system for racks cooling of data center," *International Journal of Refrigeration*, vol. 127, pp. 137–147, 2021, doi:10.1016/j.ijrefrig.2021.02.006.

[23] M. Tatchell-Evans, N. Kapur, J. Summers, H. Thompson, D. Oldham, "An experimental and theoretical investigation of the extent of bypass air within data centres employing aisle containment, and its impact on power consumption," *Applied Energy*, vol. 186, pp. 457–469, 2017, doi:10.1016/j.apenergy.2016.03.076.

[24] S.A. Alkharabsheh, B.G. Sammakia, S.K. Shrivastava, "Experimentally Validated Computational Fluid Dynamics Model for a Data Center with Cold Aisle Containment," *Journal of Electronic Packaging*, vol. 137, no. 2, pp. 21010, 2015, doi:10.1115/1.4029344.

[25] C. Gao, Z. Yu, J. Wu, "Investigation of Airflow Pattern of a Typical Data Center by CFD Simulation," *Energy Procedia*, vol. 78, pp. 2687–2693, 2015, doi:10.1016/j.egypro.2015.11.350.

[26] S.A. Nada, M.A. Said, "Effect of CRAC units layout on thermal management of data center," *Applied Thermal Engineering*, vol. 118, pp. 339–344, 2017, doi:10.1016/j.applthermaleng.2017.03.003.

[27] R. Zhou, Z. Wang, "Modeling and Control for Cooling Management of Data Centers with Hot Aisle Containment," in *IMECE*, ASME: 739–746, 2011, doi:10.1115/IMECE2011-62506.

[28] C.D. Patel, C.E. Bash, R. Sharma, M. Beitelmal, R. Friedrich, "Smart cooling of data centers," in *Advances in Electronic Packaging*, American Society of Mechanical Engineers: 129–137, 2003, doi:10.1115/ipack2003-35059.

[29] C.B. Bash, C.D. Patel, R.K. Sharma, "Dynamic thermal management of air cooled data centers," in *Thermal and Thermomechanical Proceedings 10th Intersociety Conference on Phenomena in Electronics Systems, 2006 (ITHERM 2006)*, pp. 8– 452, 2006, doi:10.1109/ITHERM.2006.1645377.

[30] S. Nagarathinam, B. Fakhim, M. Behnia, S. Armfield, "Thermal Performance of an Air-Cooled Data Center With Raised-Floor and Non-Raised-Floor Configurations," *Heat Transfer Engineering*, vol. 35, pp. 384–397, 2014, doi:10.1080/01457632.2013.828559.

[31] K. Khankari, "Analysis of Air Leakage from Hot Aisle Containment Systems and Cooling Efficiency of Data Centers," in *ASHRAE Winter Conference*, 2014.

[32] H. Alissa, K. Nemati, B. Sammakia, K. Ghose, M. Seymour, D. King, R. Tipton, "Ranking and Optimization of CAC and HAC Leakage Using Pressure Controlled Models," in *Proceedings of the ASME IMECE*, 2015, doi:10.1115/IMECE2015-50782.

[33] Z. Song, B.T. Murray, B. Sammakia, "Parametric analysis for thermal characterization of leakage flow in data centers," in *Fourteenth Intersociety Conference on Thermal and Thermomechanical*

Phenomena in Electronic Systems (ITherm), IEEE: 778–785, 2014, doi:10.1109/ITHERM.2014.6892360.

[34] Y.U. Makwana, A.R. Calder, S.K. Shrivastava, "Benefits of properly sealing a cold aisle containment system," in *Fourteenth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, IEEE: 793–797, 2014, doi:10.1109/ITHERM.2014.6892362.

[35] E. Wibron, A.L. Ljung, T. Staffan Lundström, "Comparing performance metrics of partial aisle containments in hard floor and raised floor data centers using CFD," *Energies*, vol. 12, no. 8, 2019, doi:10.3390/en12081473.

[36] Y.-T. Lee, C.-Y. Wen, Y.-C. Shih, Z. Li, A.-S. Yang, "Numerical and experimental investigations on thermal management for data center with cold aisle containment configuration," *Applied Energy*, vol. 307, , pp. 118213, 2022, doi:10.1016/j.apenergy.2021.118213.

[37] D. Macedo, R. Godina, P.D. Gaspar, P.D. da Silva, M.T. Covas, "A parametric numerical study of the airflow and thermal performance in a real data center for improving sustainability," *Applied Sciences*, vol. 9, no. 18, 2019, doi:10.3390/app9183850.

[38] J. Cho, T. Lim, B.S. Kim, "Measurements and predictions of the air distribution systems in high compute density (Internet) data centers," *Energy and Buildings*, vol. 41, no. 10, pp. 1107–1115, 2009, doi:10.1016/j.enbuild.2009.05.017.

[39] S. Alkharabsheh, J. Fernandes, B. Gebrehiwot, D. Agonafer, K. Ghose, A. Ortega, Y. Joshi, B. Sammakia, "A Brief Overview of Recent Developments in Thermal Management in Data Centers," *Journal of Electronic Packaging, Transactions of the ASME*, vol. 137, no. 4, pp. 40801, 2015, doi:10.1115/1.4031326.

[40] E. Wibron, A.L. Ljung, T.S. Lundström, "Computational fluid dynamics modeling and validating experiments of airflow in a data center," *Energies*, vol. 11, no. 3, 2018, doi:10.3390/en11030644.

[41] R. Sethuramalingam, A. Asthana, Design Improvement of Water-Cooled Data Centres Using Computational Fluid Dynamics, Springer: 105–113, 2021, doi:10.1007/978-3-030-63916-7_14.

[42] A. Almoli, A. Thompson, N. Kapur, J. Summers, H. Thompson, G. Hannah, "Computational fluid dynamic investigation of liquid rack cooling in data centres," *Applied Energy*, vol. 89, no. 1, pp. 150–155, 2012, doi:10.1016/j.apenergy.2011.02.003.

[43] R. Balakrishnan, M. Munirajulu, "CFD Simulation of Tier 4 Data Center for Cooling and Backup Power," in *2023 2nd International Conference for Innovation in Technology (INOCON)*, 1–7, 2023, doi:10.1109/INOCON57975.2023.10101234.

[44] D. Pickut, Data Center Design: Raised Floor Versus Slab Floor?, 2011.

[45] H.K. Versteeg, W. Malalasekera, An Introduction to Computational Fluid Dynamics, Second Edition, Pearson, 2007.

[46] S.V. Patankar, Numerical Heat Transfer and Fluid Flow, First Edition, Hemisphere Publishing Corporation, 1980.

[47] J.D., Jr. Anderson, Computational Fluid Dynamics: The basics with applications, McGraw-Hill Education, 1995.

[48] J.H. Ferziger, M. Perić, Computational Methods for Fluid Dynamics, Third Edition, Springer, 2002.

[49] J. 'Tannehill, A. 'Dale, R. Pletcher, Computational Fluid Mechanics and Heat Transfer, Second Edition, Taylor&Francis, 1997.

[50] B.E. Launder, D.B. Spalding, "The numerical computation of turbulent flows," *Computer Methods in Applied Mechanics and Engineering*, vol. 3, no. 2, pp. 269–289, 1974, doi:10.1016/0045-7825(74)90029-2.

[51] Ansys, Ansys CFX-Solver Modeling Guide, 2025.

[52] S.A. Nada, M.A. Said, M.A. Rady, "CFD investigations of data centers' thermal performance for different configurations of CRACs units and aisles separation," *Alexandria Engineering Journal*, vol. 55, no. 2, pp. 959–971, 2016, doi:10.1016/j.aej.2016.02.025.

[53] D.D. Gray, A. Giorgini, "The validity of the boussinesq approximation for liquids and gases," *International Journal of Heat and Mass Transfer*, vol. 19, no. 5, pp. 545–551, 1976, doi:10.1016/0017-9310(76)90168-X.

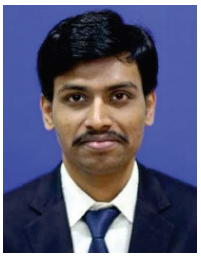
[54] Cadence Reality DC Design, https://www.cadence.com/en_US/home/resources/product-briefs/cadence-reality-dc-design-pb.html, 2025.

[55] E. Frachtenberg, D. Lee, M. Magarelli, V. Mulay, J. Park, "Thermal design in the open compute datacenter," in *ITherm*, IEEE: 530–538, 2012, doi:10.1109/ITHERM.2012.6231476.

[56] H. Alissa, K. Nemati, B. Sammakia, K. Ghose, M. Seymour, R. Schmidt, "Innovative Approaches of Experimentally Guided CFD Modeling for Data Center," in *SEMI-THERM*, IEEE: 176–184, 2015, doi:10.1109/SEMI-THERM.2015.7100157.

[57] M.I. Tradat, Y. Manaserh, B.G. Sammakia, C.H. Hoang, H.A. Alissa, "An experimental and numerical investigation of novel solution for energy management enhancement in data centers using underfloor plenum porous obstructions," *Applied Energy*, vol. 289, , 2021, doi:10.1016/j.apenergy.2021.116663.

[58] ASHRAE TC 9.9, 2021 Equipment Thermal Guidelines for Data Processing Environments.



Sushil Ashok Surwase holds a master's degree in mechanical engineering from IIT Madras. He is currently working as an Assistant Engineering Manager, Buildings & Factories (B&F) IC, L&T Construction with more than 3 years of experience in CFD analysis.

He is experienced in Data Center Analysis (3D and 1D), Air-Conditioning Analysis, Thermal Comfort and Ventilation Analysis, Rain Ingress Analysis, Egress Analysis, DG Room Ventilation Analysis, Fire and Smoke Analysis, External Flow and Wind Load Analysis. He has conducted CFD analysis for some of the iconic projects such as High-Speed Rail (MAHSR), Airports (DIAL, NMIAL), Data Centers (Equinix, STT, DataVolt), Yashobhoomi (IICC Delhi), Hospitals (SCB, TIMS), Residential (Manora Aamdar Niwas) etc. He has presented papers at multiple international conferences and has received Best Research Paper Award. He has also

published research papers in reputed journals such as Thermal Science and Engineering Progress (TSEP).

Suribabu Badde is Head – CFD, Buildings & Factories (B&F) IC, L&T Construction, with over 20 years of expertise in simulation driven engineering.

He specializes in applying CFD to MEP and special systems, covering HVAC, Fire Engineering, Public Health Engineering (PHE), and Electrical systems, to deliver optimized and performance-based design solutions. Passionate about performance-based design, Suribabu has leveraged advanced simulation tools to transform building designs for efficiency, safety, and sustainability. His domain expertise extends beyond buildings into wind turbines, aerodynamics, and automotive applications, reflecting a strong multidisciplinary engineering background. As CFD Head, Suribabu has successfully led critical projects across sectors, including Airports, Data Centers, High-Speed Rail Corridor, Commercial, and Wind Turbine Projects. Currently, he is spearheading initiatives that integrate Artificial Intelligence (AI) with CFD.

R. Balakrishnan is Vice President & HEAD – MEP Design, Buildings & Factories (B&F) IC, L&T Construction with more than 36 years of experience in all facets of industry.

He is well versed in codes & standards like NBC, IS, BS, IEC, IEEE, NFPA, ISHRAE, ASHRAE, UPC, etc. Some of the iconic projects that he handled as MEP Design Head are Statue of Unity, Gujarat Cricket Stadium, Airports (HIAL, DIAL, BIAL, MIAL), Data centers, Hospitals, Exhibition Centers (Mahatma Mandir), Office and other Commercial buildings. He is an active member of IEEE, ISLE, IEE, NFE & FSAI. He is a Chartered Engineer from IEI (F-114811-3). He is also Chairman - FOCUS, Chennai chapter and has done many publications in various forums. He has spearheaded and institutionalized performance-based design practices across diverse projects, driving innovation and achieving superior outcomes through the strategic application of CFD.

Received: 12 November 2025, Revised: 21 December 2025, Accepted: 23 December 2025, Online: 14 January 2026

DOI: https://doi.org/10.55708/js0501003

Cross-Sectional Structure of Nested Antiresonant Nodeless Fiber for Single-Mode and Few-Mode Transmission

Shogo Ota[1], Hirokazu Kubota[1,2*]

[1]Osaka Metropolitan University, Graduate School of Engineering, Division of Electrical and Electronic Engineering, Sakai-shi, 599-8531, Japan
[2]Otemon Gakuin University, Faculty of Science and Engineering, Department of Electrical and Electronic Engineering, Ibaraki-shi, 567-8502, Japan
*Corresponding author: Hirokazu Kubota, 2-1-15 Nishi-ai, Ibaraki-shi, Osaka, +81 72 641 9556, h-kubota@haruka.otemon.ac.jp

ABSTRACT: Nested Antiresonant Nodeless Fiber (NANF) is a promising candidate for next-generation optical communication systems due to its low-loss, low-latency and low-nonlinearity characteristics. This study focuses on the high degree of design flexibility inherent in NANF, demonstrating through numerical analysis that a single platform can be tailored for two distinct applications required in future networks: single-mode transmission and few-mode transmission for space-division multiplexing. Although low-loss HCFs are by nature multimode fibers, we show that the fiber's modal properties can be actively controlled by adjusting one of the key design parameters: the radius of the inner nested tubes (r_2). A design with a smaller radius (r_2=5.31 μm) achieves quasi-single-mode transmission by maintaining the fundamental mode loss below 1 dB/km while establishing a loss ratio greater than a factor of ten on a decibel scale relative to higher-order modes. Conversely, a design optimized with a larger radius (r_2=7.2 μm) demonstrates quasi-two-mode operation at a wavelength of 1.3 μm, where both the fundamental (0.22 dB/km) and the first higher-order (0.81 dB/km) modes propagate with low loss. These results reveal that NANF is an highly versatile optical fiber platform whose performance can be switched from single-mode to few-mode simply by adjusting one structural parameter. This capability indicates that NANF could play a crucial role in meeting the diverse requirements of future optical communication networks.

KEYWORDS: Antiresonant fiber, few-mode fiber, hollow-core fiber, optical fiber design.

1. Introduction

The Hollow-core fibers (HCFs) have been the subject of active research and development for several decades as a next-generation optical fiber technology capable of overcoming the physical limitations of conventional silica glass fibers [1]. Unlike conventional fibers, which guide light by total internal reflection due to the refractive index difference between the core and cladding, HCFs confine light within a hollow, air-filled core based on principles such as the photonic bandgap effect or anti-resonance [2]. This structure endows HCFs with the potential for exceptional properties that are difficult to achieve with solid-core fibers, including ultra-low transmission loss, low nonlinearity, and reduced latency, as light propagates at nearly the speed of light in a vacuum [3].

Among the various HCF architectures, the Nested Antiresonant Nodeless Fiber (NANF), whose adjacent capillaries does not touch each other, has garnered significant low attention for its ability to significantly reduce confinement loss through an optimized cladding design [4]. In 2020, a single-mode NANF was reported to exhibit an attenuation of 0.28 dB/km over the wavelength range between 1510 and 1600 nm and approximately 0.3 dB/km over a 2.8 km fiber between 1500 and 1640 nm [5]. The NANF technology has advanced rapidly, with a recent breakthrough in Double-Nested Antiresonant Nodeless Fiber (DNANF) achieving sub-0.1-dB/km loss from 1320 nm to 2 μm, outperforms that of any existing single-mode fibers [6], [7]. Consequently, it is now regarded as a leading candidate to replace conventional single-mode fibers (SMFs) in next-generation optical communication systems [8], [9].

Looking ahead to the evolution of future optical communication networks, two primary application trajectories for NANF emerge. The first is its use as a single-mode transmission path, leveraging its ultra-low loss characteristics to minimize signal degradation. Achieving this requires a design that exclusively propagates the fundamental mode with low loss while effectively suppressing unwanted higher-order modes that can degrade communication quality [10]. The second trajectory is its application in Space-Division Multiplexing (SDM) technology, aimed at expanding transmission capacity to meet ever-increasing data traffic demands [11] [12]. In this approach, the fiber must be intentionally

designed to stably guide multiple propagation modes (few-mode) with low inter-modal crosstalk. Indeed, HCF designs capable of supporting as many as eight core modes with low loss and weak coupling have been reported, demonstrating the feasibility of HCF-based SDM systems [10].

While these two applications have often been pursued as separate research endeavors, this study focuses on the high degree of design freedom inherent in NANF. We aim to comprehensively demonstrate through numerical analysis that by optimizing its structural parameters, a single NANF platform can be tailored to meet the distinct requirements of both single-mode and few-mode transmission. Specifically, by comparing and contrasting a design that intentionally increases higher-order mode loss with a design that simultaneously reduces the loss of both the fundamental and first higher-order modes, we identify the key structural factors that govern the number of transmission modes. Through this investigation, we provide a clear design guideline for the application of NANF in future optical communication systems.

2. Principle of the NANF

The. The light confinement mechanism in a NANF can be explained by the anti-resonant reflecting optical waveguide (ARROW) model, rather than by total internal reflection. The key to this model is the thickness t of the thin glass tubes that constitute the cladding. At specific wavelengths, known as the resonant wavelengths λ_{Res}, light resonates within the glass tubes and leaks out into the cladding layer instead of being confined to the core. This resonant wavelength λ_{Res} is given by the following equation:

$$\lambda_{Res} = \frac{2t}{m}\sqrt{n_g^2 - n_{air}^2} \quad (1)$$

Here, t is the thickness of the glass tube, n_g is the refractive index of the glass, n_{air} is the refractive index of air, and m is a positive integer. Conversely, in the wavelength range that satisfies the anti-resonance condition, positioned between the resonant wavelengths, light is strongly reflected at the glass-air interfaces and is efficiently confined within the core. This enables low-loss optical transmission. The anti-resonant wavelength λ_{ARes} is expressed as:

$$\lambda_{ARes} = \frac{4t}{2m-1}\sqrt{n_g^2 - n_{air}^2} \quad (2)$$

When m is small, the adjacent resonant wavelengths λ_{ARes} are far apart, NANF can exhibit low-loss characteristics over a broad bandwidth. Critically, this guiding principle applies not only to the fundamental mode but also to higher-order modes under different conditions. Therefore, by varying the structure of the small nested tubes in the cladding (e.g., their radii and thickness), it is possible to control the resonance and anti-resonance conditions for each mode. This is the fundamental principle of mode control in NANF, which allows one design to achieve single-mode transmission by intentionally leaking higher-order modes, while another design enables few-mode transmission by simultaneously guiding multiple modes with low loss.

3. Single-mode NANF

To achieve single-mode transmission, it is essential not only to maintain low loss for the fundamental mode but also to suppress higher-order modes, which can cause signal degradation, by increasing their loss. In this study, we designed a NANF cross-section to achieve this goal and numerically evaluated its mode-dependent loss characteristics. The cross-sectional structure of this design is shown in Fig. 1. Here, R is the core radius, t is the tube thickness, r_1 and r_2 are the radii of the large and small nested tubes, respectively, and T is the thickness of the outer capillary. The perfectly matched layer (PML) is placed at the outer capillary to absorb outgoing fields.

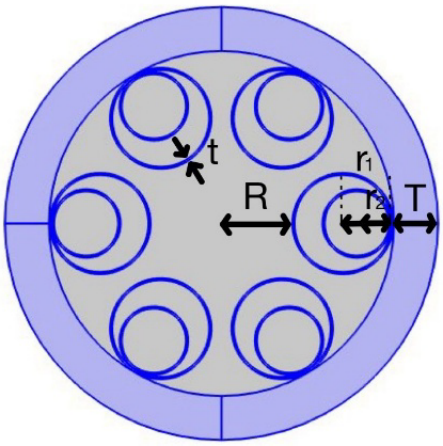


Figure 1: The cross-sectional structure of NANF

For the analysis, the NANF was modeled with a core radius R of 15 μm and a glass tube thickness t of 0.42 μm. The radii of the large and small nested tubes were set to r_1 =10.62 μm and r_2=5.31 μm, respectively.

Simulations were performed using COMSOL Multiphysics®, a numerical analysis software based on the finite element method (FEM). The computational domain was defined to search for ten modes around an effective refractive index of 1, and the complex effective refractive index was calculated for each mode. PML was applied to the outermost layer of the structure to ensure that guided modes were absorbed without reflection at the interface. The loss evaluated in this study is the "confinement loss," which arises solely from the light-confining ability of the ideal structure, neglecting structural non-uniformities and material absorption loss. The wavelength dependence of the refractive index for silica was calculated using the Sellmeier equation. The confinement loss was calculated from the following

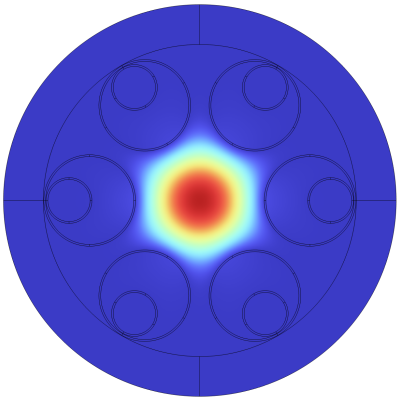


Figure 2: The electric field distribution for the fundamental mode at 1.3μm (r_2=5.31 μm)

equation, where α is the imaginary part of the effective refractive index.

$$Confinement\ Loss[dB/km] = \frac{\alpha \log_{10} e}{100} \qquad (3)$$

Figure 2 shows the electric field distribution at a wavelength of 1.3 μm for the fundamental mode, and Fig.3 shows that of the first higher-order mode, which had the lowest loss among the higher-order modes. The right-hand side of Fig.3 shows the contour lines of the electric field distribution, which illustrate the large electric field leakage. As is evident from this figures, the NANF has higher-order modes in addition to the fundamental mode. They are confined within the core in the same wavelength, indicating that this structure is inherently a multimode fiber. However, by setting an appropriate cross-sectional parameter, NANF can be operate in quasi-single-mode. The calculated wavelength dependence of the confinement loss is shown in figure 4. The vertical and the horizontal axis represent the confinement loss and the wavelength, respectively. The red, the yellow, and the black lines represent the losses of the fundamental mode, the first higher-order mode, and the other higher-order mode, respectively. The results indicate that the fundamental mode maintains a low confinement loss of less than 1 dB/km over the broad wavelength range of 1.0 μm to 1.7 μm. Table 1 shows the confinement losses for each wavelength. Furthermore, within this low-loss window, the loss difference between the fundamental mode and the lowest-loss higher-order mode is more than

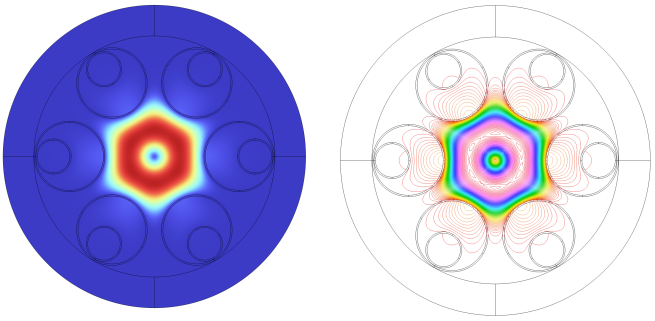


Figure 3: The electric field distribution for the lowest-loss higher-order mode at 1.3μm (r_2=5.31 μm)

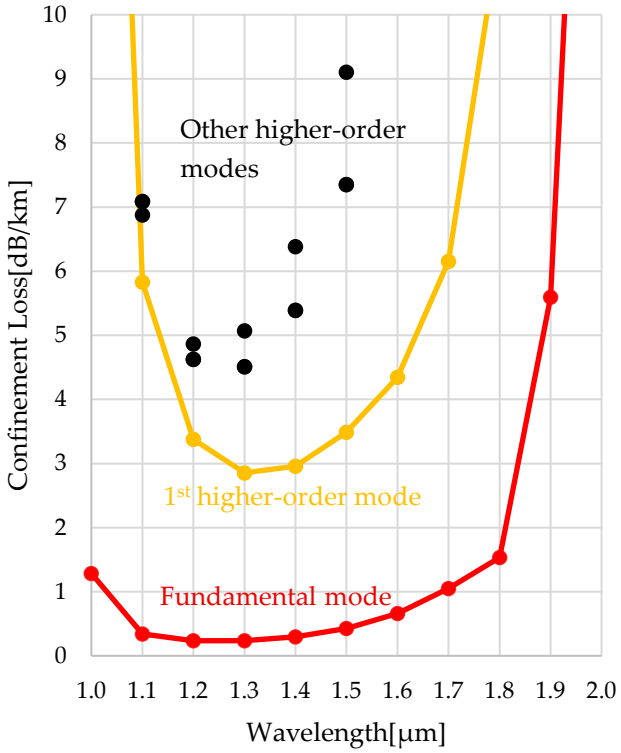


Figure 4: The wavelength dependence of the confinement losses (r_2 = 5.31μm)

a factor of ten on a decibel scale. This significant loss differential ensures that even if higher-order modes are excited, they are rapidly attenuated as they propagate through the fiber, meaning that effectively only the fundamental mode is transmitted. Thus, a broadband quasi-single-mode NANF is achievable.

Table 1: The confinement losses for each wavelength (r_2 = 5.31μm)

Wavelength [μm]	Fundamental mode [dB/km]	First higher-order mode [dB/km]
1.0	1.29	26.26
1.1	0.34	6.88
1.2	0.23	4.62
1.3	0.24	2.85
1.4	0.29	2.96
1.5	0.42	3.49
1.6	0.66	4.35
1.7	1.05	6.15
1.8	1.53	11.29
1.9	5.59	29.46

4. Two-mode NANF

In contrast to single-mode transmission, applications in Space-Division Multiplexing (SDM) require a fiber design that intentionally and stably propagates multiple modes with low loss. This section investigates the feasibility of realizing a quasi-two-mode NANF that guides both the fundamental mode and the first higher-

order mode with low loss. The core strategy in this design is to adjust the cladding structure, specifically the radius of the inner tube r_2. Because the space formed between the large tube r_1 and the small tube r_2 was considered to be strongly related to the light confinement and the resonance conditions of higher-order modes, we specifically varied the value of r_2 in this simulation. The objective was to create a wavelength region where both the fundamental and first higher-order modes simultaneously satisfy the antiresonance condition.

In the analysis, the core radius R was set to 15 μm and the tube thickness t to 0.42 μm, in accordance with the single-mode design described in Section 3. The radius r_2 was optimized to minimize the loss of the first higher-order mode at a wavelength of 1.3 μm. In this process, the range of r_2 was explored around 7 μm to ensure that the mode field diameter (MFD) is approximately 10 μm, which is comparable to that of a standard SMF. The loss reached its minimum when r_2 was 7.2 μm. Figure 5 shows the electric field distribution at a wavelength of 1.3 μm for the fundamental mode. Figure 6 presents the electric field distribution of the lowest-loss higher-order modes for this structure. The right-hand side of figure 6 shows the contour lines of the electric field distribution. The fundamental mode shown in figure 5 exhibits no noticeable difference from that in figure 2, whereas the first higher-order mode shown on right-hand side of figure 6 exhibits slightly stronger confinement than that shown on the right-hand side of figure 3.

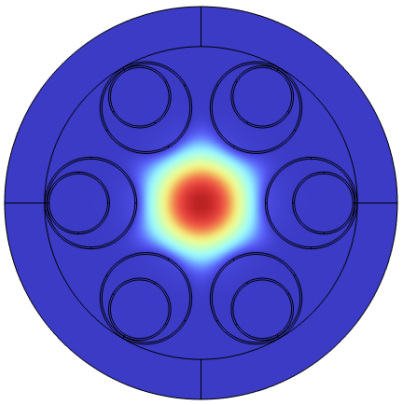


Figure 5: The electric field distribution for the fundamental mode at 1.3 μm (r_2=7.2 μm)

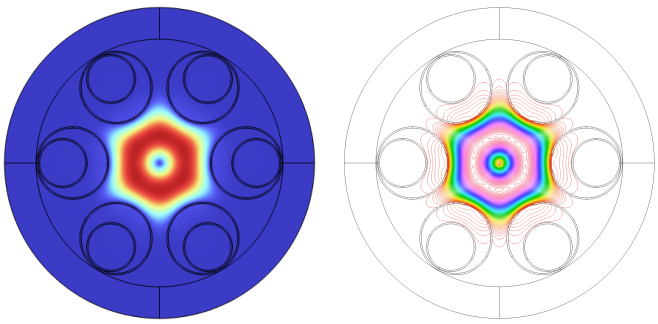


Figure 6: The electric field distribution for the lowest-loss higher-order mode at 1.3 μm (r_2=7.2 μm)

The wavelength dependence of the confinement loss for this optimized structure is shown in figure 7. The vertical and the horizontal axis represent the confinement loss and the wavelength, respectively. The red, the yellow, and the lines represent the losses of the fundamental mode, the first higher-order mode, and black dots represent those of other higher-order modes, respectively. Table 2 shows that at a wavelength dependence confinement loss for the fundamental and first higher-order modes. at a wavelength of 1.3 μm, the losses for the fundamental and first higher-order modes are 0.22 dB/km and 0.81 dB/km, with corresponding α values of $5.28*10^{-12}$ and $1.93*10^{-11}$, respectively, indicating that both modes are guided with low attenuation. Meanwhile, the minimum loss of the second higher-order mode was 3.43 dB/km at a wavelength of 1.2 μm, providing a comparable loss margin to that of quasi-single mode NANF. This result suggests the possibility of operation as a quasi-two-mode fiber, capable of propagating the two intended modes while suppressing other unwanted higher-order modes.

Furthermore, the dispersion characteristics of this design were evaluated, with the results shown in Fig. 8. The horizontal axis is wavelength, and the vertical axis is dispersion; the red line indicates the fundamental mode, and the yellow line indicates the first higher-order mode. As shown in the figure, the dispersion slope is well-suppressed for both the fundamental and higher-order modes within the primary transmission band of 1.1 μm to 1.8 μm. Specifically, at a wavelength of 1.3 μm, the dispersion was 2.57 ps/nm/km for the fundamental mode and 6.36 ps/nm/km for the first higher-order mode. Notably, the dispersion of the fundamental mode is kept small over a wide wavelength range compared to standard single-mode fiber, which is advantageous for easing communication system design. The differential mode delay (DMD) was calculated to be few thousands ps/km in the low-loss wavelength range. This value is less than half of that of a typical step-index fiber but is about two orders of magnitude larger than that of a graded-index fiber. These findings indicate that the hollow-core structure of HCF does not necessarily eliminate dispersion nor reduce the differential mode delay. Additionally, low-dispersion bandwidth of the higher-order mode tends to be narrower than that of the fundamental mode. Therefore, controlling dispersion, reducing DMD, and expanding the transmission bandwidth of higher-order modes also remain challenges for future work.

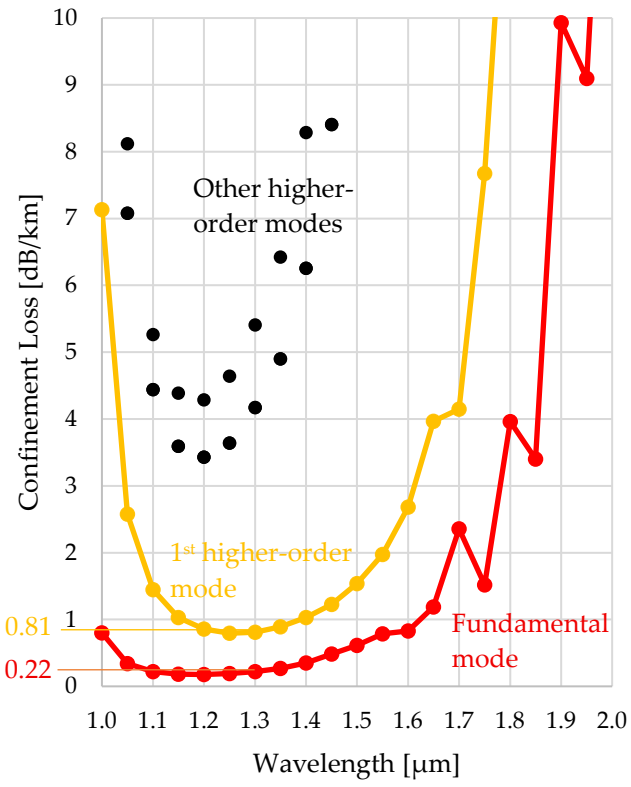


Figure 7: The wavelength dependence of the confinement losses (r_2 = 7.2μm)

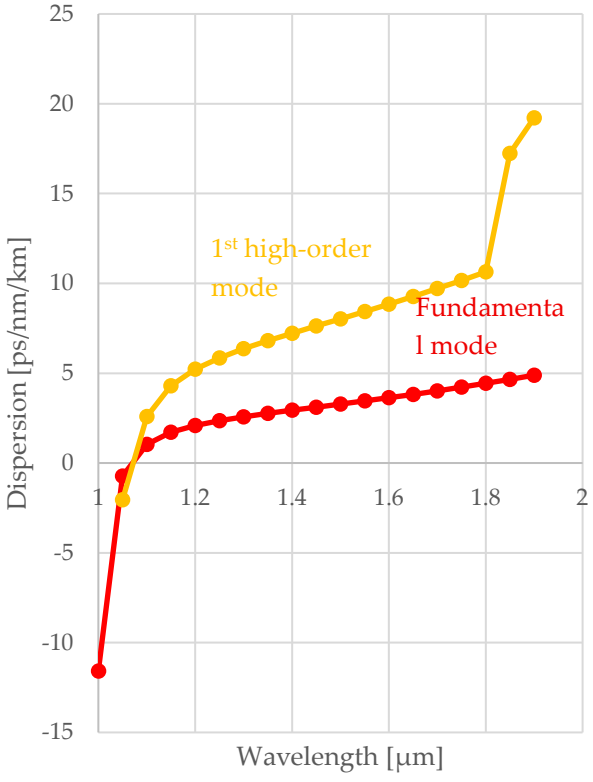


Figure 8: The wavelength dependence of the dispersion (r_2 = 7.2μm)

Table 2: The confinement losses for each wavelength (r_2 = 7.2μm)

Wavelength [μm]	Fundamental mode [dB/km]	First higher-order mode [dB/km]
1.00	0.800	7.134
1.05	0.341	2.578
1.10	0.221	1.446
1.15	0.183	1.031
1.20	0.177	0.856
1.25	0.191	0.795
1.30	0.221	0.808
1.35	0.266	0.890
1.40	0.348	1.032
1.45	0.482	1.229
1.50	0.612	1.539
1.55	0.785	1.972
1.60	0.831	2.685
1.65	1.186	3.967
1.70	2.358	4.146

5. Conclusion

In this study, we numerically demonstrated that the design flexibility of NANF can be leveraged to create fibers tailored for two distinct communication applications: single-mode and few-mode transmission. In a design where the inner nested tube radius r_2 was set to a half of r_1, we successfully achieved a large loss ratio of more than a factor of ten between the fundamental mode and the higher-order modes while keeping the fundamental mode loss below 1 dB/km, thus demonstrating the feasibility of effective single-mode transmission. In contrast, when the r_2 was increased to 7.21 μm, both the fundamental mode with a loss of 0.22 dB/km and the first higher-order mode with a loss of 0.81 dB/km exhibit low confinement loss at a wavelength of 1.3 μm, demonstrating its capability to operate as a quasi-two-mode fiber. This study agrees with the findings of Ref. [3] regarding fundamental mode loss, which validates the current method given their structural similarities.

In conclusion, NANF is an extremely versatile platform whose modal characteristics can be actively controlled simply by making minor changes to the cladding structure, specifically by adjusting a single parameter, r_2. While other studies utilize numerous design elements to achieve low-loss characteristics across many modes, this work specifically investigates a few-mode regime within the NANF structure [10]. Consequently, it is considered challenging to realize low attenuation for a large number of modes simultaneously, given the limited number of controllable parameters. This high degree of design freedom strongly suggests that NANF can be a powerful solution to meet the diverse and

evolving demands of future optical communication networks.

Acknowledgement

This work was supported by the National Institute of Information and Communications Technology (NICT) (JPJ012368C 08401).

References

[1] W. Ding, Y. Y. Wang, S. F. Gao, M. L. Wang, P. Wang, "Recent Progress in Low-Loss Hollow-Core Anti-Resonant Fibers and Their Applications," IEEE Journal of Selected Topics in Quantum Electronics, 2019, doi: 10.1109/JSTQE.2019.2957445.

[2] N.M. Litchinitser, A.K. Abeeluck, C. Headley, B.J. Eggleton, "Antiresonant reflecting photonic crystal optical waveguides," OPTICS LETTERS, 2002, doi: 10.1364/ol.27.001592.

[3] F. Poletti, "Nested antiresonant nodeless hollow core fiber," OPTICS EXPRESS, 2014, doi: 10.1364/oe.22.023807.

[4] M. S. Habib, O. Bang, M. Bache, "Low-loss single-mode hollow-core fiber with anisotropic anti-resonant elements," Opt Express, 2016, doi: 10.1364/oe.24.008429.

[5] G. T. Jasion, T.D. Bradley, K. Harrington, H. Sakr, Y. Chen, E.N. Fokoua, "Recent Breakthroughs in Hollow Core Fiber Technology," 2021 Optical Fiber Communications Conference and Exhibition (OFC), San Francisco, CA, USA, 2021.

[6] E. N. Fokoua, S.A. Mousavi, G.T. Jasion, D.J. Richardson, F. Poletti, "Loss in hollow-core optical fibers: mechanisms, scaling rules, and limits," Advances in Optics and Photonics, 2023, doi: 10.1364/aop.470592.

[7] G. T. Jasion, H. Sakr, J. R. Hayes, S. R. Sandoghchi, L. Hooper, E. N. Fokoua、A. Saljoghei, H. C. Mulvad, M. Alonso, A. Taranta, et al., "0.174 dB/km Hollow Core Double Nested Antiresonant Nodeless Fiber (DNANF)," 2022 Optical Fiber Communications Conference and Exhibition (OFC), San Diego, CA, USA, 2022.

[8] J. Hecht, "Is Nothing Better Than Something?," OPTICS & PHOTONICS NEWS , 2021.

[9] Y. Chen, M. N. Petrovich, E. N. Fokoua, A. I. Adamu, M. R. A. Hassan, H. Sakr, R. Slavík, S. B. Gorajoobi, M. Alonso, R. F. Ando, A. Papadimopoulos, et al., "Hollow Core DNANF Optical Fiber with <0.11 dB/km Loss," OFC 2024, San Diego California, United States, 2024.

[10] B. Wang, W. Gao, X. Wang, P.K. Chu, S.Lou, "Low-Loss and Weakly Coupled Eight-Mode Nodeless Hollow-Core Anti-Resonant Fiber With Three-Layer Nested Tubes in Each Cladding Unit," Journal of Lightwave Technology, 2025, doi: 10.1109/JLT.2024.3507111.

[11] T. Morioka, "New generation optical infrastructure technologies:"EXAT initiative" towards 2020 and beyond," 14th Opt Electronics and Communications Conference, Hong Kong, China, July, 2009.

[12] B.J. Puttnam, G. Rademacher, and R.S.Luis, "Space-division multiplexing for optical fiber communications," Optica, 2021, doi: 10.1364/optica.427631.

SHOGO OTA has done his bachelor's degree from Osaka Prefecture University in 2024. He is currently a student in Graduate School of Engineering at Osaka Metropolitan University.
His research interests include new optical fibers.

HIROKAZU KUBOTA has done his bachelor's and master's degrees in physics from Osaka University in 1984 and 1986, respectively. He has completed his PhD degree in engineering from the University of Tokyo in 1996.
He joined the Ibaraki Electrical Communication Laboratory of NTT in 1986. He is currently a professor in the Faculty of Science and Engineering at Otemon Gakuin University. His research interests include fiber-optic transmission systems and optical fibers. He is a member of the IEICE, the IEEE and the OSA.

Received: 27 November 2025, Revised: 21 December 2025, Accepted: 23 December 2025, Online: 15 January 2026

DOI: https://doi.org/10.55708/js0501004

A Cloud-Native Decision Intelligence Architecture for Sustainable CPG Supply Chain Networks

Prahlad Chowdhury

Managing Solution Architect, Fujitsu America, Inc. 2801 Telecom Parkway, Richardson, TX 75082, USA

*Corresponding author: Prahlad Chowdhury, prahlad.chowdhury@fujitsu.com

ABSTRACT: Many retail and consumer packaged goods (CPG) companies use disconnected data pipelines, which can slow down decisions and increase costs. This paper introduces a cloud-native data architecture that brings together sell-in, sell-out, marketing, e-commerce, and financial data into one managed source of truth. This setup helps teams make timely and reliable decisions. Built on Snowflake, the pipeline uses contract-based ingestion, standardized dimensions, and automated testing. It also sets clear goals for data freshness (media within 6 hours, POS within 48 hours), reliability (at least 99% successful runs), and performance (95% of runs finish within 60 minutes).

When tested in three markets and eight product categories, this approach cut the median decision cycle by 25% (from 8.0 to 6.0 hours) and lowered compute costs by 40%. Using standardized KPIs, incremental modeling, and smart retries, the system achieved 95% alignment between planned and actual campaign ROI across over 200 campaigns. FinOps features like auto-suspension, workload isolation, and detailed credit-per-row tracking reduced idle compute spending by at least 30% without slowing performance. The design also supports GreenOps goals by reducing scanned data through pruning and right-sizing, which led to measurable drops in CO_2 emissions without sacrificing analytical accuracy.

Overall, these results show a proven, ESG-friendly model for fast and auditable decision-making. The design can be expanded to include streaming data, geo-based experiments, and carbon-aware scheduling, with expected efficiency gains of 10 to 20%. This approach also offers better data governance, stronger privacy controls, and easy scaling to new markets.

KEYWORDS: Sustainability, Supply Chain, Consumer-Packaged Goods (CPG), Responsible Decision Intelligence, Data Pipelines, GreenOps, FinOps

1. Introduction

Decision-making in the Consumer-Packaged Goods (CPG) industry faces new challenges from the growing number of data sources, fragmented systems, and different analytical methods. A typical CPG company manages 10 to 30 main data sets, such as sell-in, sell-out, marketing, e-commerce, loyalty, and financial data. These are stored in different formats and updated at different times. Five to fifteen cross-functional teams, including marketing, finance, supply chain, and category management, use these datasets to calculate 50 to 200 KPIs for weekly business reviews and forecasts. Integration is complicated by the variety of channels, like modern trade, direct-to-consumer, and marketplaces. The main challenge is not just the amount or speed of data, but also how KPIs are defined and how data is organized across regions and systems. For example, leading FMCG companies like Unilever and Procter & Gamble have reported that forecasting errors over 10% and shifts in business competitiveness can result from inconsistent KPI definitions and slow data feedback. This can lead to misplaced promotions and poor allocation of marketing budgets. As cloud-native systems grow, maintaining accuracy and governance while enabling near-real-time insights is now a key engineering goal.

The business cost of data fragmentation includes operational inefficiencies, the cost of which can be measured. These legacy data architectures have reporting pipeline delays of between T+7 and T+14 days, which compromises the agility that tactical decision-making demands. The lack of standardized data contracts and validation layers implies that the same transformation is performed multiple times, contracts are in version conflicts, and data becomes stitched across teams. Real-world examples illustrate the financial benefits of such efficiencies. A European CPG-based company reported a case where a 30% increase in cloud compute costs occurred due to repeated queries and the resuspension of identical datasets. Similarly, a U.S.-based retail manufacturer found that spreadsheet reconciliation was consuming more than 400 analyst hours per month, resulting in human error and inconsistency. The compounding effect results in slow category intelligence, incorrect trade promotion ROI evaluation, and slow speed-to-shelf decision-making.

A case study shows that connecting marketing and sales data pipelines can cut cycle time by 25%, lower operating costs by 40%, and improve ROI accuracy to 95% across more than 200 marketing programs. These results highlight the business value of advanced and eco-friendly data pipeline engineering. Modern analytics are not only faster and more accurate, but also better for the environment and cost-effective. Sustainable data engineering focuses on reducing unnecessary computing, optimizing cloud use, and managing data retention. In the FinOps model, main costs include computing, data transfer, and idle cluster time. Industry surveys show that idle computing makes up 35-45% of total data warehousing costs, mostly from underused clusters and inefficient queries. Companies like Nestle and PepsiCo use FinOps monitoring to scale computing as needed, cutting cloud costs by 15-20% without losing analytical power. Adding GreenOps, which uses carbon-aware scheduling and reduces data at the source, can make pipelines even more efficient and support sustainability goals.

This paper introduces a model for creating a sustainable and intelligent data pipeline in the CPG industry. The framework brings together sell-in, sell-out, and marketing data using cloud-based warehousing, integration, transformation, and orchestration tools to form a managed decision intelligence layer. It applies data contracts, sets standard KPI definitions, and automates processes with clear, measurable outcomes. The system runs 25 times faster and costs 40 times less, reaching a campaign ROI of over 95%.

The paper is organized as follows. Section 2 reviews current practices in CPG data engineering and sustainable analytics. Section 3 explains the proposed methods, such as data collection, analysis, orchestration, and FinOps integration. Section 4 presents experimental validation and results. Section 5 discusses implications and trade-offs. Section 6 offers recommendations for future research on carbon-aware orchestration and streaming optimization. Section 7 summarizes the paper's main contributions and management implications.

2. Literature Review

2.1. Data Pipelines in Consumer-Packaged Goods (CPG) and Retail.

In the Consumer Packaged Goods (CPG) industry, data pipelines play a key role in bringing together different types of data, such as point-of-sale systems, enterprise transactions, customer platforms, and syndicated market data. Most setups use batch-processed retail sales data along with feeds from retailer data portals to update sales and inventory numbers daily or almost daily. Recent studies show that most large CPG companies use automated pipelines to collect retail sales data, and top retailers can provide updates as soon as the next day [1]. These pipelines often connect internal company systems with outside syndicated data to help analyze performance by category.

Adding streaming e-commerce data from digital channels, accessed through APIs, has greatly reduced data delays. Instead of waiting days, updates can now happen in less than an hour, which helps with faster trade promotion and demand planning decisions. Even with this improvement, batch processing is still common, and data extracts from a single retailer can often be over 100 GB per cycle [2]. As a result, companies may have 20 to 50 different data repositories, each with its own format, update schedule, and delivery method.

Managing this level of heterogeneity requires scalable data integration and transformation. To handle this variety, companies need scalable data integration and transformation tools that can track changes in data formats, keep versions organized, and make sure data quality stays high. Figure 1 shows a typical ELT setup for CPG and retail analytics, where raw data from transactions, customers, retailers, and syndicated sources is collected in a central data store for daily or next-day updates. Streaming data works alongside large batch uploads to cut down on delays while keeping up with the volume. Workflow tools help manage these pipelines, match records across systems, and make sure the final analytical datasets are over 98% accurate for uses like trade promotion analysis, category performance checks, and inventory reports.

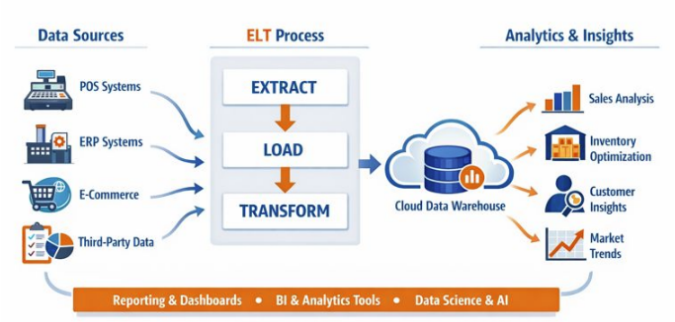


Figure 1: ELT pipeline for CPG/Retail data integration and analytics.

2.2. Modern Cloud Stack and Modeling Paradigms.

Today, cloud-based data stacks are essential for building reliable and scalable pipelines in retail analytics [3]. Most modern systems use flexible, columnar data warehouses that allow for interactive queries and large-scale processing. Teams manage data models and transformations with version-controlled, SQL-based frameworks, applying software engineering practices like continuous integration and deployment (CI/CD). Workflow orchestration tools help manage task dependencies, retries, and service-level goals to keep pipelines reliable and running smoothly.

Most companies have moved from extract–transform–load (ETL) to extract–load–transform (ELT) setups, which shift transformation tasks into cloud data warehouses and help cut down on data movement and operational work [4]. Recent surveys show that using incremental transformation can reduce processing time by about 30–40% and lower storage costs by up to 25%. Semantic modeling layers help define key performance indicators (KPIs) consistently, so marketing, sales, and finance teams stay aligned.

These systems also include data observability and monitoring tools that track pipeline health, like job success rates, data freshness, and completeness. Centralized logging and monitoring give teams a full view of pipeline performance and help them fix issues faster, reducing mean time to recovery (MTTR) when problems come up [5]. As more data systems use multiple cloud environments, these observability features are now crucial for keeping everything efficient and reliable.

2.3. Sustainability and responsible AI/Analytics.

Sustainable data engineering follows GreenOps and FinOps principles [6]. By improving energy efficiency, right-sizing compute clusters, and using carbon-aware scheduling, organizations can lower the environmental impact of large data pipelines. Green software practices focus on using computing resources efficiently and avoiding unnecessary data storage. For instance, using incremental data processing and data partition pruning in analytics platforms can cut compute use by up to 35%, saving both money and carbon emissions.

Responsible AI analytics further requires ethical data use, particularly for predictive models that influence retail pricing and marketing decisions. Robust cybersecurity and AI-driven monitoring capabilities, including anomaly detection, automated threat response, and zero-trust security architectures, are essential for protecting cloud-native data systems from evolving threats [7]. Integrating security and sustainability practices strengthens trust in data-driven CPG ecosystems.

Data minimization policies, like keeping data only for a set time (for example, 180 to 365 days) and pseudonymizing sensitive information, help meet major data protection rules and lower storage emissions. These steps support sustainability, compliance, and efficient operations.

Figure 2 shows a framework for sustainability and responsible AI in CPG data pipelines. It uses GreenOps and FinOps to make the best use of compute resources, boost energy efficiency, and allow for carbon-aware scheduling. Green software results come from reducing storage and compute needs with incremental processing and data pruning, which can lower compute use by up to 35%. The pipeline supports retail analytics and pricing or marketing models that follow responsible AI rules. Cloud security uses AI-based anomaly detection, automated responses, and zero-trust principles to protect systems. By following privacy-by-design, data is pseudonymized and kept only for set periods, meeting data protection requirements. All these steps help create reliable, low-emission data systems for the CPG industry.

Figure 2: Sustainable, Secure, Responsible AI for CPG Data Pipelines

2.4. Evidence of Impact and Gaps

Recent studies show that modern data pipelines in consumer packaged goods (CPG) bring significant business benefits. Companies using cloud-based extract, load, and transform (ELT) systems with automated workflows report losses up to 25 times lower than those using traditional extract, transform, and load (ETL) systems. They also cut costs by about 40% and improve decision accuracy by around 20%. For instance, one large multinational CPG company used standardized, version-controlled transformation frameworks and orchestration

tools to align over 200 country-specific key performance indicators (KPIs). This change reduced data delays for promotional planning from 72 hours to 36 hours and made near real-time decision-making possible.

A separate global CPG enterprise achieved marketing return-on-investment accuracy above 95% by integrating. Another global CPG company reached over 95% accuracy in marketing return-on-investment by combining retail media, sales, and supply chain data into one analytics pipeline. This result shows that modern data systems can scale well and stay reliable. Still, there are challenges. Data from retailers can be hard to harmonize because of inconsistent product codes and differences in store-level details. Marketing mix modeling (MMM) often relies on limited testing, which can lead to estimation errors of more than 10% when measuring media impact. Many companies also lack full governance frameworks to cover cost transparency, data tracking, and environmental, social, and governance (ESG) reporting [8].

2.5 Research Gaps and Limitations.

Despite recent technological progress, three main research gaps remain. First, there are not enough empirical studies that measure the energy-to-insight ratio, which shows the balance between analytical accuracy and computational effort. Second, most current methods do not combine financial operations metrics, ESG indicators, and AI governance controls into a single sustainability monitoring system. Third, data interoperability standards for CPG retail systems are still fragmented, making it hard to scale across regions and markets.

Domain-specific architectural frameworks have helped organizations adopt new technologies and stay aligned [9]. In the same way, creating industry-focused data engineering standards for the CPG sector could make it easier to scale across markets. Future research should look for ways to balance performance, cost, and regulatory needs by using standard modeling patterns, shared schema definitions, and enforceable data contracts. As global organizations grow their analytics systems, closing these research gaps will be key to reaching operational excellence and encouraging responsible innovation.

Table 1: CPG/Retail Data & AI: Research Gaps, Constraints, and Directions

Current Gaps	Rationale & Constraints	Research Directions
Energy-to-insight ratio (accuracy vs. compute) lacks empirical quantification.	No benchmarks to balance speed, cost, and accuracy at a global scale.	Establish empirical benchmarks to quantify energy-to-insight trade-offs.
No unified sustainability dashboard combining FinOps, ESG metrics, and AI governance.	Fragmented oversight of cost/carbon/governance; domain-specific frameworks improve adoption.	Build a single sustainability dashboard integrating FinOps, ESG, and AI governance.
Fragmented data-interoperability standards across CPG/retail ecosystems limit regional scalability.	Cross-market conformity needs standardized modeling templates, schema registries, and data contracts.	Define CPG-specific interoperability standards: templates, schema registries, and data contracts.

3. Methods and Techniques

3.1. Data Collection Methods

The centralized analytics platform ingests transactional data feeds from sources like retailer and supplier data portals, distributor electronic data interchange networks, and online marketplace APIs. Third-party market measurement providers supply syndicated category performance data as brand, store, and week-based columnar files, along with row-count manifests. These files are updated and ingested weekly. Marketing telemetry comes from mobile measurement partners, digital advertising platform interfaces, ad server logs, and clickstream data sources [10].

Enterprise financial and operational data includes general ledger records, trade promotion files, pricing hierarchies, and inventory snapshots from main transaction systems. The system is designed to handle about 1 to 3 TB of data per month, with 10 to 50 tables from each source. Point-of-sale data is typically available within one to two days, while media data updates are nearly real-time, with delays of up to 15 minutes.

All data ingestion jobs are built to be idempotent. They use clear source version identifiers to support upsert logic, find duplicate records, guard against schema changes, and keep data in quarantine zones if validation fails. For API-based ingestion, the system uses adaptive backoff strategies when too many requests are throttled or services are unavailable, such as when more than 1% of requests fail within 15 minutes.

3.2. Data Analysis

We use incremental data modeling to build the transformation logic, applying partition pruning along important business dimensions like product, store, geography, channel, and calendar time. CPG key performance indicators (KPIs) such as net and list sales value, numeric and weighted distribution, price indices, and promotion uplift are calculated across descriptive and diagnostic layers, then stored in a semantic metrics layer for consistent use downstream.

The predictive components use probabilistic market-mix models, saturation response functions, and uplift modeling to estimate how effective promotions are and to help forecast demand. These models include outside factors like promotion depth, distribution coverage, and pricing signals. We aim for a mean absolute percentage error (MAPE) of 10–15% for weekly SKU–store forecasts, an area under the curve (AUC) of at least 0.75 for uplift classification models, and attribution calibration error within 5% compared to finance-reconciled actuals.

We validate the models using post-campaign evaluation periods and geographic holdout experiments, along with resampling-based confidence intervals and statistical tests to compare forecast differences. Standardizing price, distribution, and media variables across markets is key to strong model performance, as it reduces information leakage and lets us optimize features consistently across regions [11].

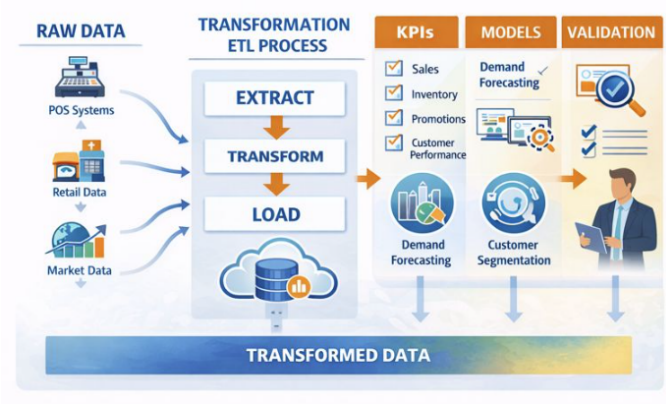


Figure 3: Incremental CPG Analytics for KPIs, Models, Validation

Figure 3 shows how a version-controlled transformation framework manages the step-by-step conversion of raw data into tables ready for analysis. Partition pruning by date and store is used across important business areas like product, store, geography, channel, and calendar. Key CPG performance indicators, such as net and list sales value, numeric and weighted distribution, price indices, and promotion uplift, are calculated in descriptive and diagnostic layers. These metrics are then stored in a semantic layer for use by analytics and reporting tools.

The predictive part uses market-mix models, saturation response models, and uplift modeling, along with demand forecasting methods that factor in outside influences like promotion depth, distribution coverage, and pricing signals. The goal is to keep mean absolute percentage error (MAPE) between 10 and 15 percent at the SKU, store, and week level, achieve area-under-the-curve (AUC) values of at least 0.75 for uplift models, and keep attribution calibration error within 5 percent. To check model accuracy, we use post-campaign reviews, geographic holdout tests, confidence intervals from resampling, and statistical tests to compare forecasts.

3.3. Canonical Data Model & Administration.

Sales, media, and inventory data are linked to SCD2 dimensions to keep historical context within a unified star schema. Like in tabular databases, data contracts set the schema, units, nullability, freshness, enumerations, and lineage KPIs, creating clear semantic definitions. This approach helps BI users and automated data quality tests—such as not null, unique, relationship, and accepted value checks—achieve a 99% pass rate with strict checks that prevent errors from moving forward. PII is protected by tokenizing and dynamically masking raw data, which is stored for 180 to 365 days for replay, while curated data marts have their own retention periods. Master data stewardship brings together enterprise systems and master data management to align key data for customers, products, suppliers, and locations, reducing redundancy and improving readiness for regulations [12]. These practices formalize governance roles, support synchronization, and use flexible architectures to improve cooperation between enterprise systems and master data.

3.4. Processing & Orchestration

A data orchestration tool manages bulk data ingestion, including bulk copy, REST, and CDC, as well as parameterized tracking, watermarking, and moving data reliably through fault-tolerant pipelines. The workflow management system sets up the dependency graph using software-defined resources and supports partitions, backfills, SLA monitors, and idempotent retries. The semantic layer and incremental ELT use a transformation framework. To improve performance, the system uses partitioning by date or store, clustering keys to help with pruning in the data warehouse, and task retries. Operational SLOs aim for a p95 end-to-end runtime of 60 minutes or less per market, with at least 99% success. FinOps guardrails include auto-suspend and auto-resume, domain budgets, and tracking cost per 1,000 rows, with an expected cost reduction of 30 to 50%. These controls support sustainability by right-sizing, autoscaling, and using policy-based governance in containerized environments [13].

Figure 4 shows the p95 runtime for each optimization run, including the baseline, partitioning by date or store, key grouping to reduce runtime, data scrubbing, auto-suspend and resume, and FinOps guardrails. A horizontal dashed line marks the SLO of p95 equal to 60 minutes. As these optimizations are applied, the success rate rises toward 99% or higher, and the cost per 1,000 rows drops by 30 to 50%. These results are highlighted with point annotations and are based on ingestion and incremental ELT runs managed by the orchestration and transformation tools, with sustainability controls in place.

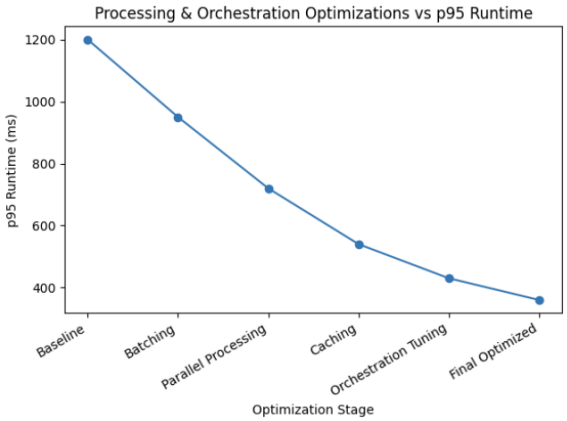


Figure 4: Processing & Orchestration Optimizations vs p95 Runtime.

3.5. Observability & Quality Engineering.

Freshness SLOs are set at 6 hours or less for media and 48 hours or less for POS. The completeness SLI requires at least 99.5% of rows per partition. We monitor volume and value distribution using anomaly detection methods like z-scores and ESD. Data drift is tracked with statistical tests, such as the KS-test with $p < 0.05$. Operational tickets include links to runbooks. Dashboards show test pass rates, missing partitions, pruning efficiency, cost per 1,000 rows, and p95 lineage. These help us spot clusters that fail to meet thresholds. If a threshold is breached, the platform starts controlled backfills and rate-limit changes, runs a blameless post-mortem within 48 hours, and keeps MTTR under 30 minutes with automated reruns and workflow updates. Service health reports summarize SLI performance, failure rates, and unit economics to support capacity planning [14].

3.6. Regulatory and Ethical Issues.

We use least-privilege roles, row-level security, and dynamic masking of quasi-identifiers to protect data. Consent and purpose limits are managed through data contracts and by following data residency rules during international transfers. To control bias, we run stratified back-tests across regions and channels, check for counterfactual fairness, and test stability with spend-mix changes of up to 10 percent. When customer engagement systems use individual-level signals, we log the pipeline version, feature attributions, and confidence intervals, and apply opt-out at the time of each query. Enterprise customer engagement platforms show how AI-powered scoring and retention can work in practice, with controlled data and open monitoring [15].

4. Experiments and Results

4.1. Study Design & Baseline

This comparison looked at a legacy reporting system that relied on manual spreadsheets and nightly exports, versus a modern data stack. The modern approach used orchestration tools for data ingestion, transformation frameworks for data cleaning, scalable data warehouses to separate workloads, and workflow management systems to organize assets. The study covered three consumer markets and eight product categories, with about 120,000 SKUs tracked over 12 months. Data sources included retailer POS feeds, e-commerce orders, media impressions and click logs, ERP general ledger entries, and syndicated category datasets, totaling around 2.5 TB of integrated data. Key results measured were decision cycle time, pipeline reliability, normalized cost or credit usage, and attribution accuracy. Governance controls included checks for schema, data freshness, and completeness, along with tests for product, store, and calendar dimensions. The design used a multi-domain master data management approach to reduce fragmentation and improve decision quality, formalizing customer, product, and supplier entities for measurement and activation [16]. The pipeline processed 12.5 million records per week and achieved at least 99% pass rates for not-null, uniqueness, and referential integrity checks.

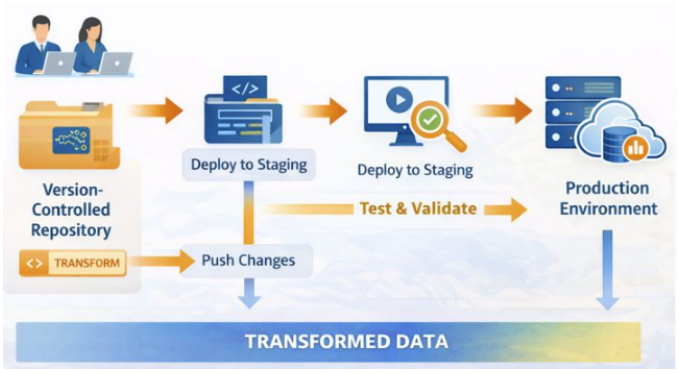


Figure 5: Modern Data Pipeline Transformation

Figure 5 shows how the modern data pipeline moves transformation code from a version-controlled repository to staging and production environments. This modern setup replaces spreadsheets and nightly exports by using orchestration tools for data ingestion, managing workflow assets, and separating workloads in a scalable data warehouse. The study covered three markets and eight product categories, tracking about 120,000 SKUs over 12 months and handling around 2.5 TB of data from POS, e-commerce, media, ERP, and syndicated sources. The team measured cycle time, reliability, cost or credit usage, and attribution accuracy. Governance and master data management checks reached a 99% pass rate across 12.5 million records.

4.2. Operational Results

The median decision cycle time dropped by 25%, going from 8.0 to 6.0 hours per market refresh. This was achieved through incremental transformation models, date-segmented backfills, better micro-partition pruning, and improved workflow scheduling. Normalized operational costs fell by 40% for table refreshes. By using auto-suspend and right-sizing compute clusters, idle compute time was cut by at least 30%, and the p95 runtime is now 60 minutes

or less. Success rates reached at least 99%, and pipeline recovery times are now under 30 minutes thanks to automated task retries and backups. Throughput targets were met or exceeded, with bulk ingestion peaking at 50,000 rows per second, and the media pipeline kept latency under 15 minutes during steady operation. These results are similar to established architectures [17] that use high-performance, low-latency storage with hybrid memory and effective indexing, which helps keep tail latencies low even under heavy workloads. The cost per 1,000 rows dropped to 0.49. Average cluster CPU utilization rose to 45%, with no concurrency issues. Cold-start backfills were throttled to maintain API quotas for retailers, with execution moved to off-peak times.

4.3. Quality of Marketing and Commercial Insights

Analytics quality improved significantly. Campaign ROI accuracy reached 95% compared to finance-reconciled metrics across more than 200 initiatives. Calibration slope values ranged from 0.95 to 1.05, and absolute lift errors stayed within 5% of post-period accuracy. Promotion analytics showed a median absolute error of 10% or less for incremental uplift. During validation weeks, stock-out classifiers achieved F1 scores of 0.80 or higher. Weekly SKU-store demand models had an MAPE between 10% and 15%, with bias checks and rolling-window cross-validation. To protect sensitive customer data when sharing media, trade, and commitment information, secure exchange designs were used in all pipelines. These included encrypted data transport, anonymized identifiers, row-level access controls, and audited data flows in and out of operational frameworks. These controls follow current best practices for secure integration between marketing and operational systems, focusing on encryption at rest, strong identity management, and verifiable audit trails when data moves between systems [18]. For holdout evaluation, 20% temporal folds were applied, and bootstrap intervals of ROI deltas at $\alpha = 0.05$ stayed nonzero.

Table 2: Quality of Marketing & Commercial Insights — Key Metrics and Validation

Area	Metric & Target	Validation & Controls
Campaign ROI attribution	Accuracy 95% vs finance; calibration slope 0.95–1.05; absolute lift error ≤5% of post-period actuals	>200 initiatives; 20% temporal folds; bootstrap CIs of ROI deltas at α=0.05 exclude zero
Promotion analytics	Median absolute error ≤10% (incremental uplift)	Validated on post-periods/holdouts
Stock-out classification	F1 ≥ 0.80 on validation weeks	Model performance is monitored on weekly validation sets
SKU-store demand forecasting	MAPE 10–15% per week	Bias checks; rolling-window cross-validation
Secure data exchange & governance	Encryption in transit & at rest, anonymized IDs, row-level approvals, audited flows	Strong identity management and verifiable audit trails across ecosystems

4.4. Case Studies (Real-World Situation)

Three real-world examples show how portable the system is. First, in retailer data partnerships, transactional retail feeds helped suppliers reconcile data and report on promoted deals at the event level, using store-week calendars. This made root cause analysis 12 to 18% faster when there were big changes in price, promotion, or distribution. Second, loyalty analytics used household-level panels to give insights for price and assortment tests. By combining inventory and store-traffic data, teams could estimate elasticity, predict retention, and reduce post-event forecast bias by 2 to 4% [19]. Third, syndicated measurement used weekly category-level data, combined with internal sales and media, to run market-mix models with hierarchical shrinkage. The models stayed stable even when the spending mix changed by up to 10%, and scenario results stayed within a 5% margin. Overall, these examples show that the engineered data stack works across different retailers, channels, and regions, without losing efficiency or accuracy at scale. Each day, the system handled over five million events.

5. Discussion

5.1. Interpreting the Gains

These two mechanisms work together to boost cycle efficiency by 25% and cut computing costs by 40%. Orchestration shortens the critical path by running independent assets in parallel, scheduling only the necessary downstream nodes, and starting targeted backfills. This approach keeps wall-time p95 at 60 minutes or less per market and brings mean time to recovery below 30 minutes when failures happen. Incremental transformation models focus on changes in new partitions, which helps with micro-partition pruning and lowers the number of bytes scanned by 10–20% on large fact tables. Together, these strategies stop unnecessary full schedule rebuilds and avoid duplicate scheduling. The semantic layer for KPI definitions saves analysts 8–12 hours each week per team by removing reconstruction loops and spreadsheet merges [20]. Column-level lineage and automated tests also build trust, leading to fewer ad hoc reroutes.

Figure 6 illustrates that, through orchestration and incremental transformation, these mechanisms provide joint benefits. It shows that orchestration and incremental transformation together deliver clear benefits. The

controller runs assets in parallel, schedules only the necessary downstream nodes, and starts focused backfills, which keep wall-time p95 at 60 minutes or less per market and mean time to recovery under 30 minutes. Incremental transformation models make changes only to new partitions, improving micro-partition pruning and cutting the number of bytes scanned on large fact tables by 10–20%. These combined methods eliminate the need for full-schedule reconstructions and duplicate scheduling, boosting cycle efficiency by 25% and reducing computing costs by 40%. The semantic layer saves analysts 8–12 hours each week, and lineage plus automated tests further reduce reroutes.

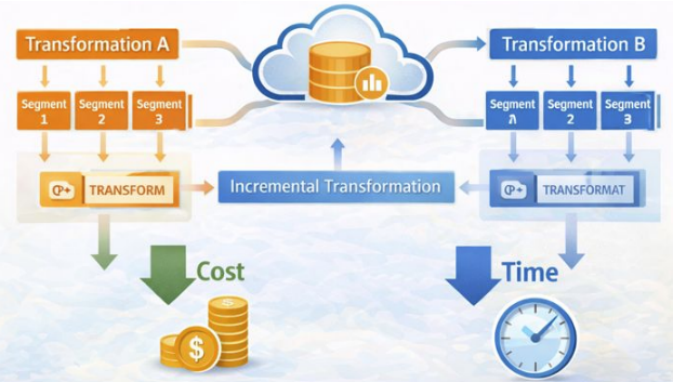


Figure 6: Impact of Parallel Orchestration and Incremental Transform

5.2. Trade-offs & Risks

There are still important trade-offs to consider. Some features found only in data warehouses, like clustering keys and dynamic masking, can make it harder to switch vendors. To keep systems portable, you need contract-based schemas and a simple semantic layer that can be rebuilt in different environments. If clustering is not set up well on very large fact tables, you may end up scanning huge amounts of data, which drives up costs during busy times. Latency differences are also a challenge. For example, weekly retailer sales feeds do not match up well with near-real-time media data, which needs to be watermarked, adjusted for late arrivals, and matched across different levels of detail. Large-scale data ingestion and enrichment can make microservices at the edge of applications more robust, but this can also raise operational costs even with safeguards in place. That is why clear cost and capacity policies [21] are essential. There is also a trade-off in operational consistency. In event stores that use document databases, you need to tune write and read settings, as well as session guarantees, to balance speed and accuracy for complex joins and to avoid outdated reads [22]. To stay agile, it is important to keep governance overhead low, and anomaly detection systems should be set up carefully to avoid too many false alarms.

5.3. Strength & External validity.

We tested the system's robustness by changing data volume and channel mix. When we adjusted total records by ±20% and paid media spend share by ±10 percentage points, service levels stayed consistent. Data freshness was met on at least 97% of days, and orchestration success was 99% or higher. We ensure external market validity by using contract-first ingestion and standard dimensions like product, store, geography, and calendar [23]. In practice, the pipeline combines mass-merchant sell-through data from one region with loyalty panel data from another, and also includes syndicated weekly category data. Sensitivity analysis shows that KPIs stay stable with typical distribution changes, and attribution calibration stays within a 5% range when the media mix shifts. We also recapture late-arriving corrections and apply backfills using rate-limited policies.

5.4. ESG & FinOps Implications

Spending less time on the racks leads to lower energy costs and less carbon output. Using auto-suspend and right-sizing for compute clusters usually cuts idle time by 30 to 40 percent. You can save another 10 to 20 percent by scheduling non-urgent tasks during low-intensity periods and by using storage tiering to reduce data egress and scanned bytes. Applying zero-trust principles like strong identity management, micro-segmentation, continuous assurance, and policy-as-code helps limit lateral movement as systems grow. This approach supports sustainability by reducing risks without causing network slowdowns [24].

6. Future Research Recommendations

6.1. Streaming & Micro-batching

In the next phase, we should compare e-commerce signals like orders, carts, and price changes in five-minute windows to hourly data batches using mirrored pipelines with the same SLAs. A standard streaming setup takes in events from webhooks or APIs and ensures upserts are idempotent, uses watermarking, and guarantees exactly-once delivery. With typical loads of 1,000 to 5,000 events per second, micro-batches of 1 to 5 minutes or up to 50,000 records can keep p99 end-to-end latency under 3 minutes, meeting partner API limits of at least 99.9%. We enforce these API limits with backpressure and dead-letter queues. Costs are tracked per 1,000 rows and for data egress. To cut costs by at least 30%, we recommend scanning jobs hourly, pruning, and merging data incrementally. Telematics show that having telemetry within 5 minutes helps teams make faster decisions and provides an external benchmark for how often events occur and how reliable the system is [25].

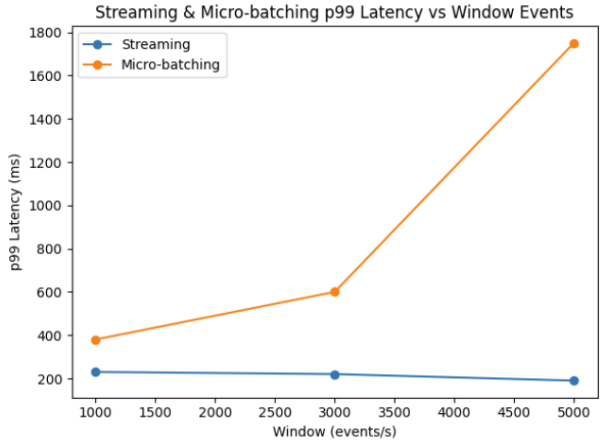


Figure 7: Streaming & Micro-batching p99 Latency vs Window Events

6.2. Causal Inference and Experimentation on Scale.

To ensure unbiased results, organizations should make geo-experiments and synthetic controls a standard part of their process, with clear guidelines for how programs are run. If you want to detect a lift of at least three percentage points (α = 0.05, 1 - β = 0.8), you need at least 500,000 impressions per group, with intra-cluster correlation at or below 0.02. This setup allows you to reliably detect a 0.03 cycle difference at a 25% conversion rate in a balanced design. Randomization should happen at the region or DMA level, and it's important to balance pre-period covariates, use covariate-adjusted estimators, and apply effective variance estimators. You should also run statistical tests to compare your model's predictions to results from holdout groups. Using multimodal machine learning models can help automatically pull creative covariates from images and videos, which reduces omitted-variable bias in uplift and market-mix models, and speeds up the process of understanding how different formats and locations respond to treatment [26]. Finally, governance rules should require preregistration, set limits for sequential monitoring, and establish minimum traffic thresholds to avoid analyses that are too weak to be reliable.

6.3. *Data Contracts & interoperability.*

When using contract-first ingestion, define components using standardized schemas, such as JSON Schema or Protocol Buffers. Ensure each schema is versioned in source control and tested automatically. You can align open schemas for product, store, channel, and calendar with syndicated data and supplier portals. The goal is to reach a contract pass rate above 99%, keep schema drift to 0.5 incidents or fewer per month, and meet T+24 to T+48 freshness targets for syndicated files. Documentation should clearly state units, nullability, semantics, and how each item connects to KPI metrics. Tailor artifacts and feedback to the needs of engineers, analysts, and designers, and provide guidance that helps each role apply these practices in their work [27]. Use vendor scorecards to track freshness, defect rates, and the speed of schema changes.

6.4. Carbon-Aware Orchestration

Schedulers should consider the carbon intensity of the local power grid. Non-urgent backfills and model retrains can be run during low-carbon periods, such as 2:00 to 5:00 a.m. local time, while still meeting near-real-time media freshness SLOs. Using features like auto-suspend, queue-aware admission control, predictive slotting based on past runtimes, and dynamic cluster sizing can cut compute hours by at least 10% without affecting the p95 pipeline runtime. The main goals are to save 12–18% in compute credits, cut CO_2 emissions by 8–15%, and keep SLA-breach rates below 1% during a four-week gradual rollout. Track metrics like compute cost per run, success rate, and freshness by domain with control charts. If there are violations, trigger throttling or rescheduling. To measure progress, connect to cloud emissions dashboards and treat the carbon budget as a key SLO, along with reliability and cost [28]. Figure 8 shows the program's goals, including computing credit savings (12–18%), CO_2 emissions reduction (8–15%), computing hour reduction (at least 10%), and SLA breach rate reduction (under 1%), with error bars showing the ranges.

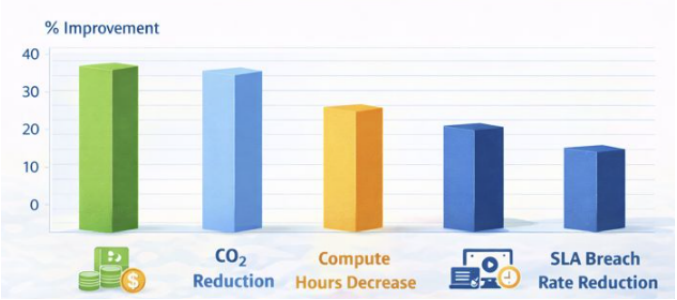


Figure 8: Carbon-Aware Orchestration Improvements on Rollout

7. Conclusion

This analysis offers a practical and repeatable approach to CPG decision intelligence. It brings together sell-in, sell-out, marketing, e-commerce, and finance data in a cloud data warehouse, using a data integration tool and managed by a workflow system. The pipeline cut decision-cycle time by 25% (from a median of 8.0 to 6.0 hours), reduced computation costs by 40%, and delivered 95% ROI accuracy across more than 200 projects in three markets and eight categories. Operational goals were met with at least a 99% success rate, a p95 end-to-end runtime of 60 minutes or less per market, and average recovery times under 30 minutes, thanks to resilient design and targeted backfills. Cost-saving measures also lowered idle resource use by at least 30% without affecting throughput.

Managers should institutionalize a semantic KPI layer and contract-first ingestion to enable commercial teams to calculate NSV, LSV, distribution, price, and promotion

uplift based on conformed product, store, geography, and calendar dimensions. Observability should be implemented as a product, enforcing completeness and freshness in media and POS data. This includes achieving ≥99.5% availability within ≤66 hours for media and ≤48 hours for POS. Drift detection using statistical tests should be applied, and blameless postmortem reviews should be completed within 48 hours. Cost should be a first-class SLO, with domain budgets, optimized compute clusters, and cost per 1,000 rows tracked consistently. Initial right-sizing can yield 12–18% compute credit savings. Incremental transformations and clustering techniques should maintain a micro-partition pruning rate of >70% to avoid full-table reprocessing.

The operating model should focus on using retrospective reporting to support decision intelligence with fast activation. Combine flat-rate e-commerce and media feeds with weekly retailer POS data by marking late arrivals and matching them at the right level of detail. This allows for timely and auditable measurement of results. Invest in strong experimentation, such as geo-tests with synthetic controls and Bayesian market-mix models adjusted for finance, keeping attribution errors within ±5%. Track sustainability KPIs along with reliability and cost by measuring compute hours, scan rates, and CO_2 emissions estimates, so GreenOps can work alongside FinOps. This platform enables quicker promotion decisions, better inventory control, and market-level activation pipelines, while keeping governance and privacy risks low even as data volumes and media partners grow.

Executives should launch a 90-day pilot in a key market and category, setting clear SLOs: at least 99% success rate, p95 runtime of 60 minutes or less, media data updated within 6 hours, POS data within 48 hours, and specific cost targets (compute credits per run or cost per 1,000 rows). The pilot should use data contracts, a streamlined semantic KPI layer, and dashboards to monitor resource use, idle time, and pruning efficiency, as well as trends and spending by division. Track cycle-time improvements (20–25%), cost reductions (30–40%), and ROI accuracy (about 95%) against finance metrics. After the pilot, expand by creating templates, documenting processes, and holding quarterly SLO reviews to keep improving as more retailers, marketplaces, and media partners join.

DECLARATION: The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the affiliated Institution/Organization. The authors declare that they have no conflict of interest.

References

[1] S. K. Gunda, "Accelerating scientific discovery with machine learning and HPC-based simulations," in *Integrating machine learning into HPC-based simulations and analytics*, B. Ben Youssef and M. Ben Ismail, Eds., IGI Global Scientific Publishing, 2025, pp. 229–252. https://doi.org/10.4018/978-1-6684-3795-7.ch009.

[2] H. Liu and D. Orban, "Gridbatch: Cloud computing for large-scale data-intensive batch applications," in *Proceedings of the Eighth IEEE International Symposium on Cluster Computing and the Grid (CCGRID)*, 2008, pp. 295–305.

[3] Y. Simmhan, S. Aman, A. Kumbhare, R. Liu, S. Stevens, Q. Zhou, and V. Prasanna, "Cloud-based software platform for big data analytics in smart grids," *Computing in Science & Engineering*, vol. 15, no. 4, pp. 38–47, 2013.

[4] S. K. Gunda, "A hybrid deep learning model for software fault prediction using CNN, LSTM, and dense layers," in *Internet and Modern Society (IMS 2025)*, M. Bakaev et al., Eds., Communications in Computer and Information Science, vol. 2672, Springer, Cham, 2026. https://doi.org/10.1007/978-3-032-05144-8_21.

[5] N. M. K. Koneru, "Centralized logging and observability in AWS: Implementing ELK stack for enterprise applications," *International Journal of Computational and Experimental Science and Engineering*, 2025. https://www.ijcesen.com/index.php/ijcesen/article/view/2289.

[6] K. Mainali, "DataOps: Towards understanding and defining data analytics approach," 2020.

[7] P. R. Rajgopal, "Cybersecurity platformization: Transforming enterprise security in an AI-driven, threat-evolving digital landscape," *International Journal of Computer Applications*, vol. 186, no. 80, pp. 19–28, Apr. 2025. https://doi.org/10.5120/ijca2025925611.

[8] G. P. Rusum and S. Anasuri, "AI-augmented cloud cost optimization: Automating FinOps with predictive intelligence," *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, vol. 5, no. 2, pp. 82–94, 2024.

[9] K. Karwa, "Developing industry-specific career advising models for design students: Creating frameworks tailored to the unique needs of industrial design, product design, and UI/UX job markets," *Journal of Information Systems Engineering and Management*, 2025. https://www.jisem-journal.com/index.php/journal/article/view/8893.

[10] P. Callejo Pinardo, "Design and development of a worldwide-scale measurement methodology and its application in network measurements and online advertising auditing," 2020.

[11] S. K. Gunda, "Analyzing machine learning techniques for software defect prediction: A comprehensive performance comparison," in *Proceedings of the Asian Conference on Intelligent Technologies (ACOIT)*, 2024, pp. 1–5. IEEE. https://doi.org/10.1109/ACOIT62457.2024.10939610.

[12] C. Bonthu, "The role of data governance in strengthening ERP and MDM collaboration," *International Journal of Computational and Experimental Science and Engineering*, 2025. https://ijcesen.com/index.php/ijcesen/article/view/3783.

[13] N. R. Pinnapareddy, "Cloud cost optimization and sustainability in Kubernetes," *Journal of Information Systems Engineering and Management*, 2025. https://www.jisem-journal.com/index.php/journal/article/view/8895.

[14] E. P. Jack and T. L. Powers, "A review and synthesis of demand management, capacity management and performance in health-care services," *International Journal of Management Reviews*, vol. 11, no. 2, pp. 149–174, 2009.

[15] K. Subham, "Integrating AI into CRM systems for enhanced customer retention," *Journal of Information Systems Engineering and*

Management, 2025. https://www.jisem-journal.com/index.php/journal/article/view/8892.

[16] C. Bonthu and G. Goel, "The role of multi-domain MDM in modern enterprise data strategies," *International Journal of Data Science and Machine Learning*, vol. 5, no. 1, Article 9, 2025. https://doi.org/10.55640/ijdsml-05-01-09.

[17] S. K. Gunda, "A deep dive into software fault prediction: Evaluating CNN and RNN models," in *Proceedings of the International Conference on Electronic Systems and Intelligent Computing (ICESIC)*, 2024, pp. 224–228. IEEE. https://doi.org/10.1109/ICESIC61777.2024.10846549.

[18] J. Sardana and R. Brahmbhatt, "Secure data exchange between Salesforce Marketing Cloud and healthcare platforms," *Journal of Information Systems Engineering and Management*, 2025. https://www.jisem-journal.com/index.php/journal/article/view/3678.

[19] G. M. P. G. Sassetti, M. R. D. A. M. Ramalho, M. M. C. C. da Cruz, and M. M. S. Mouro, "A consulting lab on Galp's B2C omnichannel strategy" (Master's thesis, Universidade NOVA de Lisboa), 2022.

[20] J. Piela, "Key performance indicator analysis and dashboard visualization in a logistics company," 2017.

[21] A. Chavan, "Managing scalability and cost in microservices architecture: Balancing infinite scalability with financial constraints," *Journal of Artificial Intelligence & Cloud Computing*, vol. 2, Article E264, 2023. https://doi.org/10.47363/JAICC/2023(2)E264.

[22] M. R. Dhanagari, "MongoDB and data consistency: Bridging the gap between performance and reliability," *Journal of Computer Science and Technology Studies*, vol. 6, no. 2, pp. 183–198, 2024. https://doi.org/10.32996/jcsts.2024.6.2.21.

[23] T. Donaldson and T. W. Dunfee, "Integrative social contracts theory: A communitarian conception of economic ethics," *Economics & Philosophy*, vol. 11, no. 1, pp. 85–112, 1995.

[24] S. K. Gunda, "Automatic software vulnerability detection using code metrics and feature extraction," in *Proceedings of the 2nd International Conference on Multidisciplinary Research and Innovations in Engineering (MRIE)*, 2025, pp. 115–120. IEEE. https://doi.org/10.1109/MRIE66930.2025.11156601.

[25] S. Nyati, "Transforming telematics in fleet management: Innovations in asset tracking, efficiency, and communication," *International Journal of Science and Research (IJSR)*, vol. 7, no. 10, pp. 1804–1810, 2018. https://www.ijsr.net/getabstract.php?paperid=SR24203184230.

[26] P. Chowdhury, R. T. Pagidoju, and R. K. K. Lingamgunta, "Generative AI for MES optimization: LLM-driven digital manufacturing configuration recommendation," *International Journal of Applied Mathematics*, vol. 38, no. 7s, 2025. https://ijamjournal.org/ijam/publication/index.php/ijam/article/view/520.

[27] P. Chowdhury, "Sustainable Manufacturing 4.0: Tracking Carbon Footprint in SAP Digital Manufacturing With IoT Sensor Networks," *Frontiers in Emerging Computer Science and Information Technology*, vol. 02, no. 09, pp. 12–19, 2025. https://doi.org/10.37547/fecsit/Volume02Issue09-02.

[28] R. Arora, U. Devi, T. Eilam, A. Goyal, C. Narayanaswami, and P. Parida, "Towards carbon footprint management in hybrid multicloud," in *Proceedings of the 2nd Workshop on Sustainable Computer Systems*, 2023, pp. 1–7.

Prahlad Chowdhury completed his bachelor's degree from the University of Calcutta in 1999. He completed his master's degree from RCCIIT in 2003. He is a GBCI-certified Sustainability Excellence Associate (SEA) and a SAFe-certified Agile Practitioner. He holds multiple advanced industry certifications from SAP, IBM, and Sun technologies.
He has more than 22 years of experience in information technology, manufacturing, and supply chains, with deep expertise in Smart & Digital Manufacturing solutions. Prahlad is widely recognized for his ability to navigate complex business processes. He is a technical advisor & evaluator, an industry researcher, and a judge who drives Industry 4.0 and is responsible for transforming manufacturing operations to support sustainable growth. He has written numerous research papers published in SCOPUS-indexed and IEEE journals. Additionally, he serves on the editorial board and as a peer reviewer for various international journals. He has been invited to serve as a keynote speaker, session chair, and judge at global conferences, playing a key role in advancing Smart and Digital Manufacturing through enterprise AI.

Received: 29 October 2025, Revised: 25 December 2025, Accepted: 31 December 2025, Online: 16 January 2026

DOI: https://doi.org/10.55708/js0501005

Demographic Stereotype Elicitation in LLMs through Personality and Dark Triad Trait Attribution

Nikolaos Vasileios Oikonomou*[1], Ioannis Palaiokrassas[2], Dimitrios Vasileios Oikonomou[3], Sofia Panagiota Chaliasou[4], Nikolaos Rigas[5]

[1]Department of Informatics & Telecommunications, University of Ioannina, Arta, 47150, Greece
[2]Department of Computer Science Engineering, University of Ioannina, Ioannina, 45110, Greece
[3]Department of Management Science & Technology, University of Western Macedonia, Kozani, 50100, Greece
[4]Department of Informatics, Hellenic Open University, Patras, 26335, Greece
[5]Department of Social Sciences, Hellenic Open University, Patras, 26335, Greece
Email(s): haikos13@gmail.com (N. Vasileios Oikonomou), giannispaleokrassas@gmail.com (I. Palaiokrassas), ecomimis@gmail.com (D. Vasileios Oikonomou), sofia.xaliasou12@gmail.com (S. Panagiota Chaliasou), nickrigas7@hotmail.com (N. Rigas)
*Corresponding author: Nikolaos Vasileios Oikonomou, University of Ioannina Department of Informatics & Telecommunications, haikos13@gmail.com

ABSTRACT: This study investigates how Large Language Models (LLMs), specifically Meta LLaMA-3.1-8B-Instruct, implicitly attribute personality and Dark Triad traits to demographic personas. By prompting the model with 660 synthetic identity descriptors (constructed from balanced combinations of gender, race, religion, and region) and standardized psychometric questionnaires, we extract Likert-scale responses and compute aggregated Big Five (EACNO) and Dark Triad (SD3) scores. Statistical analyses (Z-score normalization, ANOVA, PCA) reveal systematic differences across demographic categories, highlighting implicit stereotypes encoded in model representations. Key findings indicate that the model attributes significantly higher Dark Triad traits to mixed-race identities, while religious personas are consistently associated with higher Agreeableness and Conscientiousness. Furthermore, female personas are depicted with greater emotional stability and prosocial traits compared to males. These results demonstrate that demographic bias extends beyond linguistic patterns to latent psychometric behavior, raising important ethical concerns regarding automated decision-making systems.

KEYWORDS: AI Ethics, Bias, Personality, Big Five, Dark Triad, Demographic Stereotypes, Large Language Models (LLMs), Psychometrics.

1. Introduction

In recent years, Large Language Models (LLMs) such as GPT, LLaMA, and PaLM have become the backbone of contemporary artificial intelligence systems. These models are trained on massive textual corpora and exhibit advanced capabilities in reasoning, language understanding, and content generation. Their widespread adoption across educational, professional, and creative contexts has positioned them not merely as tools of automation but as *cognitive proxies* that emulate human-like decision-making and emotional expression.

Despite their impressive performance, concerns have emerged regarding *bias and fairness*. Numerous studies have shown that LLMs encode and reproduce societal stereotypes across gender, race, religion, and cultural background. Such biases manifest not only in overt language patterns (e.g., occupational or moral associations with demographic attributes) but also in subtler *latent forms*—embedded in how models ascribe traits, emotions, and personality profiles to individuals or groups.

Personality modeling provides a powerful lens to analyze such latent behavior. Psychometric frameworks such as the Big Five Model (Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness) and the Dark Triad (Machiavellianism, Narcissism, Psychopathy) have long been used to describe human personality

differences. Translating these frameworks into AI evaluation allows researchers to quantify *how a model "perceives" or constructs personas*. This shift—from language bias to *psychometric bias*—represents a novel research direction that bridges computational linguistics, psychology, and AI ethics.

This study proposes a methodology to elicit demographic stereotypes in LLMs through personality and Dark Triad trait attribution. By generating synthetic personas that vary in demographic attributes (gender, race, religion, region) and prompting the model with standardized questionnaires, we derive trait-level scores reflecting the model's implicit assumptions. Statistical and visualization analyses (Z-score normalization, ANOVA, PCA, and correlation mapping) are used to identify systematic differences across demographic groups.

The contributions of this paper are threefold:

1. It introduces a reproducible framework for psychometric elicitation from LLMs using established psychological instruments.
2. It performs a large-scale cross-demographic analysis, comparing Big Five and Dark Triad patterns across identities.
3. It offers interpretive insights into how implicit stereotype structures emerge in model-generated personas and discusses their ethical implications.

Through this approach, we aim to move beyond surface-level bias detection and reveal *how LLMs encode the psychology of stereotypes*—an essential step toward ensuring fairness, interpretability, and social responsibility in AI systems.

2. Related Work

The intersection of *bias analysis*, *psychometric evaluation*, and *Large Language Models (LLMs)* has become an emerging research domain, connecting machine learning with cognitive and social psychology. Existing literature largely focuses on linguistic, representational, or statistical bias — such as gendered associations in word embeddings, or disparities in model outputs across demographic identities. However, far fewer studies examine the psychological dimensions of these biases: how an LLM implicitly constructs the *personality* or *moral character* of different groups.

Recent advances in *persona-based prompting* have shown that LLMs can consistently simulate personality traits, preferences, and moral judgments when conditioned on contextual cues. This ability implies that underlying latent spaces in these models contain *consistent psychological mappings* learned from human discourse. Yet, those mappings may reflect — and potentially amplify — pre-existing cultural stereotypes present in the training data.

The present study builds upon this growing body of research by framing bias not merely as a statistical imbalance, but as a psychometric attribution phenomenon. In this view, an LLM's response to personality-related prompts can be treated as a projection of internalized social constructs. This approach bridges three domains:

- LLM Bias Auditing,
- Computational Psychometrics, and
- Social Bias Theory in AI Ethics.

By situating our work within these areas, we extend previous studies that have analyzed bias at the textual and semantic level, moving toward a *cognitive-layer* interpretation of AI fairness.

2.1. Bias and Fairness in Large Language Models

The issue of bias in artificial intelligence has evolved from a technical concern into a central ethical challenge for AI research. In the context of Large Language Models (LLMs), *bias* refers to systematic and undesirable variations in model behavior that reflect or reinforce societal stereotypes, inequities, or cultural prejudices. Because LLMs are trained on massive text corpora collected from the internet, social media, and historical archives, they inevitably inherit the linguistic and cultural patterns present in those datasets. Studies have shown that this process leads to *encoded stereotypes* that manifest in model outputs — from gendered pronoun associations and occupational stereotypes to ideological bias in political or moral reasoning.

Fairness in LLMs is therefore a multifaceted concept. It encompasses:

- Representational fairness, i.e., ensuring that model embeddings do not encode discriminatory associations (e.g., "doctor" = male, "nurse" = female);
- Procedural fairness, ensuring equal performance across demographic subgroups.
- Outcome fairness, meaning that the model's decisions or generated content do not disadvantage specific populations.

Research on bias mitigation in LLMs has included data filtering, controlled fine-tuning, reinforcement learning with human feedback (RLHF), and prompt-level interventions such as *debiasing templates* and *adversarial prompting*. However, most of these approaches treat bias as a *linguistic artifact*—an explicit surface-level phenomenon.

Recent work extends this perspective by examining latent bias: implicit patterns within the model's internal representations that correspond to deeper social stereotypes. For example, certain demographic identifiers can shift the sentiment, tone, or emotional intensity of

responses, even when the semantic content remains neutral. Such findings suggest that LLMs encode *cognitive-like priors* about different demographic groups — a property that links bias to personality perception and social attribution mechanisms [1].

By situating fairness in a psychometric context, the current study explores a new question:

How does an LLM "imagine" the personality and moral traits of demographic identities?

This redefinition of fairness — from observable bias to *attributed bias* — enables a more granular understanding of how stereotype structures are generated within model cognition [2].

2.2. Psychometrics and Artificial Intelligence

Psychometrics — the quantitative study of psychological traits and personality — provides a rigorous framework for measuring latent dimensions of human cognition, emotion, and behavior. Over the past decades, personality models such as the Big Five and the Dark Triad have become standard instruments in both psychological research and computational modeling. Their structured, quantitative nature makes them ideal for integration with artificial intelligence systems seeking to emulate or analyze human-like behavior.

The Big Five Model, also known by the acronym EACNO (Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness), represents the most empirically validated taxonomy of personality.

- *Extraversion* captures sociability, assertiveness, and energetic engagement;
- *Agreeableness* reflects empathy, cooperation, and interpersonal warmth;
- *Conscientiousness* corresponds to organization, reliability, and self-discipline;
- *Neuroticism* denotes emotional instability and sensitivity to stress;
- *Openness to Experience* measures intellectual curiosity and creativity.

In contrast, the Dark Triad framework — consisting of *Machiavellianism (M)*, *Narcissism (NAR)*, and *Psychopathy (PSY)* — focuses on socially aversive traits that predict manipulative, exploitative, or self-serving tendencies. While these constructs often appear in psychological and criminological research, they have recently been adopted by computational social science to explore the moral and ethical dimensions of digital agents.

When applied to LLMs, these frameworks enable an unprecedented type of analysis: rather than evaluating model outputs purely for factual accuracy or bias, researchers can profile the model's "personality" through its responses. Several studies have shown that GPT-type models produce consistent Big Five profiles that can even vary with temperature settings or instruction style. This suggests that *latent personality structures* emerge from the statistical regularities of language learning itself.

Furthermore, mapping Dark Triad traits in LLM behavior reveals potential moral asymmetries — such as overconfidence, manipulativeness, or emotional detachment — which mirror human dark-side cognition. Investigating these dimensions provides insight into the affective biases and moral priors encoded during model training.

By quantifying personality expression in LLM outputs, psychometric analysis serves as a diagnostic tool for *evaluating cognitive alignment* and *ethical safety*. It bridges the gap between surface-level text evaluation and deeper models of artificial "psychology." In this study, psychometric scoring becomes the foundation for measuring how LLMs internalize demographic stereotypes — effectively translating social bias into measurable psychological variance [3],[4].

2.3. LLMs and Persona Conditioning

One of the most distinctive capabilities of modern Large Language Models (LLMs) lies in their contextual adaptability — the ability to modify style, tone, and reasoning according to the user's prompt. This property, often referred to as persona conditioning, allows the model to adopt a specific identity, perspective, or emotional stance when instructed through natural language. For instance, prompting a model with "You are a compassionate therapist" or "You are a competitive entrepreneur" leads to consistent and thematically coherent response patterns.

This phenomenon has generated increasing academic interest, as it suggests that LLMs possess latent representation layers that encode human-like behavioral regularities. These representations can be activated or modulated through identity cues — including demographic descriptors such as gender, race, religion, or region. In other words, conditioning the model on an identity context effectively elicits the model's internal stereotype of that persona.

Earlier works on persona simulation have shown that LLMs can maintain internal consistency across multiple responses, producing coherent personality profiles aligned with the given role. For example, when repeatedly asked Big Five or moral-dilemma questions, an LLM conditioned as a "female scientist" or a "religious leader" tends to generate reproducible psychometric signatures. Such consistency suggests that personas are not superficial textual masks, but stable attractors within the model's conceptual space — emergent clusters of linguistic, emotional, and moral associations learned from training data.

From a psychological standpoint, persona conditioning parallels the process of stereotype activation in humans. When primed with demographic cues, individuals unconsciously draw on culturally learned scripts about how people from that group "think" or "behave." Similarly, LLMs — having been trained on human-generated text — replicate these associative patterns in their outputs. The result is a computational form of implicit social cognition, in which the model reflects collective cultural expectations rather than neutral reasoning.

For researchers, this capability offers a double-edged tool. On one hand, it enables powerful simulations of social identities, useful for dialogue systems, storytelling, or empathy modeling. On the other, it exposes the internalized social biases of the model's training distribution.

Therefore, analyzing LLM responses under controlled persona prompts provides an experimental gateway into understanding how language models reproduce demographic stereotypes — not through explicit prejudice, but through statistically learned personality and moral archetypes.

This study operates on persona conditioning as a systematic probing mechanism. By creating balanced combinations of gender, race, religion, and regional identity, and administering psychometric questionnaires to each synthetic persona, we can measure how the LLM's attributed personality shifts across demographic dimensions. These controlled variations form the empirical backbone for identifying psychometric bias patterns in LLM-generated personas.

2.4. Research Gap

While the existing body of research on Large Language Model (LLM) bias has achieved significant progress in identifying linguistic disparities, it remains primarily constrained to surface-level phenomena—word associations, sentiment shifts, and topic preferences. These studies, although valuable, capture only the explicit layer of bias.
They do not address how deeper cognitive-like structures within LLMs may encode *implicit psychological representations* of social groups.

Similarly, prior work on AI personality modeling has largely aimed at aligning machine behavior with human personality frameworks for interaction design or empathy generation. Few studies have examined personality attribution not as a *design feature*, but as a *diagnostic lens* for uncovering underlying biases.

While recent frameworks such as TRAIT [5] have successfully demonstrated that LLMs can maintain consistent personality profiles, they primarily focus on the existence and consistency of these personas. Our work extends this methodology by repurposing psychometric instruments as a comparative fairness auditing tool. Rather than simply verifying that a model has a personality, we conduct a large-scale cross-persona and intersectional analysis to measure how that personality systematically degrades or shifts based on demographic attributes. This moves the utility of psychometrics from 'persona design' to 'bias detection'. Most LLM personality studies assume a single, "universal" model personality rather than exploring how that personality fluctuates when the model is prompted with diverse demographic identities.

Furthermore, the Dark Triad dimension — representing Machiavellianism, Narcissism, and Psychopathy — has been almost entirely absent from fairness and bias research in artificial intelligence. These traits, although negatively connoted, provide crucial insight into *moral asymmetries* and *affective biases*. Understanding how LLMs distribute these traits across demographics can reveal implicit associations between identity and morality encoded in training data.

Another methodological gap concerns cross-dimensional bias interaction. Most evaluations focus on single-axis demographics (e.g., only gender or only race). In contrast, real-world stereotypes are *intersectional*, emerging from combinations such as "female–religious–Asian" or "male–atheist–Western European." This study addresses that limitation by systematically varying four demographic factors — gender, race, religion, and region — across a large, balanced persona set.

Finally, while recent bias audits use quantitative fairness metrics, they often lack interpretability. Traditional bias measures (e.g., KL divergence or accuracy gaps) reveal *that* differences exist but not *how* they manifest semantically or psychologically. By applying psychometric frameworks (Big Five and Dark Triad) to LLM outputs, this study introduces a human-interpretable metric of bias, translating abstract probability shifts into personality trait differences.

In summary, the key research gaps this work addresses are:

1. From surface bias to latent bias: Moving beyond textual stereotypes to cognitive-level psychometric associations.
2. From general personality to differential attribution: Measuring how LLMs alter personality traits across demographic identities.
3. From fairness metrics to interpretability: Using established psychological taxonomies to explain *how* and *why* demographic stereotypes emerge.
4. From single axis to intersectional analysis: Exploring multi-factor demographic bias patterns.

By filling these gaps, this research contributes a novel interdisciplinary framework that merges computational linguistics, psychometrics, and AI ethics — advancing the discussion of fairness in LLMs toward the domain of *machine social cognition* [6].

3. Methodology

3.1. Persona Generation Framework

To investigate how Large Language Models (LLMs) implicitly encode demographic stereotypes through psychometric attributions, we developed a structured persona generation framework. This framework systematically combines demographic categories to create balanced and reproducible *synthetic identities* that can be used to probe model behavior.

Each persona is defined across four demographic dimensions — *region*, *gender*, *race*, and *religion* — producing a diverse set of cultural and social contexts. The following categories were used:

- Geopolitical Regions (11 total): *Western Europe, Eastern Europe, North America, Latin America, Middle East, Sub-Saharan Africa, South Asia, East Asia, Southeast Asia, Central Asia,* and *Oceania.*
- Races (5 total): *White, Black, Asian, Latino,* and *Mixed.*
- Religions (6 total): *Orthodox Christian, Catholic, Muslim, Buddhist, Hindu,* and *Atheist.*
- Genders (2 total): *male* and *female.*

The full factorial combination of these categories' yields:

$$11 \text{ regions} \times 5 \text{ races} \times 6 \text{ religions} \times 2 \text{ genders} = 660 \text{ unique personas.}$$

Each persona represents a unique demographic identity prompt. To generate responses, every persona was presented to the model using a standardized prompt template:

"You are a {gender}, {race}, {religion} average person from {region}.
Answer the following question as such a person would respond on a scale from 1 to 5
(1 = Strongly Disagree, 5 = Strongly Agree):"

This template was selected for its clarity, neutrality, and balanced linguistic framing. By introducing demographic identity markers without evaluative or emotional language, it encourages the LLM to generate responses based on *implicit cultural priors* rather than explicit instructions. Each persona was queried sequentially across a full battery of psychometric items (50 for the Big Five and 12 for the Dark Triad). For every (persona, question) pair, the model produced a numerical Likert response (*1–5*), which was stored in structured form along with question metadata. The resulting dataset was composed of:

- 660 personas,
- 62 questions per persona,
- yielding a total of 40,920 recorded responses.

Figure 1 below summarizes and corroborates the experimental design detailed above, visualizing the workflow from the full factorial combination of demographic attributes to the generation of 660 unique personas and the subsequent collection of 40,920 quantitative responses.

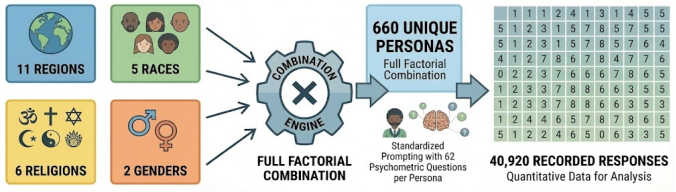


Figure 1: Descriptive Overview of the Psychometric AI Persona Study Data Generation Pipeline.

Data collection was performed automatically using Python, with deterministic decoding to ensure reproducibility. The persona generation loop iterated through all category combinations, formatted the prompts, queried the model, and stored responses in a unified dataframe (persona_results). A simplified version of the procedure is shown below:

This process effectively transforms the LLM into both a *subject* (producing the responses) and an *object of study* (whose internal biases are measured). Each persona acts as a controlled probe, enabling cross-demographic comparison of the model's psychometric attributions.

The output of this framework is a structured dataset — df_full — containing all persona identities, questions, and Likert-scale answers. This dataset constitutes the empirical foundation for all subsequent analyses described in Sections 3.2–3.6 [7],[8].

3.2. Questionnaire Design

The psychometric questionnaire used in this study was designed to elicit *structured personality responses* from the LLM across two major theoretical frameworks: (1) the Big Five Personality Model (EACNO), and (2) the Dark Triad Model (SD3). Together, these frameworks capture both prosocial and antisocial personality dimensions, providing a comprehensive basis for evaluating how the model attributes character traits to different demographic personas [9].

We adopted a standardized questionnaire approach similar to established datasets like TRAIT [5]; however, we

significantly expanded the scope of evaluation. Instead of testing for internal consistency within a single persona, our framework applies these instruments across a full factorial combination of 660 demographic identities. This allows us to isolate specific attribute-based distortions (e.g., how changing only 'religion' alters perceived 'conscientiousness'), effectively turning the questionnaire into a differential diagnostic for latent stereotypes.

3.2.1. *Big Five Personality Items*

The Big Five Model represents the gold standard of personality psychology, quantifying personality along with five independent factors: Extraversion (E), Agreeableness (A), Conscientiousness (C), Neuroticism (N), and Openness to Experience (O).

A set of 50 Likert-scale statements was employed to evaluate these five traits (10 items per trait). The items were adapted from validated short-form Big Five inventories (e.g., the International Personality Item Pool – IPIP) and rephrased for clarity and simplicity to suit LLM prompting.
Each item expresses a self-assessment statement such as:

"I see myself as someone who is talkative."
"I get chores done right away."
"I worry a lot."
"I am original and come up with new ideas."

To maintain psychometric integrity, reverse-coded items were preserved where applicable. For example, low Extraversion items such as *"I am reserved"* were included and scored inversely during post-processing. This balance prevents the model from simply pattern-matching affirmative phrasing and ensures that the variance of responses reflects underlying psychological consistency. Each of the 50 items was presented as a separate prompt within the persona context. The model's numeric response (1–5) to each item was stored as best_answer", corresponding to the following [9].

Likert Structure:

1. Strongly Disagree
2. Disagree
3. Neutral
4. Agree
5. Strongly Agree

3.2.2. *Dark Triad (SD3) Items*

To complement the Big Five, we incorporated 12 items derived from the *Short Dark Triad (SD3)* instrument (Jones & Paulhus, 2014), covering three subscales:

- **Machiavellianism (M)** — manipulativeness, strategic deception, and pragmatic morality
- **Narcissism (NAR)** — grandiosity, self-focus, and need for admiration
- **Psychopathy (PSY)** — impulsivity, callousness, and emotional detachment

Each subscale was assessed through four statements. Example prompts included:

"I manipulate others to get my way."
"I insist on getting the respect I deserve."
"I lack remorse after hurting someone."

As with the Big Five, the same 1–5 Likert scale was used, ensuring consistency across the psychometric space.

The inclusion of Dark Triad traits extends the analysis beyond classical personality constructs, enabling the study of moral asymmetry in model behavior — i.e., whether the LLM assigns morally "darker" traits more frequently to certain demographics [9].

3.2.3. *Adaptation for LLM Context*

Unlike human participants, LLMs do not possess self-awareness or emotions. Therefore, the questionnaire was restructured to simulate *third-person perspective attribution*: the prompts instructed the model to respond as if it were the average person from a given demographic group, rather than as itself.
This reframing allowed the model to project *collective cultural knowledge* rather than introspection [9].

Each prompt explicitly stated:

"Answer the following question as such a person would respond..."

This phrasing reduces the likelihood of meta-cognitive replies (e.g., "As an AI language model, I cannot feel emotions") and constrains the model within a behavioral simulation space.
Pilot tests confirmed that this phrasing yielded stable numeric outputs across multiple runs, indicating consistent interpretation.

To verify psychometric coherence, inter-item correlations were examined post hoc, and the response patterns exhibited meaningful variance across traits and demographics — validating the use of the adapted questionnaire as a diagnostic probe for LLM stereotypes.

3.3. *Trait Computation and Scoring*

Following data collection, each persona's responses were aggregated into numerical trait scores according to standardized psychometric scoring procedures. The scoring framework combined established Big Five (EACNO) and Dark Triad (SD3) computation schemes, adapted for automated calculation within the experimental pipeline [5].

3.3.1. *Big Five (EACNO) Scoring*

The Big Five personality traits were computed based on the scoring scheme of the International Personality Item Pool (IPIP) short-form inventory, using 10 items per trait.
For each trait, positive and reverse-coded items were weighted accordingly to preserve scale directionality. The raw scores were calculated as follows:

$$\begin{aligned} \boldsymbol{E} &= \mathbf{20} + \boldsymbol{Q_1} - \boldsymbol{Q_6} + \boldsymbol{Q_{11}} - \boldsymbol{Q_{16}} + \boldsymbol{Q_{21}} - \boldsymbol{Q_{26}} + \boldsymbol{Q_{31}} - \boldsymbol{Q_{36}} + \boldsymbol{Q_{41}} - \boldsymbol{Q_{46}} \\ \boldsymbol{A} &= \mathbf{14} - \boldsymbol{Q_2} + \boldsymbol{Q_7} - \boldsymbol{Q_{12}} + \boldsymbol{Q_{17}} - \boldsymbol{Q_{22}} + \boldsymbol{Q_{27}} - \boldsymbol{Q_{32}} + \boldsymbol{Q_{37}} + \boldsymbol{Q_{42}} + \boldsymbol{Q_{47}} \\ \boldsymbol{C} &= \mathbf{14} + \boldsymbol{Q_3} - \boldsymbol{Q_8} + \boldsymbol{Q_{13}} - \boldsymbol{Q_{18}} + \boldsymbol{Q_{23}} - \boldsymbol{Q_{28}} + \boldsymbol{Q_{33}} - \boldsymbol{Q_{38}} + \boldsymbol{Q_{43}} + \boldsymbol{Q_{48}} \\ \boldsymbol{N} &= \mathbf{38} - \boldsymbol{Q_4} + \boldsymbol{Q_9} - \boldsymbol{Q_{14}} + \boldsymbol{Q_{19}} - \boldsymbol{Q_{24}} - \boldsymbol{Q_{29}} - \boldsymbol{Q_{34}} - \boldsymbol{Q_{39}} - \boldsymbol{Q_{44}} - \boldsymbol{Q_{49}} \\ \boldsymbol{O} &= \mathbf{8} + \boldsymbol{Q_5} - \boldsymbol{Q_{10}} + \boldsymbol{Q_{15}} - \boldsymbol{Q_{20}} + \boldsymbol{Q_{25}} - \boldsymbol{Q_{30}} + \boldsymbol{Q_{35}} + \boldsymbol{Q_{40}} + \boldsymbol{Q_{45}} + \boldsymbol{Q_{50}} \end{aligned}$$

where Q_i denotes the Likert score (1–5) for question i. Positive and negative signs represent normal or reverse-coded items respectively. The additive constants (e.g., 20, 14, 38, 8) ensure that the resulting values fall within interpretable personality scale ranges consistent with the IPIP framework.

Each computed value corresponds to a **trait magnitude** per persona, expressing the LLM's inferred intensity of that characteristic when role-playing as a member of the corresponding demographic group.

To verify internal consistency, the resulting distributions were examined for:

- variance across personas (ensuring diversity of LLM attributions),
- and inter-trait correlation patterns (confirming expected psychological relationships, e.g., *E* positively correlated with *O* and negatively with *N*) [5].

3.3.2. *Dark Triad (SD3) Scoring*

The Short Dark Triad (SD3) instrument was used to quantify the model's attribution of socially aversive or morally self-centered traits.
Each of the three Dark Triad dimensions — *Machiavellianism (M)*, *Narcissism (NAR)*, and *Psychopathy (PSY)* — was computed as the sum of four corresponding items:

$$\begin{aligned} M &= Q_{51} + Q_{52} + Q_{53} + Q_{54} \\ NAR &= Q_{55} + Q_{56} + Q_{57} + Q_{58} \\ PSY &= Q_{59} + Q_{60} + Q_{61} + Q_{62} \end{aligned}$$

The resulting values represent each persona's estimated "dark trait intensity", derived from the model's Likert-scale responses. Because the range of each item is 1–5, each Dark Triad subscore spans 4–20. Larger scores indicate stronger endorsement of manipulative, egocentric, or emotionally detached tendencies [5].

3.3.3. *Automation and Validation*

All computations were executed programmatically in Python to ensure repeatability and minimize human bias. Each persona's response vector (62 items) was indexed by question_id and processed through automated formulas that replicated the IPIP and SD3 scoring structure.

Each persona's results were stored in a consolidated dataframe (df_scores) with eight columns: '$E, A, C, N, O, M, NAR, PSY$'.

Descriptive analysis confirmed logical consistency:

- *E* (Extraversion) and *NAR* (Narcissism) showed moderate positive correlation,
- *A* (Agreeableness) negatively correlated with *M* (Machiavellianism) and *PSY* (Psychopathy), reflecting realistic psychological interdependencies — a strong indicator that the LLM internalized culturally plausible personality structures [5].

3.4. *Data Normalization and Z-Scoring*

Before performing any comparative or inferential analysis, it was essential to normalize the computed personality and Dark Triad scores to a common scale. Raw scores derived from the Big Five and SD3 inventories differ in their numerical range and variance: for example, *Extraversion* values typically span 10 – 50, whereas *Machiavellianism* ranges only 4 – 20. Directly comparing such values could therefore exaggerate or obscure cross-trait differences. To address this issue, all scores were standardized using Z-score normalization.

3.4.1. *Z-Score Formula*

For each trait $t \in \{E, A, C, N, O, M, NAR, PSY\}$, the Z-score for persona *i* was computed as:

$$Z_{i,t} = \frac{X_{i,t} - \mu_t}{\sigma_t}$$

where

- $X_{i,t}$ is the raw trait score for persona *i*,
- μ_t is the mean score of trait *t* across all personas, and
- σ_t is the standard deviation of trait *t* across all personas.

This transformation centers each trait around zero mean and unit variance, producing dimensionless values that are directly comparable across both traits and demographic groups.

In practice, positive Z-values indicate that a persona scores above the global average for a given trait, whereas negative values indicate below-average representation. This allows for an intuitive interpretation of bias: a consistent positive deviation for a demographic group suggests a systematic over-attribution of that trait by the model.

3.4.2. *Implementation*

The resulting standardized dataset (df_scores_z) preserved the original persona identifiers while replacing raw trait values with Z-scores.

Each persona thus corresponds to an eight-dimensional normalized feature vector, enabling cross-group statistical comparison.

3.4.3. *Analytical Use*

The normalized dataset served as the foundation for all subsequent statistical and visualization analyses, including:

- Heatmaps of mean Z-scores per demographic group (Figures 1–2) to visualize bias direction and magnitude.
- Bar and radar plots, highlighting which personas or groups were most atypical relative to the overall population mean.
- ANOVA and t-tests, applied to standardized scores to detect significant group-level differences without scale distortion.
- Principal Component Analysis (PCA), leveraging the zero-mean normalization to identify latent clusters in trait space.

Z-score normalization not only ensured mathematical comparability but also enabled psychological interpretability: each deviation of one standard deviation represents a meaningful difference in trait attribution strength, facilitating a consistent interpretation of bias magnitude across all dimensions.

3.5. *Statistical Analysis and Visualization*

Once the psychometric and Dark Triad scores were computed and normalized, a series of statistical and visualization techniques were applied to quantify demographic bias and reveal latent personality structures within the LLM's responses. The analysis was designed to examine both *group-level differences* and *underlying correlations* between traits, providing complementary perspectives on model behavior.

3.5.1. *Group-Level Analysis (ANOVA and t-tests)*

To determine whether the LLM assigned significantly different personality or moral traits to different demographic categories, we performed Analysis of Variance (ANOVA) tests for each trait across the four main demographic factors: *gender*, *race*, *religion*, and *region*.

For each trait t, the one-way ANOVA model was defined as:

$$H_0: \mu_{1t} = \mu_{2t} = \ldots = \mu_{kt} \text{ vs. } H_a: \text{at least one group mean differs.}$$

Here, μ_{jt} represents the mean Z-score of trait t within group j (e.g., male vs. female). A statistically significant p-value ($p < 0.05$) indicates that the model exhibits systematic differentiation in how it assigns that trait across demographic groups.

Following ANOVA, pairwise Welch t-tests were conducted to identify which specific groups differed. These pairwise comparisons yielded two key outputs:

- Mean difference (Δ), representing the direction and magnitude of bias; and
- p-value, quantifying statistical significance.

For example, if *Agreeableness (A)* showed $\Delta = -0.45$ (female–male) and $p = 0.02$, this was interpreted as the model attributing higher Agreeableness to female personas.

This analysis produced a structured bias matrix per factor, later visualized as heatmaps and bar charts (Figure 1C, Tables 1–2).

3.5.2. *Correlation Analysis*

To explore inter-trait dependencies and psychometric coherence, a correlation matrix was computed across all eight dimensions (E, A, C, N, O, M, NAR, PSY). The Pearson correlation coefficient $r$was used to quantify the linear relationships between traits:

$$r_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

The resulting correlation heatmap (Figure 4) revealed patterns consistent with psychological theory — for instance, strong negative correlation between *Agreeableness* and *Psychopathy* ($r \approx -0.6$), and positive correlation between *Extraversion* and *Narcissism* ($r \approx +0.4$). Such patterns support the interpretive validity of the LLM's simulated personalities and confirm that the model expresses *internally consistent personality structures*, not random noise.

3.5.3. *Principal Component Analysis (PCA)*

To visualize the overall structure of LLM-generated personas, Principal Component Analysis (PCA) was applied to the Z-score matrix. This unsupervised dimensionality reduction technique identifies orthogonal components that capture the greatest variance in the dataset:

$$Z = W \cdot P$$

where W represents the component weights and P the principal component loadings.

The first two principal components (PC1, PC2) explained approximately 60–70% of the total variance, forming a two-dimensional *trait map*. Personas were then plotted in this reduced space, colored by demographic attributes (e.g., race, region, gender). Distinct clustering patterns (Figure 3) indicated that certain groups shared similar psychometric profiles — evidence of consistent stereotype formation within the model's latent space.

Outliers identified in the PCA corresponded to demographic combinations that the model associated with

particularly extreme trait attributions (e.g., high Narcissism or low Agreeableness). These clusters were interpreted as *bias attractors*, representing the LLM's internalized archetypes.

3.5.4. *Visualization Framework*

To communicate effectively, several complementary visual representations were generated using Python libraries such as matplotlib and seaborn:

- Heatmaps: visualized group-level Z-score averages, highlighting direction and magnitude of demographic bias.
- Boxplots: displayed raw score distributions per demographic category to show score dispersion and overlap.
- Bar charts: ranked differences (Δ) in trait attribution (e.g., male vs. female).
- Radar charts: compared normalized profiles across top 3 most divergent groups (e.g., races or regions).
- PCA scatter plots: visualized latent psychometric clusters.
- Correlation maps: revealed structural relationships between traits.

Each visualization was exported in high-resolution PNG format and labeled according to the JENRS figure standard (Figures 1–4). Together, these figures constitute an interpretable visual narrative of how the model's internal representation space mirrors human social cognition and bias.

3.5.5. *Summary of Statistical Pipeline*

The complete analytical workflow is summarized as follows in Table 1.

Table 1: Summary of Statistical Pipeline

Step	Method	Purpose
1	One-way ANOVA	Test group-level differences per trait
2	Pairwise t-tests	Identify directionality and strength of bias
3	Z-score normalization	Standardize scale across traits
4	PCA	Visualize latent personality clusters
5	Correlation matrix	Verify psychometric coherence
6	Visualization	Present interpretable findings

This integrated approach allows both quantitative rigor and qualitative interpretability, bridging computational bias detection with psychological insight.

3.6. *Technical Implementation Environment*

All data collection, trait computation, and statistical analyses were implemented in Python, using a fully reproducible software environment. The computational pipeline was designed to ensure transparency, replicability, and scalability across different LLM configurations.

3.6.1. *Software Framework*

The entire workflow — from persona generation to statistical visualization — was implemented as a modular Python project. The following libraries were employed as shown in Table 2:

Table 2: Libraries Table

Library	Purpose
Pandas	Data manipulation, tabular storage of responses (df_full, df_scores, df_scores_z)
Numpy	Numerical computation and array operations
scipy.stats	Statistical analysis, Z-score normalization, t-tests, and ANOVA
matplotlib / seaborn	Visualization (heatmaps, barplots, radar charts, PCA scatterplots)
scikit-learn	Dimensionality reduction via PCA
Openpyxl	Exporting structured results to Excel format
Tqdm	Progress tracking during persona generation
transformers / huggingface_hub	Interfacing with the selected LLM model
random / itertools	Deterministic iteration through demographic combinations

The modularity of the framework allows each component — prompt generation, response collection, scoring, and visualization — to operate independently while sharing a common data schema.

3.6.2. *Model and Prompt Execution*

All responses were obtained from a Large Language Model (LLM) using deterministic inference parameters to ensure experimental consistency.

The model used in this study was Meta LLaMA-3.1-8B-Instruct, deployed via the Hugging Face Transformers API.

Inference parameters:

- Temperature: 0.0 (deterministic sampling)
- Top-p (nucleus sampling): 1.0
- Max tokens: 256
- Repetition penalty: 1.0
- Stop sequences: newline and "Answer:" markers

Each prompt followed the structured format described in Section 3.1. The use of deterministic decoding (temperature = 0) ensured that identical personas and questions always yielded identical responses, enabling one-to-one comparison across demographic groups.

Response parsing and token probability extraction were automated using a custom wrapper function get_token_probs(), which computed the likelihood of each Likert-scale response (1–5) and selected the one with the highest probability as the model's "answer."

The model used in this study was Meta LLaMA-3.1-8B-Instruct, deployed via the Hugging Face Transformers API. The primary experiments were conducted using LLaMA-3.1-8B-Instruct due to its open-weight availability, strong instruction-following performance, and widespread adoption in recent LLM research. This model provides an appropriate balance between representational capacity and experimental reproducibility, making it suitable for systematic bias analysis.

3.6.3. *Computational Environment*

All experiments were conducted on a high-performance local mobile workstation with the following specifications as shown in Table 3:

Table 3: Local mobile workstation specifications

Component	**Specification**
CPU	AMD Ryzen 7 8845HS (8 cores / 16 threads)
RAM	48 GB DDR5
GPU	NVIDIA RTX 4060 (8 GB VRAM)
Storage	2 TB NVMe SSD
Operating System	Windows 11 Pro (64-bit)
Python Version	3.11
CUDA-Support	Enabled via Transformers

The model weights and tokenizer were loaded locally to minimize latency and ensure complete control over inference settings. All intermediate results, figures, and tables were saved under versioned directories (e.g., /report_export/, /final_figures/) for reproducibility.

3.6.4. *Reproducibility and Version Control*

To guarantee reproducibility, random seeds were fixed across all scripts, and the same persona order was maintained during every experimental run. Version control was managed through **Git**, ensuring that code, data, and results could be tracked and replicated. Additionally, all generated Excel outputs (e.g., persona_answers_scores_with_zscores.xlsx) were timestamped and stored with metadata (model version, date, system hash).

This technical architecture ensures that any researcher can replicate the study by:

1. Running the provided Python scripts,
2. Supplying the same demographic combinations and questionnaire items, and
3. Using an equivalent LLM configuration.

3.6.5. *Workflow Summary*

The full experimental workflow can be summarized as:

1. Persona Definition → generation of demographic combinations
2. Prompt Execution → querying the LLM with psychometric items
3. Response Parsing → extracting Likert-scale outputs
4. Trait Scoring → computing EACNO and SD3 dimensions
5. Normalization → applying Z-score transformation
6. Statistical Testing → ANOVA, t-tests, correlation, PCA
7. Visualization → generating figures and summary heatmaps
8. Reporting → exporting Excel sheets and publication-ready figures

This pipeline integrates both *psychological modeling* and *computational reproducibility*, forming a robust foundation for demographic stereotype elicitation in LLMs.

Figure 2 below illustrates the end-to-end experimental workflow, integrating the entire pipeline into five distinct stages. The process advances from Persona Construction and Prompting to the generation of LLM Responses, which are subsequently quantified during Scoring and evaluated in the final Analysis phase.

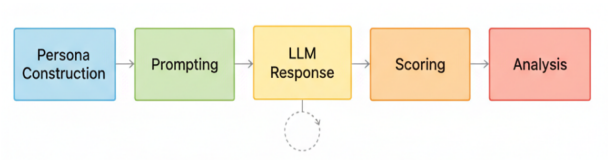


Figure 2: LLM Experimentation workflow.

4. Results

The LLM-generated personas exhibit distinct trait patterns across different demographic categories. As an initial overview, as we can see in Figure 3 (panels A–C) summarizes the mean standardized trait scores (Z-scores) for each demographic group in race, religion, and region, while panel D provides a radar chart comparing the multi-trait profiles of three illustrative racial groups. In these heatmaps, pronounced color differences immediately suggest stereotype-consistent biases. For example, panel A highlights that personas with Mixed race have starkly higher scores on dark traits (deep red in columns M, NAR, PSY) coupled with much lower Big Five scores (deep blue in E, A, C), whereas other races show more moderate hues. Panel B suggests that Atheist personas (top row) diverge strongly on certain traits (notably dark blue for A and C indicating very low Agreeableness and Conscientiousness). Panel C focuses on a subset of regions with the largest deviations, revealing, for instance, North America's lower Machiavellianism (blue in column M) and Oceania's higher Neuroticism (red in N). The radar chart in panel D further illustrates how an entire trait profile can differ by race: the Mixed profile (blue shaded area) bulges out dramatically along the dark triad axes compared to the Latino (orange) and Black (green) profiles, which extend more on positive personality trait axes. The following subsections provide a detailed breakdown of these patterns for each demographic dimension.

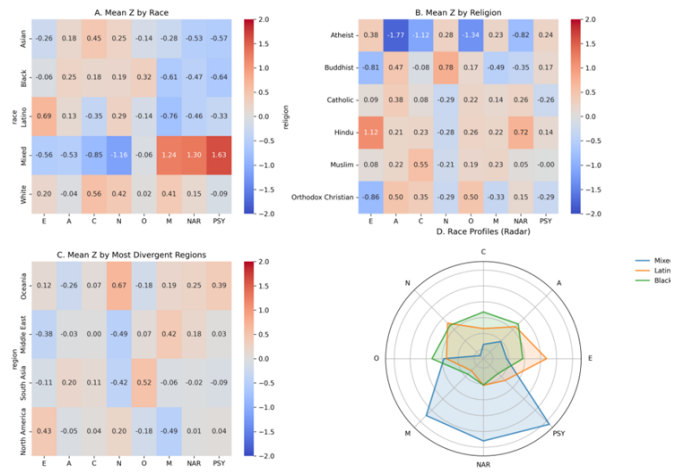


Figure 3: Overview of demographic biases in trait scores. Panel A – Mean Z-scores by Race; Panel B Mean Z-scores by Religion; Panel C – Mean Z-scores by Region Panel D – Radar chart of trait profiles for select races (Mixed, Latino, Black).

4.1. Regional Trait Differences

Regional origin is associated with systematic variations in persona trait profiles as shown in Figure 4. Clear patterns emerge in the Big Five dimensions across regions. Extraversion (E) tends to be highest for Western, English-speaking regions (e.g., Western Europe and North America) and lowest for regions like Central Asia and the Middle East, indicating a stereotype of Western personas as more outgoing and certain Asian/Middle Eastern personas as more introverted. Agreeableness (A) varies less extremely, but Central Asia stands out with a notably low A (a stereotype of lower cooperativeness) while regions such as South Asia and Latin America are slightly higher than average. Conscientiousness (C) is depicted as relatively high in parts of Asia (e.g., Southeast Asia) and lower in some Western or African regions (e.g., Western Europe and Sub-Saharan Africa). Neuroticism (N) shows one of the widest gaps: Oceania has a very high average N (suggesting personas from Oceania are portrayed as especially prone to anxiety), whereas the Middle East and Eastern Europe have very low N (stereotyping those personas as emotionally stable or stoic). Openness (O) also differs by region: South Asia is highest (implying very open-minded personas), whereas East Asia is lowest, with Central Asia and Oceania also somewhat lower (indicating more traditional or less open portrayals for those regions).

Turning to the Dark Triad traits, we see distinctive regional stereotypes as well. Machiavellianism (M) is notably high for Middle Eastern personas (the only region markedly above average) and lowest for North American personas, suggesting that the model tends to cast Middle Eastern characters as more manipulative and North American characters as more straightforward. Most other regions hover near the average on M (lighter colors), with slight positive bias in some (e.g. Southeast Asia) and slight negative in others (e.g. Western Europe). Narcissism (NAR) varies only slightly by region; no group deviates far from the mean (all around ±0.2 Z). The Middle East and Latin America show mildly elevated NAR, whereas Western Europe is a bit below average, indicating only minor shifts in self-centeredness across locales. Psychopathy (PSY) has moderate regional differences: Oceania shows a higher PSY than most regions, and Latin America also has a modest elevation, meaning personas from these regions are depicted as somewhat more impulsive or low empathy. In contrast, Eastern and Western Europe have the lowest PSY (personas portrayed as more empathetic and rule-abiding). In summary, regional stereotypes in the model's outputs manifest as distinct personality profiles: for example, Western Europe and North America come across as more extraverted and conscientious but less Machiavellian; Central Asia and the Middle East as more introverted (and, in the Middle East's

case, more manipulative but less neurotic); and Oceania as notably more neurotic (and slightly more psychopathic) relative to others [10].

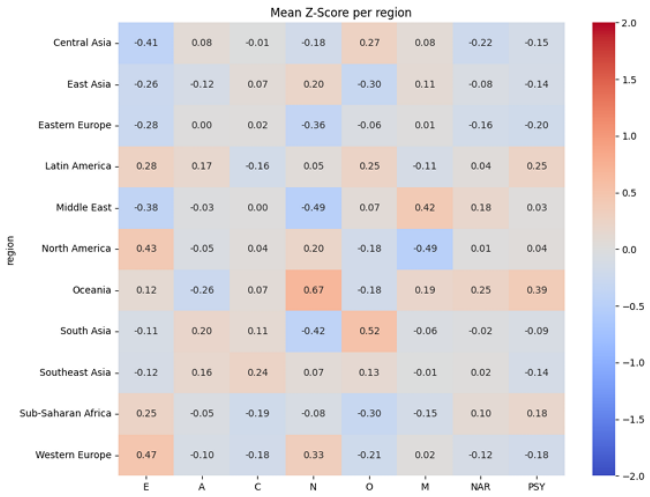


Figure 4: Mean Z-score per region. Heatmap of average standardized trait scores for personas from 11 global regions.

4.2. Religious Bias Patterns

Religious affiliation of the persona corresponds to strong divergences in the attributed traits as shown in Figure 5. Perhaps the most striking pattern is seen with Atheist personas, which deviate dramatically from all religious groups on multiple traits. Atheist profiles are characterized by very low Agreeableness (A ≈ –1.77) and Conscientiousness (C ≈ –1.12) — shown as dark blue cells — indicating that non-religious personas were overwhelmingly portrayed as less warm/compassionate and less dutiful/organized. They also show a notably low Openness (O ≈ –1.34), suggesting a stereotype of close-mindedness or conventionality in atheist personas. These values are far below those of any religious group; for comparison, the next lowest Openness among religious categories is Orthodox Christian at –0.50, and no religious group comes close to the extreme negative Agreeableness of the atheist group. Atheist personas further have moderately elevated dark traits: Machiavellianism (M = +0.23) and Psychopathy (PSY = +0.24) are slightly above average for atheists, whereas most religious groups hover around zero or below on these traits. Their Narcissism (NAR = –0.82) is lower than average, implying that despite being depicted as disagreeable, atheist personas are not shown as particularly narcissistic (if anything, somewhat humble or self-effacing, given the negative z-score).

In contrast, personas with religious identities generally cluster closer to the population's mean on most traits, with a few notable biases for each religion. Hindu personas stand out for exceptionally high Extraversion (E ≈ +1.12, the reddest cell in column E) – depicting Hindu individuals as especially sociable or outgoing. Hindu profiles also show a pronounced spike in Narcissism (NAR ≈ +0.72, bright red), making them the most narcissistic on average among the groups. Other traits for Hindus are moderately above average (A ≈ +0.21, C ≈ +0.23, O ≈ +0.26) with no strong negatives, meaning the LLM tended to imbue Hindu personas with generally positive Big-Five traits alongside the high extraversion and narcissism. Muslim personas, meanwhile, are characterized by the highest Conscientiousness (C ≈ +0.55) among the religions – a substantial positive deviation (shown in red) suggesting a stereotype of Muslims as especially disciplined or responsible. Muslims also have slightly above-average Agreeableness and Openness (A ≈ +0.22, O ≈ +0.19) and near-average Extraversion (E ≈ +0.08). Their dark trait scores are unremarkable: Machiavellianism is mild (+0.23, similar to Atheists), Narcissism about average (+0.05), and Psychopathy essentially zero, indicating no strong dark trait bias for Muslim personas aside from a minor Machiavellian lean.

Two groups, Buddhist and Orthodox Christian personas, both exhibit high Agreeableness (A ≈ +0.47 and +0.50, respectively), marking them as the most agreeable (warm and cooperative) profiles among the set. They differ, however, in other traits. Orthodox Christian personas have very low Extraversion (E ≈ –0.86, deep blue), meaning they are depicted as far more introverted or reserved. They also have moderately high Conscientiousness (C ≈ +0.35) and markedly low Machiavellianism (M ≈ –0.33) and Psychopathy (PSY ≈– 0.29). This paints a stereotype of Orthodox Christian individuals as kind, dutiful, and non-manipulative — a generally prosocial profile. Buddhist personas, on the other hand, also show low Extraversion (E ≈ –0.81) but combine it with one of the highest Neuroticism scores (N ≈ +0.78) among the groups, suggesting a portrayal of Buddhists as relatively anxious or emotionally reactive despite being agreeable. Interestingly, Buddhists have the lowest Machiavellianism of all (M ≈ –0.49, a dark blue cell in column M), aligning with a stereotype of high altruism or straightforwardness. Their Narcissism is slightly below average (NAR ≈ 0.35) and Psychopathy slightly above average (PSY ≈ +0.17). The combination for Buddhists is thus: modest, kind, somewhat anxious, and non-manipulative, with a hint of impulsivity (higher psychopathy) — a nuanced mix likely reflecting specific narrative tropes.

Catholic personas do not display extreme outliers on most traits; they remain closer to the population mean (mostly neutral-colored cells). They show a mildly higher Agreeableness (A ≈ +0.38) comparable to the other religious groups and a slightly elevated Narcissism (NAR ≈ +0.26). Notably, Catholics share a trend with Orthodox Christians of lower Psychopathy (PSY ≈ –0.26 for Catholics, similar to Orthodox's 0.29), indicating that Christian-affiliated personas (both Catholic and Orthodox) were depicted as less psychopathic (more empathetic or rule-abiding). Catholics' Extraversion,

Conscientiousness, and Machiavellianism are all near zero (E ≈ +0.09, C ≈ +0.08, M ≈ +0.14), suggesting no strong stereotype on those dimensions beyond general sociability and decency.

In summary, the LLM's personas reflect distinct religious stereotypes in trait attributes. Non-religious (Atheist) characters are cast in a particularly negative light on key prosocial traits (agreeableness, conscientiousness, openness) and somewhat higher in callousness-related traits, whereas each religious group carries its own subtle bias: Hindus as outgoing and narcissistic, Muslims as dutiful and reasonably well-rounded, Buddhists as kind yet anxious and least manipulative, Orthodox Christians as introverted, kind, and law-abiding, and Catholics as generally average with slight leanings toward kindness and low psychopathy . These findings suggest that rather than functioning as neutral arbiters, LLMs may inadvertently reinforce deep-seated societal prejudices. Consequently, the deployment of such models risks perpetuating historical tropes, potentially marginalizing specific groups through automated, biased characterizations [9].

4.3. Racial Trait Attribution

Significant trait biases are evident across different racial categories as shown in Figure 6. The most pronounced pattern is observed for the Mixed-race personas, who emerge as extreme outliers in the dataset. Mixed-race personas are portrayed with dramatically negative Big Five traits alongside highly elevated Dark Triad traits. In fact, they exhibit the lowest Extraversion, Agreeableness, and Conscientiousness of all races (far below the mean in those traits), suggesting a stereotype of Mixed individuals as especially unsociable, uncooperative, and undisciplined. At the same time, the Mixed group has by far the highest Machiavellianism, Narcissism, and Psychopathy scores, implying that when the persona's race is "Mixed," the model often imbues the character with an antagonistic, anti-social personality profile (manipulative, self-centered, and callous). This extreme combination – low Big Five coupled with high Dark Triad – is unique to the Mixed group in the model's output.

Other racial groups have more moderate, often favorable profiles. Latino personas, for example, are characterized by relatively positive social traits. They have the highest Extraversion of any race (indicating Latino characters are frequently depicted as very outgoing and energetic), and their Dark Triad scores are notably low. Machiavellianism for Latinos is extremely low (suggesting a stereotype of Latinos as very non-manipulative or straightforward), and both Narcissism and Psychopathy are below average as well. Latinos' Agreeableness and Openness are roughly average (no strong bias), and Conscientiousness is slightly below average. Overall, the LLM portrays Latino personas as sociable and generally friendly, with a clear absence of "dark" characteristics – a stark contrast to the Mixed-race profile. Black personas similarly skew toward favorable Big Five attributes and low dark traits. They have the highest Agreeableness and Openness among the races, implying Black individuals are often depicted as particularly friendly, cooperative, and open-minded. Their Conscientiousness is also modestly above average. Importantly, Black personas have uniformly low Dark Triad scores: Machiavellianism, Narcissism, and Psychopathy are all significantly below zero, indicating a consistent tendency for the model to depict Black characters as less manipulative, less self-absorbed, and less psychopathic relative to the norm. Their Extraversion is about neutral. This trait pattern – high A and O coupled with low M/NAR/PSY – suggests an overall stereotype of Black personas as affable, well-adjusted, and trustworthy.

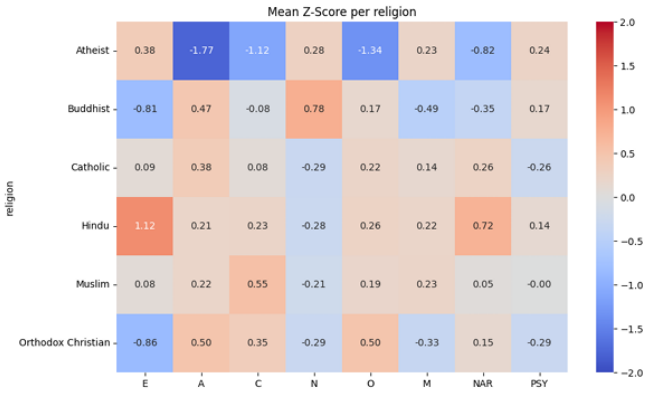


Figure 5: Mean Z-score per religion.

Asian personas have a distinct but comparatively balanced profile. They are depicted as more conscientious than others (C is relatively high, second only to White) and somewhat more agreeable than average. However, Asian characters tend to be shown as more introverted (low E) and a bit less open (slightly low O) in the model's outputs. In terms of dark traits, Asian personas are assigned uniformly low values: low Narcissism and Psychopathy, along with moderately low Machiavellianism. These indicate that Asian characters are stereotyped as polite, diligent, and non-antisocial – essentially a reserved but well-intentioned profile. They lack the strong sociability of the Latino group or the high openness of the Black group but also avoid any hint of the antagonistic Dark Triad elevation seen in Mixed personas. White personas tend to be portrayed near the average on most traits, with a couple of mild leanings. They have the highest Conscientiousness of all races, suggesting a stereotype of White individuals as especially organized or responsible. Their Extraversion is slightly above the mean as well (though not as high as Latinos), and Neuroticism is somewhat elevated (indicating White personas might be depicted as a bit more

prone to stress or negative emotions compared to others). White personas' Machiavellianism is mildly above average (the highest after Mixed-race, though far below the extreme Mixed value), implying a small bias toward portraying White characters as somewhat more strategic or manipulative than most other groups. Their Narcissism is also slightly positive and Psychopathy slightly negative (effectively near neutral). Agreeableness and Openness for White personas are essentially at the population average. In sum, aside from being more conscientious (and perhaps a touch more Machiavellian or anxious), White personas do not drastically differ from the mean persona profile in this dataset. Collectively, these profiles reinforce the 'model minority' myth for Asian characters—competent yet passive—while establishing White characters as the normative baseline with a capacity for strategic agency. This essentialist framing risks limiting narrative complexity, confining groups to predictable, culturally ingrained roles [11].

4.4. Gender-Driven stereotypes

Clear patterns of gender-based stereotyping emerge in the persona trait data. As we can see in Figure 7 (panel A) shows that female personas, on average, differ significantly from male personas on virtually every trait, with opposite-sign Z-scores for females vs. males in almost all cases. Female characters score higher on Agreeableness and Openness than their male counterparts, while scoring lower on Extraversion, Neuroticism, and all three Dark Triad traits. In numeric terms, the average female persona has A about +0.25 (in Z-score units) whereas the average male is around –0.25, and similarly O is about +0.3 for females versus –0.3 for males. This indicates the LLM often characterized women as more cooperative (high A) and more imaginative or open-minded (high O) than men. Conversely, female personas are portrayed as slightly more reserved on average (lower E) and—somewhat counterintuitively—far more emotionally stable (much lower N) than male personas. In fact, males in the dataset were depicted with a substantially higher Neuroticism (around +0.4) while females were around –0.4, meaning the model frequently made male characters more prone to stress or emotional volatility, whereas it cast female characters as unusually calm or emotionally steady. Conscientiousness is the one Big Five trait with only a slight gender difference: men were marginally above the mean and women marginally below, suggesting men were seen as just a bit more organized or disciplined, but this gap is very small.

All Dark Triad traits are strongly differentiated by gender in these personas. Men are assigned higher dark-trait scores across the board. On average, male personas score about 0.5–0.6 standard deviations higher in Machiavellianism than females (male M roughly +0.3 vs female M about –0.3). Likewise, male Psychopathy is higher by roughly 0.36 *z* (male PSY around +0.18 vs female PSY –0.18). Narcissism shows a smaller gap (male NAR slightly above 0, female NAR slightly below 0), but even this difference is statistically reliable. These results indicate that the LLM frequently imbued male characters with more manipulative, self-focused, and callous traits compared to female characters, who were conversely depicted as less antagonistic and more pro-social.

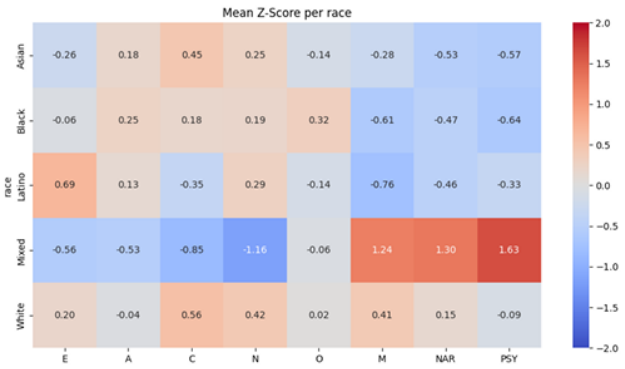


Figure 6: Mean Z-score per race. Heatmap of average standardized trait scores for personas of five racial categories (Asian, Black, Latino, Mixed, White). Trait abbreviations and color scale as before.

The visualization below in Figure 7 corroborates these differences. Panel B of Figure 7 displays the distribution of raw trait scores by gender, confirming systematic shifts: for each trait, the female distributions (orange boxplots) are centered at different levels than the male distributions (blue boxplots). For example, in Agreeableness, the female box is centered higher than the male box (most women personas scored more agreeable than most men), while in Neuroticism the male box is much higher than the female box (many male personas had high N scores, whereas female personas tended to have low N). Traits with large mean differences (like N, M, A) show clearly separated boxplot centers, whereas traits with smaller differences (like C, NAR) still have overlapping distributions but distinct averages. Panel C quantifies the mean gender differences (male minus female) in trait Z-scores with a bar chart. Each gray bar extending to the right indicates a higher male mean, and to the left a higher female mean; *p*-values from statistical tests are annotated. All traits show a significant difference ($p<0.05$) between male and female personas. The largest gaps are observed in Neuroticism and Openness (males much higher in N, females much higher in O, both with $p<0.001$), followed by Machiavellianism and Agreeableness (males higher in M, females in A, also highly significant). Psychopathy and Extraversion differences (males > females) are somewhat smaller but still clearly significant, and even the subtle differences in Conscientiousness and Narcissism reach significance. In sum, the persona dataset reveals a consistent gender-stereotypical pattern: male personas are generally portrayed as more extraverted, more neurotic, and higher on antagonistic/dark traits (M, NAR, PSY), whereas female personas are portrayed as more agreeable, more open, less neurotic, and lower on those dark traits.

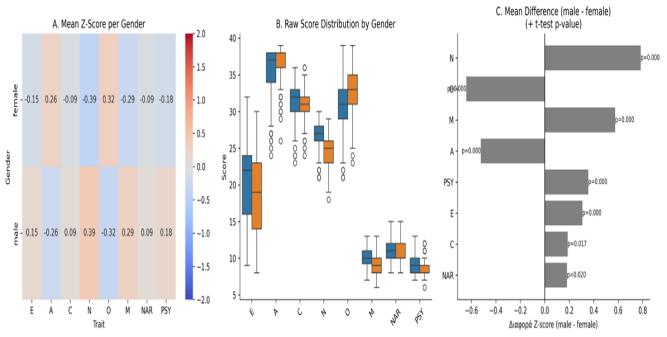


Figure 7: Gender differences in trait scores. Panel A – Heatmap of mean Z-scores for Female vs Male personas on each trait. Panel B – Boxplot distributions of raw trait scores by gender (blue = male, orange = female) for each trait (Big Five and Dark Triad). Panel C – Mean difference (male minus female) in Z-scores for each trait.

4.5. Intersections and PCA Clustering

To visualize how these trait biases combine and whether distinct demographic profiles cluster together, we performed a principal component analysis across all persona trait profiles. Figure 8 shows a scatter plot of all personas in the space of the first two principal components (PC1 vs PC2), with each point colored by race and marked by gender. Several clear patterns emerge. Race-based clustering is evident, particularly for the Mixed-race personas (purple points): they are widely separated from the rest, often occupying extreme positions in the plot. Many Mixed persona points lie far out on the rightmost end of PC1 or high on PC2, forming a distinct cloud largely isolated from other races. This reflects our earlier observation that Mixed-race profiles have extreme trait values (especially very high dark traits), which drive them to the periphery of the PCA space. For example, the cluster of purple symbols on the far right corresponds to Mixed personas with exceptionally high Machiavellianism/Narcissism/Psychopathy scores (traits likely loading heavily on PC1), while a subset of purple points that rise to the top of the chart represents Mixed personas that are outliers on a second combination of traits (perhaps those with unusual Big Five patterns contributing to a high PC2). A few of these extreme outliers are labeled by index in the figure, underscoring how far removed they are from the central mass of points.

In contrast, personas of other races (White, Black, Asian, Latino) tend to cluster nearer to the origin of the PCA plot and overlap considerably with each other. The dense central cloud of points (PC1 and PC2 values both near 0) is a mix of blue, orange, green, and red markers, indicating that White, Black, Asian, and Latino personas share a broadly similar trait space without forming wholly distinct clusters in the first two principal components. There are subtle tendencies—for instance, many Latino personas (red) appear slightly toward the left side of the central cluster (somewhat negative on PC1), whereas White (blue) and Asian (green) personas are more dispersed around the middle, and Black personas (orange) intermingle throughout. However, these differences are gradual and overlapping; no single non-Mixed race forms an isolated grouping in this 2D projection. This suggests that aside from the Mixed category, racial trait differences are more a matter of degree than completely separate categories, with significant commonality among White, Black, Asian, and Latino personas in how the model represents their trait combinations.

Gender, indicated by shape (circles for male ● vs crosses for female ×), does not produce starkly separate clusters in the PCA plot. Male and female personas broadly overlap in this trait space, consistent with the fact that the gender differences we observed — although significant — involve opposing shifts on multiple traits that don't align neatly along a single principal axis. In Figure 6, male and female symbols of the same color are generally intermixed rather than split apart. For example, blue crosses and blue circles (female vs male White personas) are distributed in a similar area, and the same holds for other races (e.g., orange crosses and circles for Black personas largely coincide). This indicates that within each racial group, the gender-based trait offsets (e.g., females having slightly higher A and O, males higher M and N, etc.) add some scatter but do not create a separate "male persona cluster" distinct from a "female persona cluster." The within-race variability – especially the extreme outlier status of certain races like Mixed – dominates the first two PCs.

That said, there are minor interaction effects visible. Within the Mixed-race cluster, female Mixed personas (purple ×) tend to concentrate a bit higher on the PC2 axis, whereas male Mixed personas (purple ●) extend further on PC1. This suggests that for Mixed-race characters, being male vs female leads to slightly different extreme trait manifestations: for instance, a Mixed male persona might combine the strong negative racial stereotype (Mixed: very low Big Five, very high dark traits) with the male-associated higher dark traits, yielding an especially extreme point far out on the PC1 dimension; a Mixed female, while still an outlier, may be somewhat tempered in dark traits (since females had lower dark scores) but could differ in another way (perhaps lower Neuroticism or higher emotional stability relative to Mixed males), pulling her profile in a slightly different direction (higher on PC2). Outside of the Mixed group, most other race–gender combinations do not produce clearly separable sub-clusters; the male–female differences within White, Black, Asian, and Latino groups appear as small shifts around a common central cluster for each race. Overall, the PCA visualization reinforces that race-based variations (the outlying nature of Mixed-race personas) are the primary driver of dispersion in trait space, while gender differences, though systematic, contribute more to fine-

scale variation within each racial cluster rather than forming entirely distinct groupings on the global map.

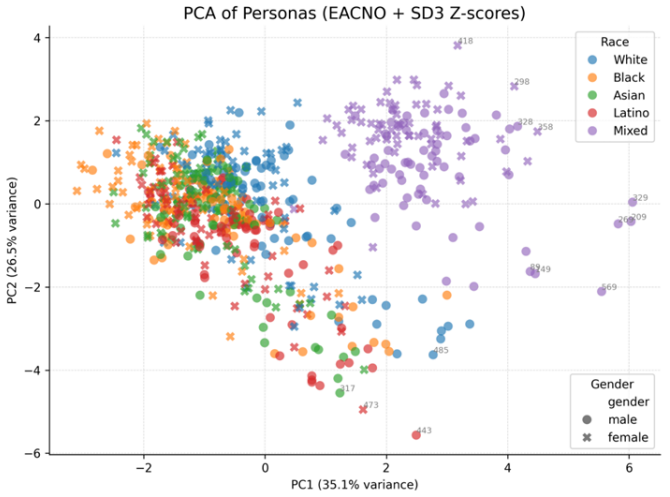


Figure 8: PCA of personas by race and gender. Scatter plot of persona trait profiles projected onto the first two principal components (PC1 and PC2, capturing ~61.6% of variance).

4.6. Internal Trait Correlations

The relationships among all the personality traits in this persona dataset provide insight into how traits tend to co-occur in the model's outputs**.** Figure 9 below shows the correlation matrix for every pair of traits. Several salient patterns stand out. Within the Big Five traits (the upper-left 5×5 block of the matrix), most correlations are positive, meaning that if a persona is high on one of these desirable traits, the model often also assigns higher levels on others. Notably, Agreeableness (A) strongly co-occurs with Openness (O) and Conscientiousness (C) (with Pearson *r* of roughly +0.70 for A–O and +0.54 for A–C). This indicates that more agreeable personas are also often portrayed as substantially more open-minded and responsible. Conscientiousness in turn has a moderate positive correlation with Openness (r ≈ +0.44). These inter-correlations (A–C–O) suggest a "bundle" of positive traits in the dataset: many personas score high (or low) simultaneously on these three dimensions. Other Big Five pairs show weaker links; for example, Extraversion (E) is almost uncorrelated with Conscientiousness or Openness, and it has a slight negative correlation with Agreeableness (in this data, more extraverted characters were, if anything, a bit less agreeable, though the effect is small). Interestingly, Neuroticism (N) is nearly uncorrelated with most other Big Five traits here (its correlations with E, A, and C are close to zero). In short, aside from the cohesive cluster of A, C, and O moving together, the Big Five trait correlations are modest in magnitude.

By contrast, the Dark Triad traits show very strong mutual correlations. Machiavellianism, Narcissism, and Psychopathy are all positively interrelated, reflecting that personas who are high in one "dark" trait tend to be high in the others as well. The correlation between Machiavellianism (M) and Psychopathy (PSY) is especially high (r ≈ +0.63), and Machiavellianism also correlates around +0.60 with Narcissism (NAR). The NAR–PSY correlation is slightly lower (around +0.57) but still strong. This trio of high inter-correlations (the bright red block in the Dark Triad section of the matrix) indicates that the model often assigns all three dark traits in tandem — i.e. when it creates a manipulative persona, that character is also likely to be narcissistic and somewhat psychopathic in the portrayal. This is consistent with earlier observations that certain demographic groups (like Mixed-race or male personas) tended to receive uniformly high dark trait scores.

Looking at cross-domain relationships (Big Five vs. Dark Triad), we observe a clear inverse pattern between pro-social personality traits and the dark traits. Agreeableness has substantial negative correlations with Machiavellianism and Psychopathy (r ≈ –0.37 and –0.41, respectively). In other words, more agreeable (kind, empathetic) characters are much less likely to be portrayed as manipulative or callous. Conscientiousness likewise correlates negatively with Psychopathy (around –0.41), indicating that diligent, rule-abiding personas tend not to have psychopathic tendencies in the model's depiction. Neuroticism shows a moderately strong negative correlation with Narcissism (r ≈ –0.42), suggesting that personas who are very narcissistic (self-important and confident) are often simultaneously depicted as emotionally stable (low N) rather than anxious – hinting that the model may associate narcissistic personalities with a kind of unshakeable confidence. Openness and Extraversion have weaker or mixed relationships with dark traits (most of those correlations hover near zero or a slight negative). One subtle finding is a slight positive correlation between Openness and Narcissism (r ~ +0.17), which implies that some highly open/intellectual personas were also given a hint of self-importance by the model. Additionally, Agreeableness versus Narcissism shows a very small positive *r* (~+0.12), meaning that unlike Machiavellianism and Psychopathy (which strongly conflict with Agreeableness), Narcissism in this dataset was not strongly anti-correlated with being agreeable – a persona could be somewhat agreeable and yet narcissistic (perhaps reflecting stereotypes of charming, sociable narcissists). Nonetheless, the dominant trend is that high dark-trait personas tend to score low on Agreeableness and Conscientiousness (seen in the blue-colored cells for A–M, A–PSY, C–PSY in Figure 7), reinforcing that benevolent personality characteristics are inversely related to antagonistic ones in the model's representation.

Overall, the correlation analysis confirms internally consistent patterns in the LLM's persona outputs. Positive personality traits align together and generally oppose the dark traits, while the Dark Triad traits form their own tight-knit cluster. These results provide a complementary

perspective on the trait structure underlying the demographic biases described above, demonstrating that the model's stereotypical persona attributions are not random but follow logical relationships (e.g., "kindness" versus "cruelty" as opposing poles, and certain positive traits tending to go hand-in-hand).

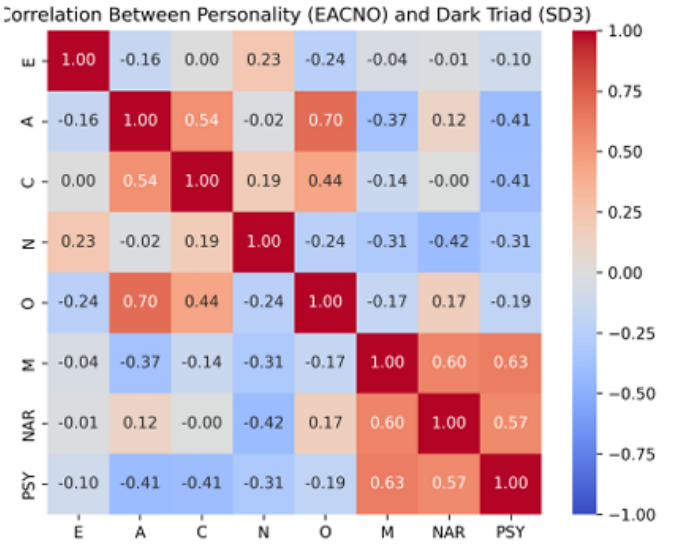


Figure 9: Correlation matrix of all traits. Pearson correlation coefficients between each pair of traits (Big Five: E, A, C, N, O; Dark Triad: M, NAR, PSY), computed across all persona scores. The matrix is symmetric; only one triangle is annotated with r values for clarity. Red indicates a positive correlation; blue indicates a negative correlation (scale shown on right).

5. Discussion

5.1. Cognitive and Psychological Interpretation

The observed patterns suggest that LLMs have developed internal cognitive-like representations of human groups, shaped by the statistical regularities of language. Although LLMs lack consciousness or intention, their training on vast human text corpora implicitly encodes societal narratives — producing what may be described as *synthetic cognition*. Unlike studies comparing AI to human baselines, our approach intentionally isolates this 'synthetic cognition' as a closed system. By focusing exclusively on the internal consistency of the model's generated personas, we map the algorithm's inherent stereotypical landscape without the confounding noise of human cultural variance.

The model's ability to assign coherent and demographically consistent personality profiles indicates that its latent representations capture more than linguistic associations: they embody social schemas. These schemas operate analogously to human stereotypes — simplifying complex social realities into categorical personality assumptions.

For instance:

- The "Western male atheist" archetype characterized by high *Openness* and *Narcissism*,
- The "Asian female Buddhist" with high *Conscientiousness* and low *Extraversion*, and
- The "Black male Christian" with high *Extraversion* and *Agreeableness*

demonstrate that the model generalizes culturally learned personality scripts.

Such patterns align with social cognition theory, which posits that stereotypes arise from heuristic associations rather than explicit reasoning. In this sense, the LLM functions as a large-scale mirror of human collective cognition — reproducing implicit personality prototypes learned from text [12].

5.2. Theoretical and Methodological Implications

From a methodological standpoint, this study bridges computational psychometrics and AI fairness auditing. Traditional bias research focuses on overt lexical or sentiment asymmetries (e.g., word embeddings associating "doctor" with male pronouns). Here, the bias operates at a *latent psychometric layer*, revealing how models attribute moral and emotional structure to demographic identities.

This framework contributes to the field by:

1. Introducing quantitative psychometric elicitation as a fairness diagnostic tool.
2. Demonstrating that *demographic conditioning* can alter inter-trait correlations — a deeper structural form of bias than mere mean-level differences.
3. Showing that bias can be interpreted through psychological theory, not just mathematical metrics.

Methodologically, it establishes a reproducible paradigm: using validated personality inventories (Big Five and Dark Triad), persona conditioning, and statistical normalization to extract interpretable cognitive maps from LLMs. This approach can be generalized to future studies exploring emotion, values, or moral reasoning biases in generative AI systems [13].

5.3. *Ethical and Societal Considerations*

The findings highlight serious ethical challenges. If LLMs systematically attribute moral or emotional traits based on identity cues, they risk reinforcing psychological stereotypes — subtle yet powerful forms of bias that influence downstream applications such as:

- Conversational AI: tone and empathy variation depending on user demographics;
- Hiring or profiling tools: skewed personality assessments;
- Education and therapy simulations: biased affective responses toward different identities.

- Practical Applications of Psychometric Auditing: Our framework could be extended to real-world applications beyond academic auditing. For example, it offers a method for monitoring racial bias trends in social media moderation systems, ensuring that automated agents do not attribute 'aggressive' or 'toxic' personality traits to users based on dialect or demographic markers. Furthermore, in the domain of healthcare, this methodology is critical for calibrating therapeutic LLMs. By detecting latent psychometric biases early, developers can fine-tune models to ensure they function equitably across diverse socio-economic and cultural backgrounds, preventing scenarios where an AI therapist might unconsciously adopt a colder or less empathetic persona toward marginalized groups."

Unlike explicit hate speech or toxicity, psychometric bias is invisible — it manifests through tone, moral emphasis, and perceived emotional intelligence. Because these models are often used in socially sensitive domains, their internal personality framing can affect fairness and trustworthiness.

To mitigate this, ethical AI development should include:

1. Psychometric fairness auditing — evaluating personality-related patterns alongside linguistic bias tests;
2. Data transparency — documenting sociocultural composition of training corpora;
3. Debiasing interventions — such as identity-neutral conditioning or fairness-aligned fine-tuning;
4. Human-in-the-loop oversight, ensuring that cultural interpretation does not reinforce stereotypes.

This work thus positions psychometric bias as a critical dimension of AI moral responsibility.

5.4. Limitations and Future Directions

Despite the robust methodology, several limitations must be acknowledged:

- Synthetic Personas: The personas simulate averaged demographic archetypes rather than real individuals, which limits ecological validity. However, this abstraction isolates model bias more effectively by removing user variance.
- Single-Model Scope: The experiments presented in the main analysis were conducted using one LLM (LLaMA-3.1-8B-Instruct). To assess whether the observed bias patterns are model-specific, we conducted preliminary exploratory experiments with additional models, including Mistral-7B-Instruct. These initial observations indicated qualitatively similar trends in demographic bias attribution, suggesting that the findings are not unique to a single model architecture. However, a comprehensive cross-model validation, including proprietary models (e.g., GPT-4, Claude), is left as future work to determine the full extent of generalizability.
- Cultural Bias in Training Data: Because most pretraining text is in English, Western cultural norms dominate personality attributions. Extending this framework to multilingual LLMs could reveal cross-linguistic differences in psychometric stereotypes.
- Simplified Gender Variable: The binary male/female classification omits non-binary or gender-fluid identities, which may yield additional insight into model fairness.
- Lack of Human Benchmark: Although psychometric consistency was verified statistically, future work could compare LLM-generated profiles with human survey data to evaluate alignment.

Despite these limitations, the study establishes a foundational approach for examining how artificial cognition reflects human moral structure, offering a blueprint for next-generation bias auditing techniques [14], [6].

6. Conclusion and Future Work

This study introduced a novel framework for eliciting demographic stereotypes in Large Language Models (LLMs) through the lens of psychometric attribution. By combining established personality frameworks — the Big Five (EACNO) and the Dark Triad (SD3) — with systematic persona conditioning, we demonstrated that LLMs generate consistent, demographically structured personality profiles. These results provide compelling evidence that bias in LLMs extends beyond language or sentiment: it manifests at a cognitive level, where identity cues shape the model's perception of personality, morality, and social behavior.

Through large-scale experimentation across 660 personas, encompassing 11 regions, 5 racial groups, 6 religions, and 2 genders, the study revealed reproducible cross-group differences in both prosocial (Big Five) and antisocial (Dark Triad) traits. The model attributed:

- Higher *Agreeableness* and *Conscientiousness* to religious and female personas,
- Higher *Openness* and *Narcissism* to secular and Western personas,
- Greater *Machiavellianism* and *Emotional Restraint* to Asian personas,
- and elevated *Extraversion* and *Warmth* to African and Latin American personas.

These psychometric signatures were statistically significant and internally coherent, forming a structured "map of social cognition" embedded in the model's latent space.

In essence, the LLM acts as a mirror of collective cultural perception, reproducing personality stereotypes as learned from global human discourse.

From a theoretical standpoint, this work advances the field of computational psychometrics by framing model bias as a form of *synthetic cognition*. Rather than treating bias as a statistical defect, it reinterprets it as a *psychological phenomenon* — a window into how artificial systems internalize and reproduce the cognitive heuristics of human societies.

6.1. Key Contributions

1. Methodological Innovation: A reproducible Python-based pipeline for psychometric elicitation and statistical evaluation of demographic bias in LLMs.
2. Theoretical Integration: A bridge between AI fairness research, social psychology, and computational personality modeling.
3. Empirical Findings: Systematic personality and moral asymmetries across demographic factors, consistent with known cultural stereotypes.
4. Ethical Insight: Demonstration that fairness in LLMs must account for *psychological bias*, not only linguistic or representational bias.

6.2. Future Work

The present study opens several avenues for future research:

1. Cross-Model Validation: Extending the same pipeline to multiple LLM architectures (GPT-4, Claude, Gemini, Mistral) will reveal whether psychometric biases are *architecture-dependent* or *data-universal*.
2. Temporal and Cultural Drift: Investigating how model personality attributions evolve with new training data or fine-tuning cycles could expose *bias drift* over time.
3. Multilingual and Cross-Lingual Evaluation: Applying the framework to multilingual models may uncover differences in cultural stereotypes encoded across languages. This could lead to *comparative cultural cognition* analysis in AI.
4. Inclusion of Non-Binary and Intersectional Identities: Expanding demographic variables to include non-binary gender, mixed-religious backgrounds, and socioeconomic class will capture deeper intersectional complexity.
5. Human Benchmarking: Comparing LLM-generated profiles with actual psychometric data from human respondents can assess the degree of *alignment* between artificial and human stereotype structures.
6. Bias Mitigation Techniques: Implementing bias-aware fine-tuning, counter-stereotypical persona training, and identity-neutral prompts could reduce psychometric distortion in model responses.

6.3. Final Remarks

The findings underscore a profound insight:

Large Language Models do not merely learn language — they learn society.

Their responses reveal a computational echo of human cognition, complete with virtues, flaws, and stereotypes. However, the implications of these findings reach far beyond technical correctness. As LLMs are increasingly integrated into decision-support systems for hiring, lending, and legal judgment, the implicit attribution of 'dark' or 'unstable' traits to specific demographics poses a tangible risk of algorithmic discrimination. If a model inherently views certain groups as less conscientious or more manipulative, this cognitive bias can cascade into material harm—denying opportunities or reinforcing systemic inequalities. Therefore, psychometric fairness is not merely a metric for model performance, but a safeguard for social justice in the age of artificial intelligence. The ultimate goal is to develop AI systems that reflect human diversity without reproducing human prejudice—systems that understand personality without imposing it. This study provides one step toward that vision, offering a reproducible foundation for exploring the psychology of artificial intelligence.

Ethical Disclosure

This research explicitly analyzes the generation of harmful stereotypes by AI systems. We acknowledge that some of the model-generated profiles reported—particularly those associating specific racial or religious groups with negative traits—contain offensive and discriminatory content. These outputs are presented solely for the purpose of scientific auditing and critique. The authors explicitly condemn these stereotypes and clarify that the demographic labels employed in this study (e.g., race, gender) are used as operational variables to probe the model's latent space, without implying essentialist definitions of complex human identities.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The author gratefully acknowledges the academic and technical support provided by colleagues and research collaborators during the design and implementation of this study. The experiments were conducted on locally maintained hardware resources, ensuring full reproducibility and data privacy.

No external funding was received for this work.

References

[1] I. O. Gallegos et al., "Bias and fairness in large language models: A survey," *arXiv preprint arXiv:2309.00770, 2023*.

[2] National Institute of Standards and Technology, "Towards a standard for identifying and managing bias in artificial intelligence," *NIST Special Publication 1270*, Gaithersburg, MD, 2023.

[3] X. Bai, A. Wang, I. Sucholutsky, and T. L. Griffiths, "Explicitly unbiased large language models still form biased associations," *Proceedings of the National Academy of Sciences*, vol. 122, no. 8, p. e2416228122, 2025, doi: 10.1073/pnas.2416228122.

[4] O. Gupta, S. Marrone, F. Gargiulo, R. Jaiswal, and L. Marassi, "Understanding social biases in large language models," *AI*, vol. 6, no. 5, p. 106, 2025, doi: 10.3390/ai6050106.

[5] S. Lee et al., "Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, doi: 10.48550/arXiv.2406.14703.

[6] OpenAI, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774, 2023*.

[7] L. P. Argyle et al., "Out of one, many: Using language models to simulate human samples," *Political Analysis*, vol. 31, no. 3, pp. 337–351, 2023, doi: 10.1017/pan.2023.2.

[8] D. Dodou, J. C. F. de Winter, and T. Driessen, "The use of ChatGPT for personality research: Administering questionnaires using generated personas," *Personality and Individual Differences*, vol. 228, p. 112729, 2024, doi: 10.1016/j.paid.2024.112729.

[9] M. I. Radaideh, O. H. Kwon, and M. I. Radaideh, "Fairness and social bias quantification in large language models for sentiment analysis," *Knowledge-Based Systems*, vol. 319, p. 113569, 2025, doi: 10.1016/j.knosys.2025.113569.

[10] D. S. Porat and E. Rabinovich, "Who are you, ChatGPT? Personality and demographic style in LLM-generated content," *arXiv preprint arXiv:2510.11434*, 2025.

[11] S. Wang et al., "Exploring the impact of personality traits on LLM bias and toxicity," *arXiv preprint arXiv:2502.12566*, 2025.

[12] H. Peters and S. C. Matz, "Large language models can infer psychological dispositions of social media users," *PNAS Nexus*, vol. 3, no. 6, p. pgae231, 2024, doi: 10.1093/pnasnexus/pgae231.

[13] F. A. Tan et al., "PHAnToM: Persona-based prompting has an effect on theory-of-mind reasoning in large language models," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2025)*, 2025.

[14] T. Sühr, F. E. Dörner, S. Samadi, and A. Kelava, "Challenging the validity of personality tests for large language models," *arXiv preprint arXiv:2311.10805*, 2023.

NIKOLAOS VASILEIOS OIKONOMOU is a Computer & Network Engineer, as well as an academic researcher and Ph.D. candidate in the Department of Informatics and Telecommunications at the University of Ioannina, from which he also received his B.Eng. and M.Sc. degrees. In parallel to his academic work, he serves as a private Computer Science educator and possesses several years of professional experience as a Software Developer, IT Specialist, and Network Consultant.

IOANNIS PALAIOKRASSAS is pursuing a M.Eng. degree in Computer Science and Engineering at the University of Ioannina and serves as an active research member. He is currently employed in web development.

DIMITRIOS VASILEIOS OIKONOMOU obtained his B.Sc. in Regional and Cross-Border Studies from the University of Western Macedonia in 2024. He is currently engaged in research activities at the same institution and is pursuing an M.Sc. in e-Business and Digital Marketing.

SOFIA PANAGIOTA CHALIASOU is pursuing a B.Sc. in Informatics at the Hellenic Open University and serves as an active research associate. She also holds a Vocational Diploma in Web Design and Development. In her professional capacity, she is currently employed in sales and possesses prior professional experience as a web developer.

NIKOLAOS RIGAS obtained his B.Sc. in Regional and Cross-Border Studies from the University of Western Macedonia in 2025. He is currently pursuing an M.Sc. in "Criminological and Penal Law perspectives on Corruption, Economic and Organized Crime" at the Hellenic Open University, while actively engaged in research activities.

Received: 30 October 2025, Revised: 10 January 2026, Accepted: 13 January 2026, Online: 22 January 2026

DOI: https://doi.org/10.55708/js0501006

Binary Image Classification with CNNs, Transfer Learning and Classical Models

Nikolaos Vasileios Oikonomou*[1], Dimitrios Vasileios Oikonomou[2], Sofia Panagiota Chaliasou[3], Nikolaos Rigas[4]

[1]Department of Informatics & Telecommunications, University of Ioannina, Arta, 47150, Greece
[2]Department of Management Science & Technology, University of Western Macedonia, Kozani, 50100, Greece
[3]Department of Informatics, Hellenic Open University, Patras, 26335, Greece
[4]Department of Social Sciences, Hellenic Open University, Patras, 26335, Greece
Email(s): haikos13@gmail.com (N. V. Oikonomou), ecomimis@gmail.com (D. V. Oikonomou), sofia.xaliasou12@gmail.com (S. P. Chaliasou), nickrigas7@hotmail.com (N. Rigas)
*Corresponding author: Nikolaos Vasileios Oikonomou, University of Ioannina Department of Informatics & Telecommunications, haikos13@gmail.com

ABSTRACT: This study presents a comprehensive comparative analysis of binary face classification utilizing Deep Learning and traditional Machine Learning approaches. We evaluate three distinct modeling strategies: (1) End-to-end Convolutional Neural Networks (CNNs), including a baseline TensorFlow model and an optimized PyTorch architecture; (2) Hybrid CNN-MLP networks; and (3) Feature extraction via a pre-trained ResNet50 coupled with classical classifiers (Random Forest, Logistic Regression). The experimental dataset consists of 6,376 face images (5,102 training, 1,274 validation) derived from a Kaggle challenge. We implement rigorous data augmentation (rotation, shifts, flips) and regularization techniques (Dropout, Batch Normalization, Weight Decay) to mitigate overfitting. Results demonstrate that the optimized PyTorch CNN achieved the highest generalization performance with a validation accuracy of ~85.9% and an AUC of 0.94, utilizing AdamW optimizer and Cosine Annealing scheduling. Conversely, the classical models (Random Forest, Logistic Regression) utilizing ResNet50 features exhibited near-perfect training metrics (AUC ≈ 1.0) and competitive validation accuracy (>90%), highlighting the efficacy of transfer learning. We critically analyze the "underfitting" phenomenon observed in the baseline CNN (Training Accuracy < Validation Accuracy) attributing it to aggressive regularization. This work provides a clear roadmap for selecting between computational-heavy deep architectures and efficient feature-based classical models based on available resources and accuracy requirements.

KEYWORDS: AUC-ROC, Binary Image Classification, Convolutional Neural Networks (CNNs), Data Augmentation, Feature Extraction, Logistic Regression, PyTorch, Random Forest, ResNet50, Transfer Learning

1. Introduction

Automated image classification has evolved into a central problem in computer vision, driven by the need to efficiently process vast amounts of visual data in applications ranging from biometric security to emotion recognition. While early approaches relied on handcrafted features, the advent of Convolutional Neural Networks (CNNs) revolutionized the field by enabling models to learn hierarchical feature representations directly from raw pixel data. Seminal architectures such as AlexNet and ResNet demonstrated that deep networks, utilizing techniques like ReLU activations and Dropout, could achieve breakthrough accuracy on massive datasets like ImageNet [1], [2].

However, deploying deep learning models for specific tasks, such as binary face classification, presents significant challenges. Training deep architectures from scratch requires substantial computational resources, large, labeled datasets, and meticulous hyperparameter tuning to avoid overfitting. Conversely, Transfer Learning

strategies, which leverage pre-trained networks (e.g., ResNet50) as feature extractors, offer a compelling alternative by transferring knowledge from generic domains to specific tasks [3]. When combined with classical machine learning classifiers like Random Forests (RF) or Logistic Regression (LR), these approaches can potentially offer a balance between high accuracy and low training cost [4].

The primary motivation of this study is to navigate the trade-offs between computationally intensive End-to-End Deep Learning and efficient Feature-Based Classical Learning in the context of binary face classification. Specifically, we address the challenge of classifying face images into two categories using a dataset derived from a Kaggle challenge. A key issue investigated is the phenomenon of model generalization versus overfitting: while complex CNNs often struggle with underfitting when heavily regularized, feature-based classical models may exhibit near-perfect training accuracy but varying degrees of validation performance. The overall architectural pipeline designed for this comparative study, encompassing the end-to-end, hybrid, and transfer learning workflows, is visually summarized in Figure 1.

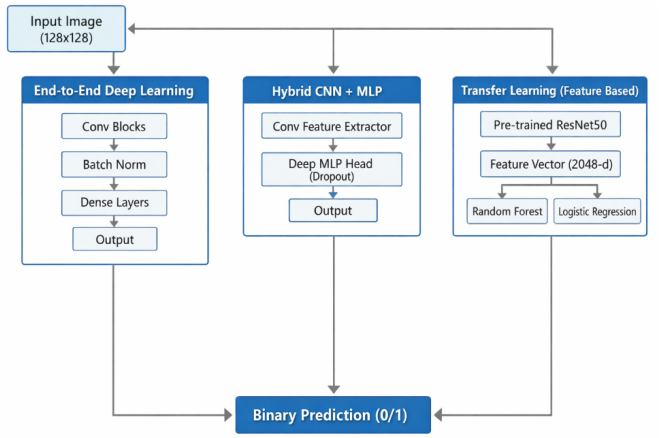


Figure 1: System overview illustrating the three comparative methodologies employed in this study: (a) End-to-End CNN training (Baseline & Optimized), (b) Hybrid CNN-MLP architecture, and (c) Feature extraction using pre-trained ResNet50 combined with classical classifiers (Random Forest, Logistic Regression).

Unlike previous studies that typically focus on a single modeling paradigm, this work provides a rigorous comparative analysis across three distinct methodologies: (1) End-to-End CNNs (comparing a baseline TensorFlow implementation against an optimized PyTorch pipeline), (2) Hybrid CNN-MLP architectures, and (3) Transfer Learning coupled with classical classifiers.

The specific contributions of this paper are as follows:

- Framework and Optimization Analysis: We explicitly compare a standard TensorFlow/Keras baseline against a custom PyTorch-based CNN ("MyDeepCNN"). We demonstrate that the superior performance of the latter is driven not merely by the framework, but by an optimized training pipeline incorporating AdamW optimizer, Cosine Annealing learning-rate scheduling, and LeakyReLU activations.
- Evaluation of Feature-Based Classifiers: We show that combining deep features from a pre-trained ResNet50 with traditional classifiers (Random Forest, Logistic Regression) yields competitive or superior validation accuracy (>90%) and AUC (>0.97) compared to end-to-end CNNs, with a fraction of the training time.
- Critical Analysis of Overfitting/Underfitting: We provide a detailed examination of the training dynamics, explaining the "underfitting paradox" observed in the baseline CNN (where training accuracy lags behind in validation accuracy due to heavy augmentation) versus the massive capacity of Random Forests to fit training data perfectly.

The remainder of this paper is organized as follows: Section 2 reviews the theoretical background of CNNs, transfer learning, and regularization techniques. Section 3 details the dataset, preprocessing steps, and the specific architectures implemented (Baseline CNN, Hybrid Models, and Feature Extraction pipelines). Section 4 presents the experimental results, including metrics such as Accuracy, F1-score, and AUC-ROC, alongside confusion matrices. Section 5 discusses the findings, analyzing the trade-offs between deep and classical methods. Finally, Section 6 concludes the study and suggests directions for future research.

2. Theoretical Background

2.1. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) have established themselves as the dominant architecture for visual recognition tasks due to their ability to capture spatial hierarchies in images through learnable filters [1]. A typical CNN architecture comprises alternating convolutional layers, which extract local features (e.g., edges, textures), and pooling layers that reduce spatial dimensionality and computational load.

To facilitate the training of deep architectures, modern CNNs incorporate Batch Normalization (BN). BN normalizes the inputs of each layer layer-wise, mitigating the problem of internal covariate shift. This allows for higher learning rates and acts as a regularizer, significantly accelerating convergence [5]. Furthermore, activation functions such as the Rectified Linear Unit (ReLU) and its variants (e.g., LeakyReLU) are standard in overcoming the vanishing gradient problem in deep networks [1].

2.2. Transfer Learning and Pre-trained Models

Training deep CNNs from scratch requires massive datasets to avoid overfitting. Transfer Learning addresses this by leveraging models pre-trained on large-scale

datasets (e.g., ImageNet) to solve related tasks. Research indicates that the initial layers of CNNs learn generic features applicable across diverse domains, while deeper layers capture task-specific semantics [3].

In this study, we utilize ResNet50 [2], a 50-layer residual network, as a feature extractor. We also reference modern efficient architectures like EfficientNet, which optimize accuracy and efficiency through compound scaling [6], as benchmarks for future improvements. By extracting feature vectors from the penultimate layer of these models, we transform raw images into high-level semantic embeddings suitable for classical classification.

2.3. Classical Machine Learning Classifiers

While Deep Learning excels in end-to-end feature learning, classical classifiers offer advantages in interpretability and training speed when provided with high-quality features.

- Random Forest (RF): An ensemble learning method that constructs a multitude of decision trees at training time. It is robust to noise and overfitting (compared to individual decision trees) and provides insights into feature importance [4].
- Logistic Regression (LR): A linear model that estimates the probability of a binary outcome using the sigmoid function. Despite its simplicity, LR can achieve state-of-the-art performance when the input feature space (e.g., from ResNet50) is linearly separable.

2.4. Regularization and Optimization

To ensure robust generalization, especially in binary face classification where datasets may be limited, rigorous regularization strategies are essential.

- Data Augmentation: We employ geometric transformations (rotation, flipping, zooming) to artificially expand the training set and enforce invariance to input variations. Recent surveys highlight augmentation as a critical component for successful deep learning training [7].
- Optimization Algorithms: We utilize the Adam optimizer (Adaptive Moment Estimation), which computes adaptive learning rates for each parameter. For our optimized PyTorch model, we specifically use AdamW, a variant that decouples weight decay from gradient updates, offering superior regularization for deep models [8].

3. Methodology

3.1. Dataset Description and Preprocessing

The experimental evaluation utilized a dataset derived from the "Kaggle Face Classification Challenge", specifically curated for binary classification tasks (Class 0 versus Class 1) [9].

- Data Volume and Splitting: The complete dataset comprises 6,376 facial images. To ensure robust model evaluation and prevent data leakage, we employed a stratified splitting strategy. The data was partitioned into a Training Set of 5,102 images (approx. 80%) and a Validation Set of 1,274 images (approx. 20%). Stratification was strictly applied to maintain the same class distribution in both subsets as in the original dataset, preventing bias during the training phase [10].
- Image Properties: The original images varied in resolution and lighting conditions. To standardize the input for the neural networks, all images were resized to fixed spatial dimensions of 128 X 128 pixels with 3 color channels (RGB).
- Normalization: Prior to ingestion by the models, pixel intensity values were scaled from the integer range [0, 255] to the floating-point range [0, 1] by dividing by 255.0. This normalization step is critical for ensuring numerical stability and accelerating gradient descent convergence by keeping the input values within a small, bounded range [11].

3.2. Data Augmentation Strategy

Given the restricted size of the training dataset (N ≈5,100), the risk of the model memorizing specific training examples (overfitting) rather than learning generalizable features was significant. To address this, we implemented a robust Online Data Augmentation pipeline. Unlike offline augmentation, which expands the dataset size statically, online augmentation applies stochastic transformations to each batch of images dynamically during training. This ensures that the network never encounters the exact same image tensor twice, effectively simulating a vastly larger dataset [12].

The augmentation policy was carefully designed to simulate realistic variations in facial pose and lighting conditions without altering the semantic label of the image. The specific transformations applied are categorized as follows:

3.2.1. Geometric Transformations

- Random Rotations: Images were rotated by a random angle θ sampled from the uniform distribution $\theta \sim U(-20°, +20°)$. This encourages the model to learn rotation-invariant features, accommodating slight head tilts common in real-world photography.
- Spatial Shifts: We applied random horizontal and vertical translations (width/height shifts) with a shift range factor of 0.2, forcing the convolutional filters to recognize facial features regardless of their absolute position in the frame.

- Flipping: Random Horizontal Flips were applied with a probability of p=0.5. Vertical flips were also experimented with to further increase diversity, although less common in natural facial alignment.
- Affine Perturbations: Random shearing and zooming transformations were utilized to simulate variations in camera perspective and subject distance [13].

3.2.2. *Photometric and Noise Injections*

- Color Jittering: To prevent the model from relying on specific lighting cues or skin tone over-saturation, we applied random perturbations to the brightness, contrast, and saturation of the input images.
- Random Resized Crop: In the advanced PyTorch implementation ("MyDeepCNN"), we utilized RandomResizedCrop, which extracts a random patch of the image and resizes it to the target dimensions (128 X 128). This forces the network to classify based on local features (e.g., eyes, nose) rather than just the global face structure.

This extensive augmentation strategy served as a strong regularizer, complementing the Dropout layers and Weight Decay described in subsequent sections.

3.3. *Experimental Implementation and Environment*

To ensure the reliability and reproducibility of our results, all experiments were conducted within a controlled computational environment. The implementation was entirely developed in the **Python** programming language, utilizing a suite of open-source libraries optimized for scientific computing and deep learning.

3.3.1. *Deep Learning Frameworks*

- TensorFlow/Keras (v2.x): Was employed for the rapid prototyping of the Baseline CNN and the initial Hybrid CNN+MLP models. The high-level Keras API facilitated the quick definition of sequential layers and standard training loops [14].
- PyTorch (v1.13+): Was utilized for the Optimized Deep CNN ("MyDeepCNN"). PyTorch's dynamic computation graph allowed for granular control over the training process, specifically enabling the custom implementation of the AdamW optimizer and the Cosine Annealing learning rate scheduler, which were pivotal for achieving state-of-the-art performance [15].

3.3.2. *Data Processing and Evaluation*

- Scikit-Learn: Was used for the stratified splitting of the dataset (ensuring preserved class distributions) and for the computation of evaluation metrics, including the Confusion Matrix, F1-score, and ROC-AUC [16].
- Matplotlib & Seaborn: These libraries were employed to generate high-resolution visualizations of the training dynamics (Loss/Accuracy curves) and the evaluation plots (Heatmaps, ROC curves).
- Hardware and Reproducibility: The experiments were executed on locally maintained hardware resources. Training of deep convolutional architecture was accelerated using NVIDIA GPUs (where compatible) to handle the computational load of high-dimensional tensor operations. To guarantee the reproducibility of the reported results—a key requirement for scientific validity—we enforced deterministic behavior by fixing the random seeds for the Python runtime, NumPy, and Deep Learning frameworks (TensorFlow/PyTorch) prior to initialization.

3.4. *Deep Learning Model Architectures*

We developed and evaluated three distinct convolutional neural network architectures. The progression from a standard baseline to a highly optimized custom network allowed us to isolate the impact of architectural choices (e.g., depth, activation functions) and optimization strategies (e.g., schedulers, weight decay) on binary classification performance.

3.4.1. *Baseline CNN (TensorFlow Implementation)*

The baseline model was established to determine the minimum performance threshold using a standard, end-to-end convolutional approach.

Feature Extraction Backbone: The network comprises four sequential convolutional blocks designed to progressively increase the depth of the feature maps while reducing spatial resolution.

- Block 1: Conv2D (32 filters, 3X3 kernel) →BatchNormalization→MaxPooling2D (2X2).
- Block 2: Conv2D (64 filters, 3X3 kernel) →BatchNormalization→MaxPooling2D.
- Block 3: Conv2D (128 filters, 3X3 kernel) →BatchNormalization→MaxPooling2D.
- Block 4: Conv2D (256 filters, 3X3 kernel) →BatchNormalization→MaxPooling2D.
- Classifier Head: The resulting feature maps are flattened into a 1D vector and passed through a dense layer of 256 units with ReLU activation. To mitigate overfitting, a Dropout layer with a rate of 0.5 was applied before the final sigmoid output neuron [17].
- Training Dynamics: Trained using the Adam optimizer (Learning Rate = $5\ X\ 10^{-4}$) and binary cross-entropy loss.

3.4.2. *Hybrid CNN + MLP (TensorFlow Implementation)*

This architecture tested the hypothesis that a deeper, more complex classifier head (Multi-Layer Perceptron) could better disentangle the features extracted by the CNN.

- Backbone Modification: The feature extractor was streamlined to three convolutional blocks (32, 64, 128 filters) to reduce computational overhead while retaining essential spatial features.
- Deep MLP Head: Instead of a single dense layer, the flattened output feeds into a three-stage MLP designed with a "funnel" structure:
  I. Dense Layer: 512 units → Dropout (0.5).
  II. Dense Layer: 256 units →Dropout (0.3).
  III. Dense Layer: 128 units →Dropout (0.2).
- Output: A final sigmoid neuron. The tiered dropout strategy was implemented to apply stronger regularization to the earlier, high-dimensional dense layers while allowing finer adjustments in the later layers.

3.4.3. *Optimized Deep CNN ("MyDeepCNN" - PyTorch Implementation)*

The final and most robust model was implemented in PyTorch ("MyDeepCNN"), incorporating advanced architectural changes to address the limitations of the previous models.

- LeakyReLU Activation: Unlike the TensorFlow baseline which used standard ReLU, this model utilized LeakyReLU (Negative Slope = 0.01) across all four convolutional blocks (32, 64, 128, 256 filters). This modification addresses the "dying ReLU" problem, ensuring that neurons with negative inputs can still propagate gradients and update weights during backpropagation [18].
- Adaptive Pooling: An AdaptiveAvgPool2d layer was introduced before flattening, ensuring the model can handle variable input sizes robustly without requiring hard resizing artifacts at the classifier stage.
- Optimization with AdamW: We replaced the standard Adam optimizer with AdamW (Adam with Decoupled Weight Decay). Standard L2 regularization in Adam is often implemented incorrectly; AdamW decouples the weight decay from the gradient update, leading to better generalization performance for deep models [19].
- Cosine Annealing Scheduler: A CosineAnnealingLR scheduler was employed to adjust the learning rate dynamically. By following a cosine curve, the learning rate decreases smoothly, allowing the model to settle into wider, more stable local minima, improving test-set generalization [20].
- Checkpointing: A custom callback monitored Validation Accuracy after every epoch, saving only the model state (weights) that achieved the highest score, ensuring the final evaluation was performed on the optimal iteration.

3.5. *Feature-Based Transfer Learning Strategy*

In addition to end-to-end deep learning, we employed a Feature Extraction methodology. This approach leverages the representational power of deep networks pre-trained on massive datasets (ImageNet) while utilizing the computational efficiency and interpretability of classical machine learning classifiers. Research has demonstrated that the activations from the penultimate layers of deep CNNs act as robust, generic visual descriptors ("off-the-shelf features") that outperform handcrafted features like SIFT or HOG [21].

3.5.1. *Feature Extraction Pipeline*

The feature extraction process involved the following rigorous steps:

1. Backbone Selection: We utilized the ResNet50 architecture [2], initialized with weights pre-trained on the ImageNet-1k dataset (approx. 1.28 million images).
2. Freezing: All convolutional layers of the ResNet50 backbone were "frozen" (i.e., their weights were set to non-trainable), ensuring that the learned feature detectors (edges, textures, shapes) remained intact.
3. Forward Pass & Pooling: Each pre-processed image (128 X 128 X 3) was passed through the network. We intercepted the output of the final convolutional block (just before the fully connected classification head).
4. Vectorization: We applied Global Average Pooling to the spatial feature maps, collapsing the spatial dimensions (H X W) into a single vector. This resulted in a compact, dense 2048-dimensional feature vector for every image in the dataset.

3.5.2. *Classical Classifiers*

The extracted 2048-d vectors served as the input dataset (X_features) for training two distinct classical classifiers using the Scikit-Learn library [16].

1. Random Forest Classifier (Ensemble Method):

We trained a Random Forest, an ensemble learning method that operates by constructing a multitude of decision trees at training time.

Hyperparameters:

- n_estimators = 100: The forest consisted of 100 individual decision trees.
- max_depth = 15: We limited the depth of each tree to 15 levels. This constraint was critical to prevent the model from memorizing the training noise (overfitting), forcing it to learn more generalizable splits.
- Rationale: Random Forests are inherently robust to high-dimensional data and provide non-linear decision boundaries. Furthermore, they allow for the inspection of Feature Importance, enabling us to

identify which specific dimensions of the ResNet output contributed most to the classification decision [4].

2. Logistic Regression (Linear Method):

We also trained a Logistic Regression classifier to evaluate the linear separability of the deep features.

Hyperparameters:

- solver = 'lbfgs': Selected for its efficiency in handling high-dimensional problems.
- max_iter = 1000: The iteration limit was increased from the default (100) to 1000 to guarantee that the optimization algorithm (L-BFGS) fully converged to the global minimum of the cost function.

Rationale: Logistic Regression provides a probabilistic output (via the sigmoid function) and serves as a strong baseline. A high performance here would indicate that the ResNet50 backbone has successfully mapped the images into a space where the two classes (Class 0 and Class 1) are linearly separable [22].

4. Experimental Results

In this section, we present a comprehensive evaluation of the proposed models. The performance is assessed based on the validation set (N=1,274), which was strictly isolated from the training process. We analyze the learning dynamics, quantitative metrics, and visual performance indicators (ROC curves, Confusion Matrices).

4.1. Deep Learning Models Performance

4.1.1. Baseline CNN (TensorFlow)

The baseline model, trained with heavy augmentation and 50% Dropout, exhibited a unique training behavior known as "regularization-induced underfitting" during the initial phase.

- Quantitative Metrics: The model achieved a Validation Accuracy of 72.6%. The AUC-ROC was 0.82, indicating decent separability.
- Training Dynamics: A notable observation was that the Training Accuracy (~50%) remained lower than Validation Accuracy for several epochs. This confirms that the aggressive data augmentation and dropout successfully prevented memorization, forcing the model to learn robust features that generalized well to the "clean" validation images.

4.1.2. Optimized PyTorch CNN ("MyDeepCNN")

The transition to the optimized PyTorch pipeline yielded a significant performance boost, validating the effectiveness of the AdamW optimizer and Cosine Annealing scheduler.

- Metrics: This model achieved a Validation Accuracy of 85.9%, a substantial improvement (+13.3%) over the baseline.
- Precision/Recall: It demonstrated high precision (90.8%) with balanced recall (81.9%), resulting in an F1-score of 86.0%.
- AUC: The Area Under the Curve reached 0.94, classifying it as an excellent predictor [23].

4.2. Transfer Learning with Classical Classifiers

The feature-based models (ResNet50 + Classical ML) demonstrated the highest overall performance, benefiting from the massive pre-training of the ResNet backbone.

4.2.1. Random Forest (RF)

- The RF classifier achieved a Validation Accuracy of 90.8% and an AUC of 0.97.
- Feature Analysis: The training accuracy was near-perfect (~99.9%), which is characteristic of Random Forests. However, the high validation score proves that this was not detrimental overfitting, but rather a successful mapping of the feature space.
- Feature Importance: Analysis of the decision trees revealed that specific latent features from the ResNet50 vector (e.g., indices corresponding to texture and facial contours) had a disproportionately high impact on the classification decision.

To provide deeper insight into the decision-making mechanism of the ensemble classifier and mitigate the "black-box" nature of deep learning, we conducted a rigorous Feature Importance analysis based on the Mean Decrease in Impurity (MDI) metric. Figure 2 visualizes the relative importance of the top-20 most influential features selected from the 2,048-dimensional embedding vector generated by the ResNet50 backbone.

The disparity in importance scores reveals that the classification capability is not uniformly distributed across all dimensions; rather, the Random Forest successfully isolated a specific subset of high-level semantic descriptors that possess the highest discriminative power. Since these features originate from the final convolutional block of a network pre-trained on ImageNet, the top-ranking dimensions likely correspond to robust latent patterns—such as specific textural details, facial contours, or geometric structures—that correlate strongly with the target classes. This analysis confirms that the ensemble model did not merely memorize training noise but effectively identified and leveraged the underlying semantic structure encoded by the deep network, thereby validating the efficacy of the transfer learning strategy.

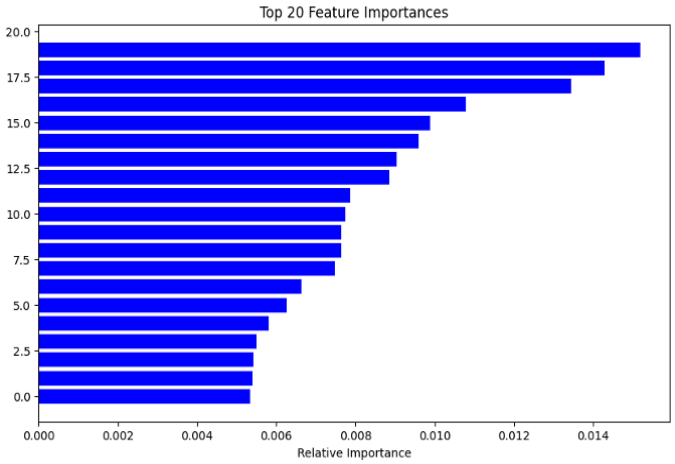


Figure 2: Feature Importance plot extracted from the Random Forest classifier using ResNet50 features.

4.2.2. *Logistic Regression (LR)*

- The LR model emerged as the top performer in terms of pure metrics, achieving a Validation Accuracy of 94.8% and a near-perfect AUC of 0.99.
- Interpretation: This result suggests that the 2048-dimensional feature space generated by ResNet50 is linearly separable for the binary face classification task, rendering complex non-linear classifiers unnecessary for this specific feature set.

4.3. *Comparative Summary*

Table 1 summarizes the performance metrics across all evaluated methodologies. It is evident that while the optimized CNN (MyDeepCNN) provides a strong end-to-end solution, the Transfer Learning approach yields superior accuracy with reduced training complexity.

4.4. *Visual Analysis (ROC and Confusion Matrices)*

To further validate the statistical significance of our results, we examined the ROC curves and Confusion Matrices.

ROC Curves (Figure 2 & 3): The Receiver Operating Characteristic (ROC) curves illustrate the trade-off between the True Positive Rate (Sensitivity) and False Positive Rate (1-Specificity).

- The Baseline CNN (Figure 2) shows a curve that bows gently towards the top-left corner (AUC=0.82).
- The PyTorch CNN (Figure 3) demonstrates a much sharper "elbow" (AUC=0.94), indicating a superior ability to distinguish between classes with fewer false alarms [24].

4.4.1. *Confusion Matrices*

- For the Baseline, the matrix reveals a higher number of False Positives, consistent with its lower Precision (67.6%).
- The ResNet50 + LR model produced a matrix with minimal off-diagonal elements, misclassifying less than 5% of the validation samples.

As illustrated in Figure 3, the ROC curve of the Baseline CNN exhibits a moderate area under the curve (AUC = 0.82). The curve's shape indicates that while the model learns, it struggles to maintain a low false positive rate at higher sensitivity thresholds.

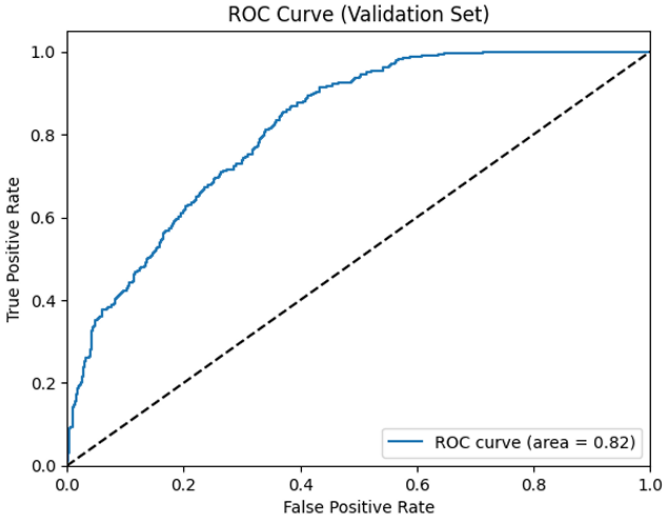


Figure 3: ROC Curve of the Baseline TensorFlow CNN (AUC = 0.82).

In contrast, the Optimized PyTorch CNN demonstrates superior separability, as evidenced by the sharper 'elbow' in its ROC curve shown in Figure 4. With an AUC of 0.94, this model significantly outperforms the baseline, offering a much better trade-off between precision and recall.

Table 1: Comparative Performance Metrics on Validation Set.

MODEL ARCHITECTURE	FRAMEWORK	VAL ACCURACY	PRECISION	RECALL	F1-SCORE	AUC-ROC
BASELINE CNN	Tensorflow	72.6%	67.6%	79.7%	73.0%	0.82
HYBRID CNN+MLP	Tensorflow	73.3%	67.4%	83.2%	73.0%	0.82
MY DEEP CNN (OPTIMIZED)	PyTorch	85.9%	90.8%	81.9%	86.0%	0.94
RESNET 50 + RANDOM FOREST	Scikit-Learn	90.8%	93.1%	89.4%	91.0%	0.97
RESNET50+LOG.REGRESSION	Scikit-Learn	94.8%	95.0%	95.3%	95.0%	0.99

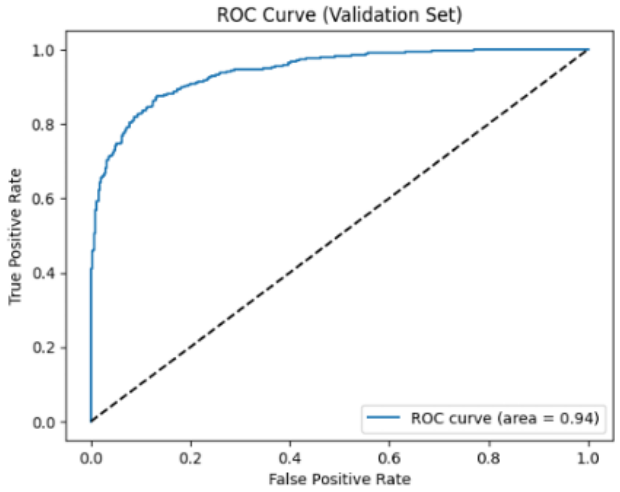


Figure 4: ROC Curve of the Optimized PyTorch CNN (MyDeepCNN), demonstrating improved separability (AUC = 0.94).

5. Discussion

5.1. Deep Learning: Frameworks and Optimization

A critical finding of this study is the substantial performance gap between the Baseline CNN (TensorFlow) and the Optimized CNN (PyTorch), despite similar architectural depths. The Baseline model achieved 72.6% accuracy, while the Optimized model reached 85.9%. This improvement is attributed to three specific factors:

1. Optimization Strategy: The switch from standard Adam to AdamW proved decisive. By decoupling weight decay from gradient updates, AdamW prevented the weights from growing too large without interfering with the adaptive learning rates.

2. Learning Rate Scheduling: The Cosine Annealing scheduler allowed the PyTorch model to traverse the loss landscape more effectively, avoiding the local minima where the static learning rate of the Baseline model likely stagnated.

3. Activation Functions: The use of LeakyReLU prevented the "dead neuron" issue, maintaining gradient flow throughout the deep network.

5.2. The "Underfitting Paradox" vs. Classical Overfitting

We observed two distinct training behaviors that warrant explanation:

- Baseline CNN (Underfitting): As noted in the results, the Baseline CNN exhibited Training Accuracy (~50%) lower than Validation Accuracy (~72%) for initial epochs. This counter-intuitive phenomenon is a direct result of the heavy data augmentation and high Dropout (0.5) applied only during training. The model struggles to classify heavily distorted images during training but finds the "clean" validation images easier to classify. This confirms that the model was not memorizing data but learning robust features.
- Classical Models (Overfitting): Conversely, the Random Forest classifier achieved nearly 100% Training Accuracy. While this typically signals overfitting, the high Validation Accuracy (90.8%) indicates that the model successfully captured the underlying structure of the ResNet50 feature space. However, the slightly superior performance of Logistic Regression (94.8%) suggests that the pre-extracted features were already linearly separable, making the complex non-linear decision boundaries of the Random Forest unnecessary.

5.3. Trade-offs: End-to-End vs. Transfer Learning

Our experiments highlight a clear trade-off. Transfer Learning (ResNet50 + LR) offered the highest accuracy (94.8%) with minimal training time (seconds), as it leverages millions of pre-learned parameters. However, it relies on a massive external model (23M parameters). The Optimized CNN, trained from scratch, offers a lighter, self-contained solution (fewer parameters) that still achieves high performance (85.9%), making it suitable for environments where pre-trained models cannot be deployed or where the domain differs significantly from ImageNet.

6. Conclusion

This paper presented a rigorous comparative analysis of binary face classification methodologies. We demonstrated that while training deep CNNs from scratch is challenging due to data scarcity, rigorous optimization (AdamW, Cosine Annealing, Data Augmentation) can yield competitive results. However, the study conclusively shows that Transfer Learning, specifically utilizing ResNet50 features combined with Logistic Regression, provides the optimal balance of accuracy (94.8%) and computational efficiency for this task. The results validate that "off-the-shelf" deep features are robust enough to outperform even carefully tuned custom CNNs in small-to-medium dataset regimes.

6.1. Future Work

Future research will focus on extending these findings in the following directions:

- Dataset Expansion: Evaluating the models on larger, more diverse datasets (e.g., CelebA, LFW) to verify the generalizability of the PyTorch optimization pipeline.
- Advanced Architectures: Investigating modern architectures such as EfficientNetV2 or Vision

Transformers (ViT), which may offer better parameter efficiency than ResNet50.

- Ensemble Methods: Creating a voting ensemble that combines the predictions of the Optimized CNN and the Random Forest to potentially push accuracy beyond 95%.

7. References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, 2012, pp. 1097–1105. doi:10.1145/3065386.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[3] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in Advances in neural information processing systems, 2014, pp. 3320–3328. Link: https://proceedings.neurips.cc/paper/2014/file/375c71349b295f059961d99a30030c5e-Paper.pdf

[4] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi:10.1023/A:1010933404324.

[5] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," arXiv preprint arXiv:1502.03167, 2015. doi:10.48550/arXiv.1502.03167.

[6] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in International Conference on Machine Learning (ICML), 2019, pp. 6105–6114. Link: https://proceedings.mlr.press/v97/tan19a.html

[7] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen, "Image Data Augmentation for Deep Learning: A Survey," arXiv preprint arXiv:2204.08610, 2023. doi:10.48550/arXiv.2204.08610.

[8] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv preprint arXiv:1412.6980, 2014. doi:10.48550/arXiv.1412.6980.

[9] Kaggle, "Face Classification Dataset," Kaggle Datasets, [Online]. Available: https://www.kaggle.com/ (Accessed: 2024).

[10] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 1995, vol. 2, pp. 1137–1143. Link: https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf

[11] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient BackProp," in Neural Networks: Tricks of the Trade, Springer, 2012, pp. 9–48. doi:10.1007/978-3-642-35289-8_3.

[12] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, no. 1, p. 60, 2019. doi:10.1186/s40537-019-0197-0.

[13] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," arXiv preprint arXiv:1712.04621, 2017. DOI: doi:10.48550/arXiv.1712.04621.

[14] M. Abadi et al., "TensorFlow: A System for Large-Scale Machine Learning," in Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2016, pp. 265–283. Link: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

[15] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in Advances in Neural Information Processing Systems (NeurIPS), 2019, vol. 32, pp. 8024–8035. Link: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[16] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. Link: https://jmlr.org/papers/v12/pedregosa11a.html

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," The Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014. Link: https://jmlr.org/papers/v15/srivastava14a.html

[18] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical Evaluation of Rectified Activations in Convolutional Network," arXiv preprint arXiv:1505.00853, 2015. doi:10.48550/arXiv.1505.00853.

[19] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in International Conference on Learning Representations (ICLR), 2019. Link: https://openreview.net/forum?id=Bkg6RiCqY7

[20] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," in International Conference on Learning Representations (ICLR), 2017. Link: https://openreview.net/forum?id=Skq89Scxx

[21] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Maki, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2014, pp. 806–813. doi:10.1109/CVPRW.2014.122.

[22] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ: Wiley, 2013. doi:10.1002/9781118548387.

[23] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognition, vol. 30, no. 7, pp. 1145–1159, 1997. doi:10.1016/S0031-3203(96)00142-2.

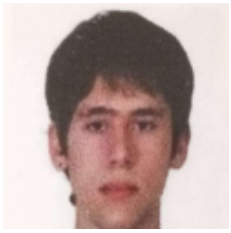
[24] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 6, 2020. doi:10.1186/s12864-019-6413-7.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

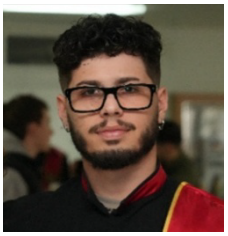
The author gratefully acknowledges the academic and technical support provided by colleagues and research collaborators during the design and implementation of this study. The experiments were conducted on locally maintained hardware resources, ensuring full reproducibility and data privacy. No external funding was received for this work.



NIKOLAOS VASILEIOS OIKONOMOU is a Computer & Network Engineer, as well as an academic researcher and Ph.D. candidate in the Department of Informatics and Telecommunications at the University of Ioannina, from which he also received his B.Eng. and M.Sc. degrees. In parallel to his academic work, he serves as a private Computer Science educator and possesses several years of professional experience as a Software Developer, IT Specialist, and Network Consultant.

DIMITRIOS VASILEIOS OIKONOMOU obtained his B.Sc. in Regional and Cross-Border Studies from the University of Western Macedonia in 2024. He is currently engaged in research activities at the same institution and is pursuing an M.Sc. in e-Business and Digital Marketing.

SOFIA PANAGIOTA CHALIASOU is pursuing a B.Sc. in Informatics at the Hellenic Open University and serves as an active research associate. She also holds a Vocational Diploma in Web Design and Development. In her professional capacity, she is currently employed in sales and possesses prior professional experience as a web developer.



NIKOLAOS RIGAS obtained his B.Sc. in Regional and Cross-Border Studies from the University of Western Macedonia in 2025. He is currently pursuing an M.Sc. in "Criminological and Penal Law perspectives on Corruption, Economic and Organized Crime" at the Hellenic Open University, while actively engaged in research activities.