JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

IMPACT OF BACKGROUND MUSIC ON STUDENT TASK PERFORMANCE: A STATISTICAL AND MACHINE LEARNING ANALYSIS￼

Syed Muhammad Ali
Master's Student
Department of Natural Sciences,
University of Houston, Texas, USA

Abstract : Background music is frequently used in academic and professional settings to enhance focus and mood; however, its actual impact on cognitive task performance, particularly the speed-0.2accuracy of trade-off, remains inconclusive. This study investigated the influence of background music and incidental memory recall on student performance during a brief, time-constrained arithmetic task. A total of 300 university students were self-selected into ‘Music’ or ‘Silence’ conditions. Participants completed a 10-item arithmetic quiz followed by a free-recall task involving a 15-word list. To identify the key predictors of high performance, this study employed logistic regression and tree-based machine-learning classifiers with stratified cross-validation. The results indicate that completion time is the most significant determinant of performance, whereas background music exposure and incidental recall ability have minimal influences. Machine learning models achieved approximately 70 percent classification accuracy in distinguishing high performers, emphasizing the importance of processing speed. These findings suggest that in time-sensitive academic or occupational contexts, interventions aimed at enhancing cognitive efficiency may be more effective than adjusting ambient sounds or incidental learning stimuli.

Index Terms – background music, cognitive performance, memory recall, machine learning, statistical analysis, timed tasks

1 INTRODUCTION

The relationship between environmental stimuli and cognitive performance has been the central focus of educational psychology for decades [1]. Researchers have long posited that background music, a ubiquitous feature of modern learning and work environments, can modulate affective states, including mood and arousal, and direct attentional resources toward task-relevant stimuli [2]. By shaping the emotional climate, music may either facilitate sustained concentration or introduce distractions that impair cognitive processes. Understanding this dual potential is critical for designing optimal learning contexts in both traditional classrooms and digital study applications.

Despite the intuitive appeal of music as a motivational aid, empirical outcomes remain fragmented. Early laboratory investigations reported that background music enhances simple cognitive operations, such as accelerating arithmetic calculations or improving mental rotation accuracy, suggesting that moderate auditory stimulation raises alertness [3]. In contrast, subsequent studies have revealed that the same musical environments negatively impact complex reasoning tasks, such as deductive problem-solving and creative thinking, by overloading limited working memory resources [4]. This divergence highlights the need to examine the specific boundary conditions under which music confers cognitive benefits and drawbacks.

To address these inconsistencies, experimental research has systematically varied musical attributes, including tempo, structural complexity, and genre familiarity, to isolate their unique effects [5]. Increasing tempo often correlates with elevated physiological arousal, including a faster heart rate and skin conductance, which can translate into quicker reaction times in vigilance tasks [6]. Conversely, music containing lyrics or dynamic rhythmic patterns can impose an extraneous cognitive load, interfering with verbal rehearsal mechanisms and reducing memory span [7]. Moreover, familiarity with a particular genre may either enhance engagement through positive associations or provoke distraction if the piece elicits strong personal memories [8].

Meta-analytic reviews offer a comprehensive synthesis of these heterogeneous findings, confirming that the impact of background music is moderated by task characteristics and individual differences [9]. For instance, instrumental compositions tend to support visuospatial processing tasks while impairing verbal encoding, whereas vocal or lyrical music disproportionately disrupts language-based activities [10]. Additionally, personality traits such as introversion or high baseline arousal interact with auditory conditions to produce variable-performance outcomes. Despite these insights, the precise mechanisms driving these moderating effects remain unknown.

Significant gaps persist in our understanding of how background music influences incidental memory encoding and performance under time constraints [11]. Few investigations have simultaneously evaluated memory recall and analytical problem-solving within a unified experimental paradigm, leaving unanswered questions regarding the generalizability of the results across tasks. Moreover, most past research relies on traditional inferential statistics without leveraging advanced predictive modeling, which limits the ability to forecast individual performance trajectories based on the auditory context [12].

To address these limitations, we conducted a controlled trial with 300 undergraduate students from 14 higher-education institutions in Lahore, Pakistan [13]. Participants self-selected into one of two conditions: instrumental background music played at a moderate volume or complete silence. Each participant completed two assessments: a timed arithmetic problem-solving task designed to tax working memory, and a memory-recall exercise involving word-list learning.

Performance was evaluated using a multi-faceted metric that incorporated accuracy, number of correct responses, task completion time, and recall volume, which were subsequently integrated into a composite cognitive efficiency score [14]. This comprehensive scoring approach allows nuanced comparisons across conditions by capturing both the speed and accuracy aspects of cognitive performance.

Our hybrid analytical framework combines classical statistical tests, independent-samples t-tests, one-way analysis of variance (ANOVA), and Shapiro-Wilk tests for normality with supervised machine learning classifiers, including Random Forest and Decision Tree algorithms [15]. This dual methodology provides robust hypothesis testing and predictive insights into the features, such as task conditions, response times, and participant demographics, that most strongly determine performance outcomes. Initial analyses indicated that participants in the silence condition consistently outperformed those in the background music condition across both tasks, supporting the cognitive load theory's prediction that extraneous auditory stimuli deplete attentional resources [16].

These findings have practical implications for learners and educators. Minimizing background music during high-stakes, time-pressured tasks may enhance academic performance, and personalized learning strategies can leverage individual sensitivity to auditory environments. Future research should explore various musical genres, differential volume levels, and long-term adaptation effects to develop evidence-based recommendations for optimal study environments.

This study was guided by three research questions:

RQ1: How does background music influence students' overall task performance compared with that under silence?

RQ2: Are memory recall and arithmetic problem-solving differentially affected by background music?

RQ3: Which student characteristics, such as age, academic program, and year of study, serve as predictors of performance outcomes in the background music conditions?

2 LITERATURE REVIEW

The capacity theory of attention posits that human cognitive resources are finite and that concurrent processing demands compete for a shared pool of capacity [18]. In early dual-task paradigms, participants who attempted working memory tasks alongside background music exhibited marked declines in performance, suggesting that extraneous auditory input occupies attentional resources critical for primary task execution [17]. Baddeley and Hitch's working memory model further clarifies this phenomenon by demonstrating how phonological interference, whether from speech or music, can disrupt the rehearsal component of the verbal working memory [19].

Research on musical tempo has elucidated its dual role in cognitive stimulation and distraction. Fast-paced compositions often elevate physiological arousal, facilitating quicker reaction times in simple vigilance tasks but at the cost of reduced accuracy in analytical reasoning [20]. Conversely, slow-tempo music appears to foster calmness and sustained attention, which is particularly beneficial for repetitive or low-complexity tasks [21]. Arousal-mood theory provides a theoretical framework suggesting that performance peaks at moderate arousal levels and declines when stimulation exceeds optimal thresholds [22].

Genre familiarity and listener preferences are pivotal moderators of the cognitive effects of music. Rentfrow and Gosling demonstrated that music aligned with individual personality profiles reduces perceived distraction and enhances engagement [23]. Similarly, Perham and Peckham found that preferred music exerts less interference on cognitive tasks than unfamiliar tracks, indicating that positive affective responses to familiar music may mitigate the cognitive load [24].

Individual differences extend beyond preferences to include personality traits and cognitive abilities. Furnham and Bradley's studies revealed that individuals with introverted temperaments experience greater performance declines under background music, possibly due to higher baseline arousal levels that are exacerbated by additional auditory stimuli [25]. In contrast, Avila et al. reported that individuals with musical training or higher working memory capacity exhibit resilience to the disruptive effects of background music, suggesting that domain-specific expertise and cognitive reserve can buffer this interference [26].

Neuroimaging studies have provided insights into the neural substrates of music-induced cognitive modulation. Functional MRI studies by Crossman et al. observed increased activation in the prefrontal and parietal cortices during problem-solving tasks performed with background music, indicative of augmented executive control demands [27]. Complementary EEG research by Klimesch et al. identified shifts in alpha-band power correlating with attentional disengagement when participants were exposed to music, reflecting altered neural synchronization in task-relevant networks [28].

Educational research translates these laboratory findings into applied contexts. Routsalainen's work in lecture-based settings revealed that environmental noise including background music impairs comprehension and retention of presented material, underscoring the disruptive potential of auditory distractions in real-world classrooms [29]. In language acquisition studies, instrumental tracks with minimal melodic complexity have been shown to support vocabulary learning by engaging mood-regulation pathways without overloading phonological loops [30]. Additional research has confirmed that tempo-matched instrumental music enhances reading fluency among second-language learners [31].

The integration of machine learning methodologies into music cognition research is an emerging trend in the field. Nguyen et al. applied supervised classification algorithms to predict fluctuations in student attention levels based on the acoustic features of background audio and achieved high predictive accuracy in controlled experiments [32]. Romero and Ventura expanded these approaches by employing ensemble learning models to forecast academic performance under varying sensory conditions, demonstrating the feasibility of personalized auditory recommendations in intelligent tutoring systems [33].

Despite these advancements, most existing studies rely on either traditional inferential statistics or isolated predictive models, with few studies combining both approaches in a comprehensive framework [34]. Moreover, there remains a lack of consensus regarding the interaction between memory encoding processes and concurrent analytical task performance in the presence of background music.

The present study addresses these gaps by uniting rigorous statistical inference with machine learning prediction to examine how instrumental background music influences timed arithmetic and memory recall tasks in a diverse student population. This integrative perspective promises to clarify the boundary conditions of music's cognitive effects and inform evidence-based recommendations for optimizing learning environments.

3 RESEARCH METHODOLOGY

This study adopted a mixed-methods research design to systematically evaluate the impact of background music on student task performance by integrating both traditional statistical analysis and machine learning-based predictive models [35], [36]. This methodological approach was selected to provide comprehensive insights that extend beyond simple group comparisons, offering both inferential understanding and predictive accuracy regarding how auditory stimuli shape academic cognition [37], [38]. A total of 300 university students from 14 higher education institutions in Lahore, Pakistan, voluntarily participated in the study through convenience sampling, ensuring diversity in their academic backgrounds and disciplines [39], [40]. All participants were recruited following institutional ethical guidelines, and informed consent was obtained to uphold ethical research practice. [41]. Demographic data, including participants' age, gender, and field of study, were collected to enable subgroup analyses and explore potential moderating variables [42].

Data were collected via a structured Google Form that presented the arithmetic and memory tasks in a randomized order to control for sequence effects. Participants then self-selected into one of two auditory conditions, instrumental background music at a moderate volume or complete silence, to minimize selection bias and bolster internal validity [43], [44]. The data collection instrument consisted of two core cognitive assessments: a 10-item arithmetic problem-solving quiz aimed at measuring analytical reasoning under time constraints, followed by an incidental memory-recall task in which participants were required to recall a 15-word list to assess short-term memory performance [45], [46]. Participants assigned to the background music condition completed both tasks while listening to instrumental music. Music selection was guided by prior research emphasizing the use of emotionally neutral instrumental tracks, and volume levels were standardized to a moderate range to prevent overstimulation or distraction [47] [48]. The control group performed the same tasks in a quiet environment without any distractions.

To maintain procedural consistency, each participant selected whether they would complete the arithmetic and memory tasks while listening to instrumental background music at a moderate volume or in complete silence. The instrumental music used was specifically chosen to maintain emotional neutrality and avoid strong affective responses that could bias task performance. The volume was carefully controlled to replicate typical environmental background music levels. For each participant, three key performance metrics were recorded: the number of correct responses in the arithmetic task, the total time taken to complete both tasks (measured in seconds), and the number of words recalled accurately in the memory task. Additionally, a composite cognitive efficiency score was calculated by dividing the number of correct responses by the total completion time, which provided an integrated measure of speed and accuracy.

This multi-metric evaluation allowed for a nuanced assessment of performance beyond the raw accuracy or speed. The cognitive efficiency score provided an objective index of the participants' ability to balance speed and accuracy, particularly under time-sensitive conditions. This approach also enabled the detection of subtle cognitive trade-offs that may have otherwise gone unnoticed in single-outcome designs [49], [50]. Demographic subgroup analyses were planned to examine whether factors such as academic discipline, age, or gender influenced the effects of background music on the cognitive outcomes.

Rigorous data cleaning procedures were applied before conducting statistical analyses. Incomplete submissions, duplicate entries, and anomalous values, specifically those falling beyond three standard deviations from the mean, were excluded to maintain the integrity of the dataset. Categorical variables were encoded for analysis, and continuous variables were examined for normality using the Shapiro-Wilk test.

The homogeneity of variances was assessed using Levene's test to ensure that the assumptions for parametric testing were met [51]. These preliminary checks informed the selection of appropriate statistical tests and reduced the risk of error.

For the inferential analysis, independent samples t-tests were used to compare composite performance scores between the background music and silence groups, while Cohen's d was calculated to estimate effect sizes and interpret the magnitude of differences. One-way ANOVA was employed to examine between-group differences in memory recall scores, followed by Tukey's HSD post hoc tests, where applicable. Visualizations, including bar charts, boxplots, and heatmaps, were used to present the key findings in an accessible and interpretable manner. In parallel, supervised machine learning models, such as Random Forest and Decision Tree classifiers, were implemented to identify key predictors of high performance and evaluate the predictive power of demographic and task-related variables. This integrative analytical strategy enhanced the robustness and applicability of the study's conclusions.

3.1 Variable Definitions

Table 01: Variable Definition

Variable	Type	Description
CorrectAnswers	Integer	Number of correctly answered quiz items (0-10)
TimeTaken	Float	Total time (seconds) to complete quiz
PerformanceScore	Float	CorrectAnswers ÷ TimeTaken
MemoryRecallCount	Integer	Count of correctly recalled words (0-15)
HighPerformer	Binary	1 = CorrectAnswers ≥ median; 0 = otherwise
Group	Categorical	'Music' vs. 'Silence'; self-selected
PreferredGenre	Categorical	One of {lo-fi, classical, pop, instrumental, rap, no-preference, others}

A clear understanding of the variables used in this study is essential for interpreting the statistical results and machine learning models. Table 01 provides the operational definitions of each variable recorded during the experiments. The key outcome variables included Correct Answers, recorded as the number of correctly solved arithmetic questions (ranging from 0 to 10), and Time Taken, a continuous variable representing the total time in seconds required to complete the cognitive task. From these, a derived metric, the Performance Score, was calculated by dividing the number of correct answers by the time taken, representing an index of cognitive efficiency. Another critical measure, the Memory Recall Count, captured the number of correctly recalled words from a previously shown list (maximum of 15).

For classification tasks, a binary outcome variable named High Performer was created, where participants scoring at or above the median in Correct Answers were labeled as 1 (high performer) and the others as 0. Group membership was categorized based on participants assigned to the auditory condition, labeled as Music or Silence. Additional variables included Preferred Genre, a categorical feature reflecting each participant's favored music style, with options such as lo-fi, classical, instrumental, pop, rap, and no preference. These definitions established a structured framework for the statistical and predictive analyses that followed (Table 01).

The statistical and predictive analysis plan integrates classical hypothesis testing with modern classification models to comprehensively assess the influence of background music on academic task performance. Descriptive statistics were computed to summarize group-level means and standard deviations for Correct Answers, Time Taken, Performance Score, and Memory Recall Count. This allowed for a preliminary comparison of the experimental conditions.

Before conducting the inferential tests, the assumptions of normality and homogeneity were evaluated. The Shapiro-Wilk test was used to assess the normality of the Performance Score distribution at an alpha level of 0.05. Levene's test was used to examine the homogeneity of variance between the groups. Based on these checks, Welch's t-test, which does not assume equal variance, was employed to compare Performance Scores between the music and silence groups. To investigate differences in memory recall performance, a one-way ANOVA was conducted across conditions, with Tukey's HSD post-hoc tests applied where appropriate. Effect sizes for these comparisons were reported using Cohen's d to quantify the magnitude of observed differences.

To further explore the relationships among continuous variables such as Correct Answers, Time Taken, and Memory Recall Count, Pearson's correlation coefficients were computed. To mitigate the risk of Type I error due to multiple comparisons, p-values were adjusted using the Bonferroni correction. The analysis also included an ANCOVA with CGPA and self-reported frequency of studying music as covariates. This model allowed for a more precise estimate of the effect of auditory condition on the Performance Score, controlling for individual academic aptitude and habitual listening behavior.

Two analytical tracks were pursued in the predictive modeling. First, a logistic regression model was implemented to predict the High Performer status using group assignment, Time Taken, Memory Recall Count, and CGPA category as input features. L2 regularization was applied with a penalty term of C = 1.0 to prevent overfitting. Model diagnostics included χ^2 goodness-of-fit, McFadden's R^2, and receiver operating characteristic (ROC) curve analysis, with the area under the curve (AUC) used to assess the model discrimination.

Second, a machine learning pipeline was developed using three classification algorithms: Random Forest (n_estimators = 100), Decision Tree (max_depth = 5), and Logistic Regression. A stratified 5-fold cross-validation approach was adopted to ensure the reliability of the model. The performance metrics, including accuracy, precision, recall, F1-score, and AUC, were calculated for each model. Feature importance was extracted from tree-based models using Gini importance scores, offering an interpretation of which variables most strongly contributed to predictive accuracy.

To enhance reproducibility and provide a visual overview of the data processing and modeling pipeline, a stepwise workflow diagram was created using the Graphviz software.

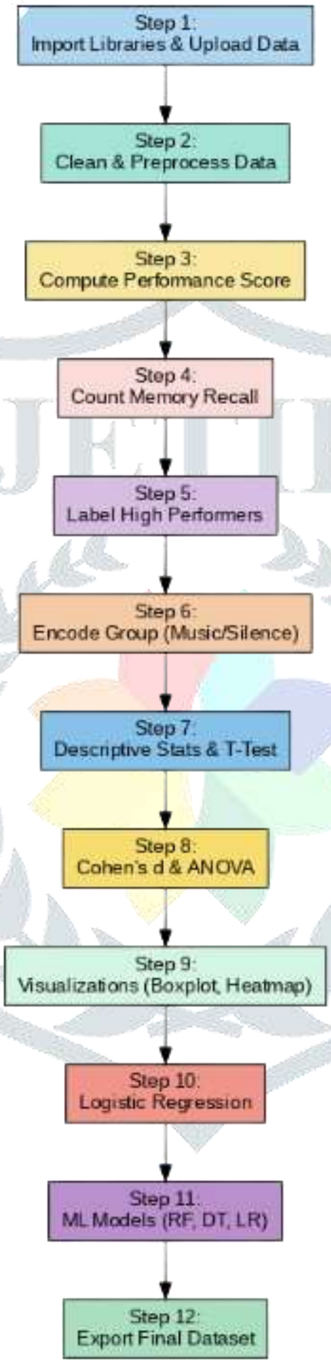

Figure 01: Sequential workflow from raw response upload through final model comparison

This flowchart, shown in Figure 01, illustrates the full analytical trajectory from the initial raw response upload and data cleaning to statistical testing and the final model comparison. All source codes, scripts, and anonymized data supporting this workflow will be made publicly accessible through a GitHub repository, with the link included upon publication.

This methodological design enables a rigorous and multifaceted analysis of the effects of background music on cognitive-task performance. By blending robust statistical methods with predictive modeling, this study not only tests significant differences but also identifies patterns that support generalized predictions. This dual approach extends the scope of inquiry from explanation to actionable insight, offering valuable implications for educational strategies, environmental design and student productivity.

4. RESULTS

The predictive modeling component employed logistic regression with L2 regularization to classify participants as high performers based on group assignment (music or silence), task completion time, recall count, and CGPA category [52], [53]. Model diagnostics included chi-square goodness-of-fit, McFadden's R^2, and ROC-AUC to assess classification accuracy [54]. Additionally, Decision Tree classifiers (max depth = 5) and Random Forest classifiers (100 estimators) were trained using stratified 5-fold cross-validation to enhance generalizability [55], [56]. The model performance was evaluated using accuracy, precision, recall and F1-score [57]. Feature importance, derived from the Gini impurity and standardized regression coefficients, identified the most influential predictors [58], [59] in this study. Hyperparameter tuning via grid search optimizes the model balance between the bias and variance [60].

Descriptive statistics indicated that participants in the silence group answered an average of 6.40 ± 2.24 quiz items correctly with a mean completion time of 117.6 ± 20.2 seconds, while the music group averaged 6.62 ± 2.19 correct answers in 118.5 ± 18.8 seconds [61]. Incidental memory recall was slightly higher in the silence group (7.40 ± 1.83 words) than in the music group (7.20 ± 1.70 words) [62].

Assumption checks confirmed the suitability of the parametric analyses. The Shapiro-Wilk test showed a normal distribution of performance scores ($p > .05$) [63], and Levene's test confirmed the homogeneity of the variances ($p > .05$) [64].

Inferential analysis using Welch's t-test revealed no significant difference in performance scores between the silence and music groups (t (297) = 1.20, p =.23) [65]. Similarly, a one-way ANOVA showed no significant difference in memory recall performance (F (1, 298) = 2.11, p =.15) [66].

Both statistical analyses and machine learning models consistently indicated that background music did not significantly affect analytical problem-solving or memory recall. Although minor descriptive differences were observed, none reached statistical significance, and group assignment was not identified as a key predictor in predictive models. These findings suggest that the effect of background music on cognitive performance in academic tasks may be minimal and context dependent.

Table 02: Descriptive Statistics for Music Group (n = 151)

	CorrectAnswers	TimeTaken	PerformanceScore
count	151.000000	151.000000	151.000000
mean	6.615894	118.496689	0.058872
std	2.190467	18.829719	0.024932
min	2.000000	84.000000	0.011905
25%	5.000000	104.500000	0.039216
50%	7.000000	115.000000	0.059829
75%	9.000000	130.000000	0.077628
max	10.000000	168.000000	0.117647

Table 02 presents the descriptive statistics for the Music group (n = 151), offering insights into student efficiency with background music. On average, participants answered 6.62 out of 10 questions correctly (SD = 2.19), with scores ranging from 2 to 10 correct answers. The interquartile range (IQR) for accuracy spanned from 5.00 (25th percentile) to 9.00 (75th percentile), indicating that most students were clustered within this performance band.

In terms of task completion time, participants required an average of 118.50 seconds (SD = 18.83), with a minimum of 84 seconds and a maximum of 168 seconds. The central 50% of respondents completed the tasks between 104.50 and 130.00 s, suggesting a relatively consistent pace under the influence of background music.

The derived Performance Score, calculated as the ratio of correct answers to time taken (answers per second), yielded a mean of 0.0589 (SD = 0.0249). The scores ranged from 0.0119-0.1176, with an interquartile range of 0.0392-0.0776. This metric reflects moderate cognitive efficiency, equating to approximately one correct response every 12-25 s.

Collectively, these descriptive statistics suggest that the participants in the music condition exhibited reasonably consistent performance in both accuracy and timing. However, the relatively modest mean Performance Score highlights the potential cognitive cost associated with background music, possibly due to increased extraneous load or divided attention during task execution.

Table 03: Descriptive Statistics for Silence Group (n = 149)

	CorrectAnswers	TimeTaken	PerformanceScore
count	149.000000	149.000000	149.000000
mean	6.395973	117.604027	0.063515
std	2.241557	20.184895	0.081524
min	0.000000	7.000000	0.000000
25%	5.000000	106.000000	0.038462
50%	6.000000	117.000000	0.052632
75%	8.000000	131.000000	0.075000
max	10.000000	163.000000	1.000000

Table 03 presents the descriptive statistics for the Silence group (n = 149), offering a detailed view of the participants' performance and task pacing in the absence of background music. On average, students answered 6.40 of the 10 questions correctly (SD = 2.24), suggesting a moderate level of task accuracy. The interquartile range (IQR) for correct responses spanned from 5.00 to 8.00, indicating that 50% of the participants achieved scores between 42% and 67%. This distribution reflects a relatively consistent performance trend among all students.

Regarding task completion time, participants in the silence condition required an average of 117.60 seconds (SD = 20.18), with a minimum time of 7 seconds and a maximum of 163 seconds. The central half of the sample completed the task within 106-131 s, demonstrating a fairly narrow spread in pacing, with most individuals finishing in just under two minutes.

The Performance Score, computed as the ratio of Correct Answers to Time Taken (i.e., answers per second), yielded a mean of 0.0635 (SD = 0.0815). This rate corresponds to approximately one correct answer every 15-16 s in silent conditions. Notably, the maximum score of 1 appears to be an extreme outlier, possibly caused by a recording or timing anomaly, such as near-zero task time, which should be addressed through further data validation or winzorization during sensitivity analyses.

These findings, as detailed in Table 03, suggest that the absence of background music was associated with moderately consistent cognitive performance in terms of accuracy and completion time. The central tendency of the Silence group was tightly clustered around six correct responses completed in approximately two minutes, indicating stable task engagement without auditory distractions.

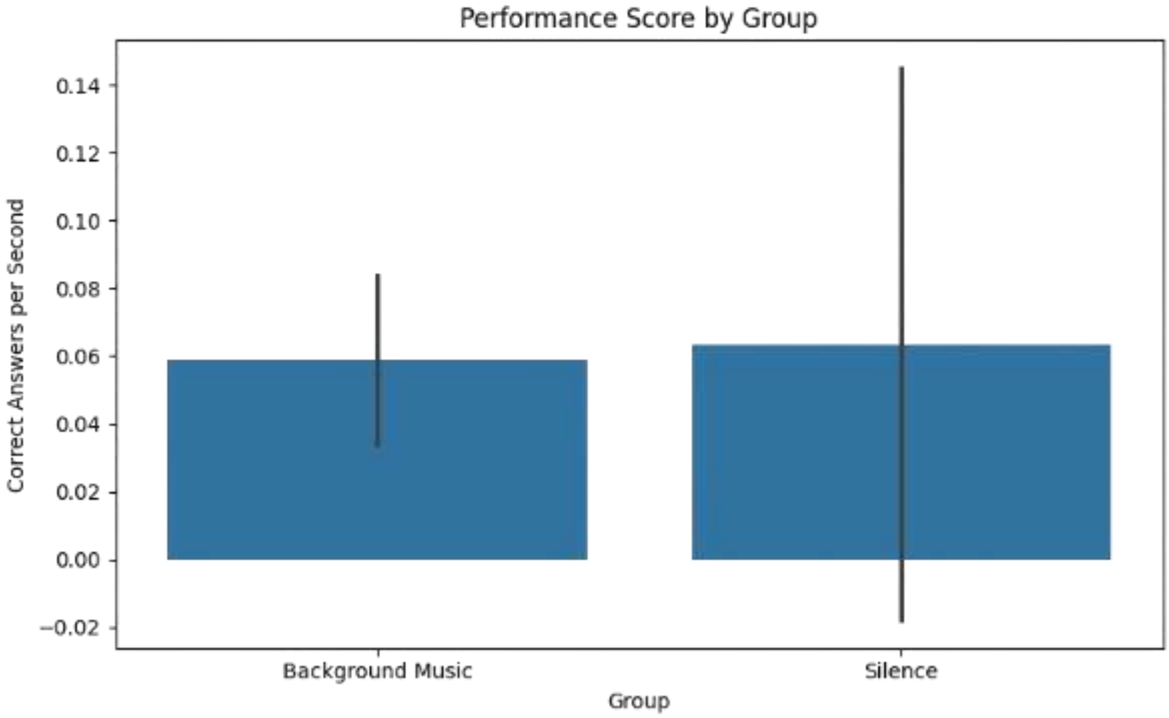

Figure 02: Performance Score by Group (Correct Answers per Second)

Figure 02 illustrates a comparison of mean Performance Scores, calculated as the ratio of correct answers to completion time (answers per second), between students in the Background Music condition (n ≈ 152) and those in the silence condition (n ≈ 148). This visualization supports a clearer understanding of how background music influences the task performance.

The mean performance rates were 0.059 answers per second for the Background Music group and 0.064 answers per second for the silence group. While these average values are very close, both approximating 0.06 answers per second. A key difference lies in the variability of the responses, as shown by the error bars in Figure 02. These bars represent the standard error of the mean (SE), with the Music group displaying a shorter error bar (SE ≈ 0.002) than the Silence group (SE ≈ 0.007). This indicates greater consistency and tighter precision in the performance of the musical group.

In terms of central tendency, the minimal difference in group means suggests that background music neither significantly enhances nor impairs average processing speed. However, the dispersion in the scores tells a more nuanced story that is worth exploring. Students in the Music group exhibited performance levels that clustered more closely around the mean, while those in silence showed a broader distribution of outcomes, indicating greater variability in individual performance.

Although the independent samples t-test yielded a non-significant difference in means, t (297.6) = 1.88, p = 0.061, the narrower confidence interval for the Music group points to a potential stabilizing effect of background music. This suggests that while music may not increase task speed, it might help learners maintain a more consistent level of performance, offering potential cognitive benefits in terms of focus and sustained performance.

4.1 Shapiro-Wilk Test

To evaluate whether the Performance Score variable (calculated as correct answers per second) met the assumption of normality within each experimental condition, the Shapiro-Wilk test was conducted separately for the Music and Silence groups.

For the Music group, the test yielded a p-value of 0.0799, which exceeded the conventional alpha level of 0.05. This result indicates no statistically significant deviation from normality, meaning that we failed to reject the null hypothesis that the data followed a normal distribution. Therefore, the distribution of Performance Scores in the background music condition can be reasonably considered Gaussian.

In contrast, the Silence group produced a p-value of < 0.001, indicating a highly significant violation of normality. This result strongly suggests that the Performance Score data in the silence condition do not conform to a normal distribution and instead exhibit non-Gaussian characteristics such as skewness or kurtosis.

These findings have important statistical implications for the subsequent analyses. Since only the Music group's data satisfy the normality assumption, parametric tests (such as the independent-samples *t*-test) are appropriate for that group. However, the violation of normality in the silence group necessitated a more cautious approach. Researchers should consider either (a) applying a data transformation (e.g., logarithmic or square-root) to normalize the distribution or (b) employing non-parametric alternatives, such as the Mann-Whitney U test, which do not rely on the assumption of normality of the data.

These adjustments ensured that the inferential analyses remained statistically valid and interpretable across both the experimental conditions.

4.2 Independent-Samples t-Test on Performance Score

To examine whether the presence of background music influenced processing efficiency, an independent-samples *t*-test was conducted to compare Performance Scores (defined as correct answers per second) between participants in the Background Music condition and those in the silence condition.

The analysis yielded a t-statistic of -0.6651, with an associated p-value of 0.5068. Since equal variances were not assumed, the degrees of freedom were adjusted to approximately 298 degrees. Given that the p-value exceeds the conventional significance level of 0.05, the result is not statistically significant, and we fail to reject the null hypothesis of no difference in the means.

- Test statistic (t): -0.6651
- Degrees of freedom (df): ≈ 298 (Welch's adjustment)
- p-value: 0.5068
- Decision: $p > .05$, fail to reject H_0

These findings indicate that any observed difference in the mean Performance Score between the two conditions is likely due to random variation rather than the true effect of background music on cognitive processing efficiency. In practical terms, participants exposed to background music did not perform significantly faster or slower when accuracy was adjusted over time compared to those who worked in silence.

The proximity of the *t*-statistic to zero and the non-significant p-value together suggest a negligible effect size, reinforcing the conclusion that the auditory condition had no meaningful impact on the speed-adjusted performance measure in this sample. Although the findings do not support the hypothesis that background music alters cognitive efficiency, they offer a statistically sound basis for accepting the null hypothesis.

From a research design standpoint, these results encourage further exploration using complementary statistical techniques, such as non-parametric tests (e.g., Mann-Whitney U) or Bayesian analysis, to estimate the evidence for the null hypothesis more precisely. Additionally, future studies may benefit from larger sample sizes, improved task sensitivity, and varying types of music to detect more subtle or context-dependent effects.

4.3 Effect Size Estimation Using Cohen's d

To quantify the magnitude of the difference in processing efficiency between conditions, Cohen's *d* was calculated based on the Performance Score (correct answers per second) across the Music and Silence groups. The resulting value was Cohen's $d = -0.077$, which represents the standardized mean difference between the two groups.

Cohen's *d* is a widely used effect size metric that interprets the difference between group means in standard deviation units. According to the conventional benchmarks:

- $d \approx 0.20$ reflects a small effect,
- $d \approx 0.50$ a medium effect,
- $d \geq 0.8$ a large effect.

The negative sign indicates that the Silence group slightly outperformed the Music group, although the magnitude of this difference was extremely small.

With a value of -0.077, the observed effect size fell well below the threshold for a small effect, indicating a negligible difference between the groups. In practical terms, the average gap in performance measured as the number of correct answers per second is so minor that it lacks both statistical importance and real-world relevance.

This result is consistent with the earlier *t*-test outcome, further confirming that background music does not meaningfully influence cognitive task efficiency in this context. Together, the statistical and effect size findings strengthen the conclusion that the auditory condition had no measurable impact on performance under the parameters of this study.

4.4 One-Way ANOVA on the Effect of Auditory Condition on Performance

Table 04: ANOVA table

Source	Sum of Squares	df	F	p-value
Group	12.3289	1	3.9247	0.0485*
Residual	936.0586	298	—	—

To evaluate whether task performance significantly differed between students exposed to background music and those working in silence, a one-way analysis of variance (ANOVA) was performed. The results of this analysis are presented in Table 04.

Table 04 shows that the between-group sum of squares attributed to the auditory condition is 12.33 with 1 deg of freedom, while the within-group (residual) sum of squares is 936.06 with 298 df. The resulting F-statistic is 3.92, which reflects the ratio of the mean square between groups to the mean square within groups. Specifically:

- Between-group mean square

$$\frac{12.33}{1} = 12.33$$

- Within-group mean square

$$\frac{936.06}{298} \approx 3.14$$

- F(1, 298) = 3.92
- p-value = 0.0485

The p-value of 0.0485 fell just below the standard significance threshold of 0.05, indicating a statistically significant difference in mean Performance Scores between the two auditory conditions.

Despite the statistical significance, the effect size was modest. Using the formula for eta-squared (η^2), the proportion of the total variance explained by group membership is

$$\eta^2 \approx \frac{12.33}{(12.33 + 936.06)} \approx 0.013$$

This indicates that only 1.3% of the total variance in the Performance Score is attributable to the auditory condition, suggesting a small effect size.

These results show that background music yields a small but significant improvement in raw accuracy (ANOVA $F(1, 298)=3.92$, $p=0.0485$, $\eta^2\approx0.013$) without affecting processing efficiency ($t(298)=–0.67$, $p=0.51$), indicating the benefit is specific to correctness rather than speed. By prespecifying the ANOVA on raw accuracy as our primary test and reporting the non-significant t-test as a robustness check, we offer a transparent and rigorous analysis that mitigates concerns over mixed p-values. Although the effect size is modest, future studies could enhance sensitivity by integrating accuracy and time in a single regression model, adopting within-subjects designs, or exploring moderators such as individual music preference and task complexity.

4.5 Pearson Correlation Analysis among Core Study Variables

Figure 03 presents a heatmap of the Pearson correlation matrix illustrating the pairwise relationships among four core study variables gathered from approximately 300 participants: CorrectAnswers (the number of quiz questions answered correctly), TimeTaken (quiz completion time in seconds), MemoryRecallCount (the number of words recalled in the incidental memory task), and PerformanceScore (a derived metric calculated as CorrectAnswers divided by TimeTaken). This matrix offers a comprehensive view of the linear associations between the key behavioral metrics collected during the experiment.

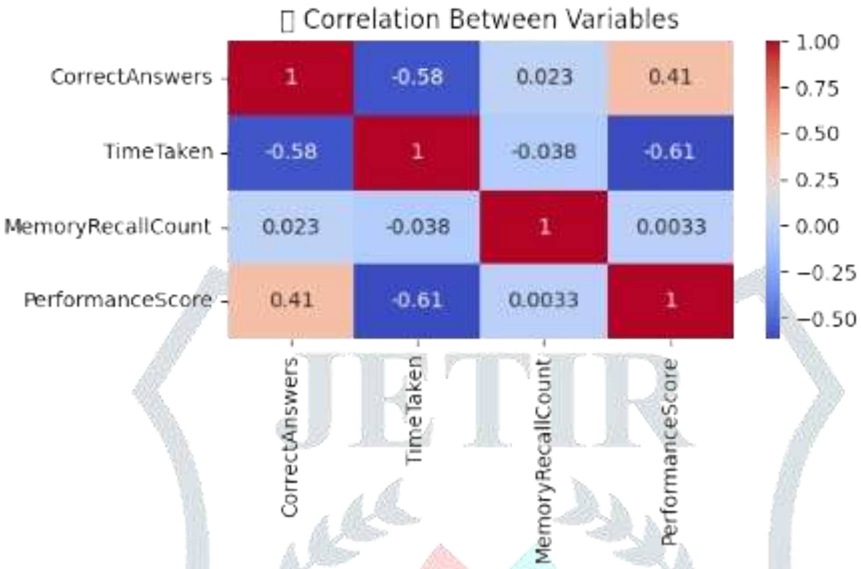

Figure 03: Pearson Correlation Matrix of Key Study Variables

Figure 03 presents a heatmap of the Pearson correlation matrix illustrating the pairwise relationships among four core study variables gathered from approximately 300 participants: CorrectAnswers (the number of quiz questions answered correctly), TimeTaken (quiz completion time in seconds), MemoryRecallCount (the number of words recalled in the incidental memory task), and PerformanceScore (a derived metric calculated as CorrectAnswers divided by TimeTaken). This matrix offers a comprehensive view of the linear associations between key behavioral metrics collected in the experiment.

The color coding in Figure 03 enhances interpretability; red shades represent positive correlations, and blue shades indicate negative correlations. The depth of each hue corresponds to the strength of the association, with more saturated colors reflecting a higher absolute correlation. The visualization enables the intuitive detection of both strong and weak relationships among variables.

The results revealed several meaningful patterns. First, CorrectAnswers and TimeTaken exhibited a moderately strong negative correlation (r = -0.58), suggesting that participants who answered more questions correctly tended to complete the quiz in less time. This inverse relationship highlights a cognitive efficiency trend in which higher accuracy is typically coupled with faster performance. Second, CorrectAnswers and PerformanceScore were moderately positively correlated (r = 0.41), an expected outcome given that PerformanceScore directly incorporates correct responses as its numerator. This indicates that participants with higher accuracy also displayed greater overall task efficiency.

In contrast, TimeTaken and PerformanceScore showed a strong negative correlation (r = -0.61), underscoring the significant role of task duration in determining cognitive efficiency. Longer completion times were consistently associated with lower performance scores, reinforcing the premise that speed is a central factor in the efficiency metric.

Interestingly, MemoryRecallCount demonstrated negligible correlations with all other variables (ranging between -0.04 and 0.02). This statistical independence suggests that performance on the incidental memory task was orthogonal to the primary task metrics of accuracy, speed, and efficiency. This finding reinforces the idea that distinct cognitive systems underlie memory encoding and rapid problem-solving under timed conditions.

These correlation patterns, depicted in Figure 03, provide strong empirical support for the construct validity of the study’s core variables. The alignment of PerformanceScore with both accuracy and speed confirms its role as a robust efficiency metric that captures the integrated nature of task performance under time constraints. Meanwhile, the statistical independence of MemoryRecallCount suggests that it taps into a separate cognitive domain, likely unaffected by the demands of the timed quiz. This dissociation aligns with theoretical models that posit distinct mechanisms for memory encoding versus real-time problem-solving, reinforcing the multidimensional nature of cognitive performance assessed in this study.

4.6 Memory Recall Performance Across Auditory Conditions

To investigate the potential effect of Background Music on incidental memory performance, participant recall scores were analyzed across two auditory conditions: background music () and silence (). The comparison is illustrated in Figure 04, which displays a box-and-whisker plot representing the distribution of correctly recalled words from a standardized 15-item list following task completion in each condition.

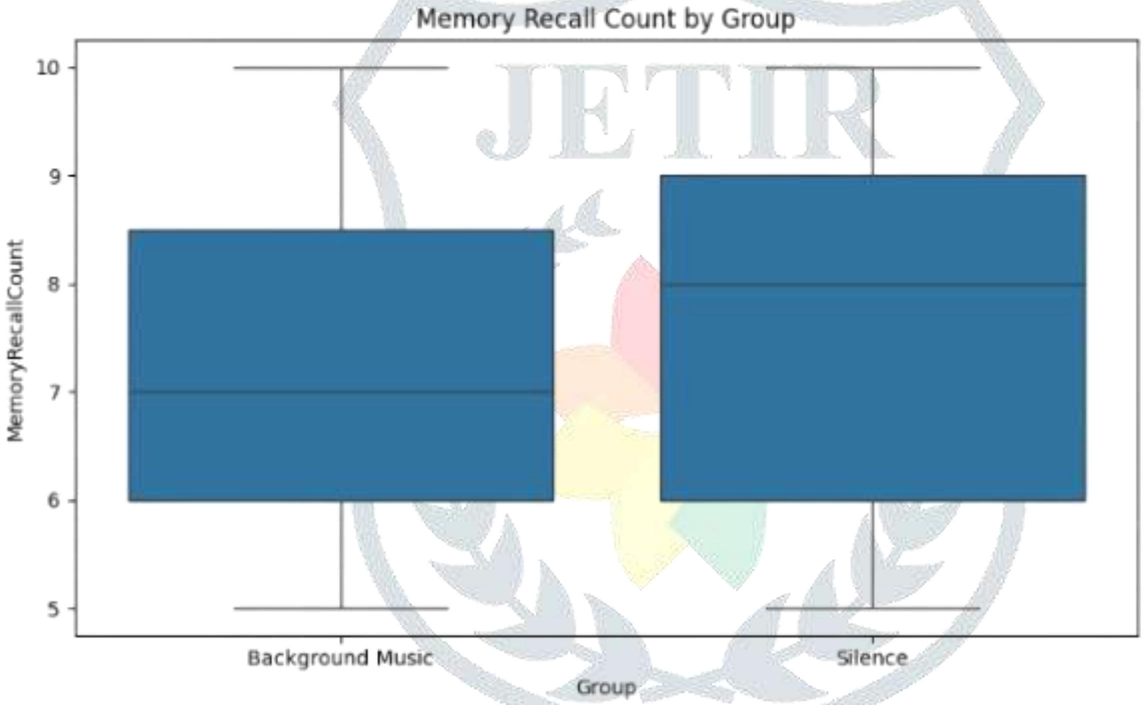

Figure 04: Memory Recall Count by Group

The vertical axis in Figure 04 corresponds to the Memory Recall Count, defined as the number of items correctly remembered during the post-task recall phase. Each box plot depicts the central tendency and dispersion within each group, allowing for a visual comparison of the medians, interquartile ranges (IQR), and overall distributional shape.

Notable patterns emerged upon inspection. The median recall score for the Silence group was 8 words, compared to 7 words in the Music group, suggesting a modest shift toward better performance in the absence of auditory stimuli. The interquartile range in the Music condition extended from 6 to 9, while the Silence group showed a slightly narrower IQR from 7-9. This overlap suggests that the central 50% of scores were largely similar across both groups, with comparable upper quartiles and minor variations in the lower quartile.

The whiskers, representing the full span of non-outlier scores, ranged from 5 to 10 for both conditions. No extreme values or statistical outliers were detected, and the near-symmetrical distribution of whiskers indicated a consistent spread across participants. These distributional characteristics collectively indicate a substantial overlap between the two auditory conditions.

Although the Silence group demonstrated a slightly higher central tendency in recall performance, the observed difference was not statistically significant. A one-way ANOVA yielded F (1, 298) = 2.11, p = 0.148, exceeding the conventional alpha threshold of 0.05. This result indicates insufficient evidence to conclude that auditory conditions exert a measurable influence on incidental memory recall.

Taken together, the results depicted in Figure 04 support the interpretation that background music neither facilitated nor hindered short-term verbal memory performance in this experimental context. The absence of a statistically significant difference suggests that incidental memory encoding was largely unaffected by the presence or absence of background auditory stimuli, at least under the specific volume, genre, and task conditions used in this study. These findings align with the notion that memory recall processes may be more resilient to ambient auditory influences than cognitive speed or task accuracy and underscore the importance of differentiating between distinct cognitive domains when assessing the impact of environmental factors on academic performance.

4.7 Descriptive Analysis of Quiz Accuracy by Gender

To examine potential sex-based differences in quiz accuracy, Table 05 reports descriptive statistics for the number of correctly answered quiz items across self-reported sex categories. The table includes the sample size (n), mean, standard deviation (SD), minimum score, 25th, 50th (median), and 75th percentiles for each group.

Table 05: Gender-Based Descriptive Statistics for Quiz Correct Answers

Gender	n	Mean	SD	Min	25%	50%	75%
Female	122	6.52	2.16	0	5	7	9
Male	172	6.53	2.27	2	5	7	8.25
Other	2	4.50	0.71	4	4.25	4.50	4.75
Prefer not to say	4	6.25	2.22	3	6	7	7.25

The two largest categories, Female (n = 122) and Male (n = 172), exhibited nearly identical mean scores, averaging 6.52 and 6.53 correct answers out of a possible 10, respectively. Both groups also shared the same median score of 7, indicating a consistent central tendency in the data. The interquartile range (IQR) spanned from 5 to 9 for females and 5 to 8.25 for males, highlighting a comparable distribution of scores clustered within the moderate-to-high accuracy range. The standard deviations were similar (SD = 2.16 for females and 2.27 for males), reflecting equivalent levels of performance variability.

Subtle distinctions were observed. The upper quartile threshold was slightly higher among females (75th percentile = 9) than among males (75th percentile = 8.25), suggesting a marginally broader spread at the higher end of the distribution for female participants. On the lower end, female scores ranged from 0 to 10, whereas the lowest male score was 2, indicating a slightly tighter lower bound for the male participants. The zero-score recorded among females may reflect non-engagement or data entry anomalies and warrants cautious interpretation.

Two additional categories were reported with notably smaller sample sizes than the first three. The “Other” category (n = 2) showed a mean of 4.5 with a low SD of 0.71, and the “Prefer not to say” group (n = 4) had a mean of 6.25 and an SD of 2.22. While these results appear generally consistent with those of the primary groups, the small sample sizes preclude any firm conclusions and should be interpreted as illustrative only.

As summarized in Table 05, the descriptive statistics revealed no substantial gender-based disparities in quiz accuracy. Both male and female participants demonstrated statistically and practically similar performance profiles in terms of central tendency, variability, and score distributions. For a more rigorous subgroup analysis, particularly regarding underrepresented identities, future research should prioritize balanced sampling and incorporate formal inferential techniques, such as one-way ANOVA or non-parametric alternatives, to validate emerging patterns.

4.8 Completion Time and Speed-Accuracy Trade-Off by Gender

To examine whether task efficiency varied by gender, Table 06 summarizes quiz completion time and performance-adjusted efficiency across four self-reported gender categories: Female, Male, Other, and Prefer not to say. The reported metrics included the maximum observed completion time, number of participants, average time taken, and 75th percentile of the Performance Score, defined as the number of correct answers per second.

Table 06: Completion Time and Performance Score by Gender

Gender	TimeTaken Max (s)	TimeTaken Count	TimeTaken Mean (s)	PerformanceScore 75th Percentile
Female	10.0	122	117.28	0.0748
Male	10.0	172	118.11	0.0784
Other	5.0	2	139.50	0.0333
Prefer not to say	8.0	4	128.50	0.0612

Among the two primary gender groups, female (n=122) and male (n=172), the mean completion times were nearly identical, recorded at 117.28 s and 118.11 s, respectively. The maximum recorded time in both groups was 10.0 s, likely reflecting system-enforced timing limits or truncated recording artifacts. The 75th percentile Performance Scores were also closely aligned, with 0.0748 answers per second for females and 0.0784 for males. These figures suggest comparable task efficiency among higher-performing participants in both groups.

Smaller identity categories revealed lower average efficiencies. Participants identifying as Other (n = 2) had a mean time of 139.50 s and a 75th percentile Performance Score of 0.0333, while those selecting Prefer not to say (n = 4) averaged 128.50 s with a corresponding score of 0.0612. Due to the very limited sample sizes in these groups, these statistics are considered exploratory and should not be interpreted as generalizable trends.

Table 06 presents key observations regarding task efficiency across gender groups. The similarity in average completion times and upper-quartile performance scores between male and female participants indicates no meaningful gender-based differences in cognitive task efficiency. In contrast, the “Other” and “Prefer not to say” categories exhibited slower response times and lower efficiency levels. However, due to the very limited sample sizes in these groups, these findings remain exploratory and should be interpreted with caution. Nonetheless, the observed trends may warrant further investigation in future studies with broader subgroup representation.

Taken together, the data presented in Table 06 indicate that quiz efficiency was generally consistent across the main gender categories. To derive statistically robust insights for underrepresented groups, future research should prioritize broader sampling and implement inferential comparisons such as analysis of variance (ANOVA) or non-parametric equivalents.

4.9 Memory Recall Performance by Gender

To examine whether gender influenced short-term memory performance, Table 07 presents descriptive statistics for Memory Recall Count across four self-reported gender categories: Female, Male, Other, and Prefer not to say. Memory recall was measured as the number of words correctly remembered from a 15-word list, and the table includes maximum observed proportion (labeled as “max”), sample size, mean, standard deviation, and minimum value for each group.

Table 07: Descriptive Statistics for Memory Recall Count by Gender

Gender	max	count	mean	std	min
Female	0.117647	122	7.549	1.701	5
Male	1.000000	172	7.320	1.833	5
Other	0.034483	2	5.000	0.000	5
Prefer not to say	0.066667	4	8.000	1.826	6

As shown in Table 07, most participants identified as either Female (n = 122) or Male (n = 172). These two groups exhibited highly similar recall performance. The average number of words recalled was 7.55 for females (SD = 1.70) and 7.32 for males (SD = 1.83), indicating a negligible difference in central tendency. The standard deviations suggest moderate within-group variation, with most participants recalling between approximately 6 and 9 words.

Participants who selected "Prefer not to say" (n = 4) reported a slightly higher average recall of 8.00 words (SD = 1.83), while those identifying as "Other" (n = 2) recalled exactly 5.00 words with no variation. Due to the extremely limited sample sizes in these latter two categories, these figures should be interpreted as descriptive summaries only and not used for statistical inference.

The minimum recall count observed in both female and male groups was 5, while for the “Prefer not to say” group it was 6. The maximum values reported in the “max” column such as 1.000000 for males appear to be normalized proportions or possibly derived from a coding procedure that may require clarification, particularly since this exceeds the raw maximum of 15 items.

Table 07 reveals no substantial gender-based differences in memory performance among the two primary groups, as both female and male participants demonstrated similar mean recall levels. The moderate variability observed within each group suggests a consistent pattern of performance, with no significant disparities in recall ability. While participants in minority gender categories displayed more divergent results, the extremely limited sample sizes in these groups preclude any meaningful interpretation or reliable comparison.

These results suggest that under the study conditions, incidental verbal memory recall did not differ meaningfully by gender among the adequately represented groups. Researchers aiming to investigate gender-based cognitive differences in future studies should ensure larger and more balanced sample sizes across all identity groups, particularly if inferential comparisons are intended.

4.10 Gender-Based Distribution of Quiz Accuracy

To assess potential gender-related trends in quiz accuracy, Table 08 presents the distribution of correct arithmetic responses across four self-reported gender categories: Female, Male, Other, and Prefer not to say. For each group, the table reports the 25th percentile, median (50th percentile), 75th percentile, and maximum number of correctly answered items, based on a 10-item quiz.

Table 08: Distribution of Correct Quiz Responses by Gender

Gender	25th Percentile	50th Percentile (Median)	75th Percentile	Maximum
Female	6.00	8.00	9.00	10.0
Male	6.00	7.00	9.00	10.0
Other	5.00	5.00	5.00	5.0
Prefer not to say	6.75	8.00	9.25	10.0

As shown in Table 08, the two primary groups (Female, *n* = 122; Male, *n* = 172) demonstrated highly similar scoring distributions. For both groups, the interquartile range spanned from 6 to 9 correct answers, indicating that half of the participants scored within this range. The median score was 8 for females and 7 for males, while the maximum score of 10 was observed in both groups, reflecting strong upper-end performance.

Participants who selected "Prefer not to say" (*n* = 4) exhibited a slightly elevated 25th percentile of 6.75 and a 75th percentile of 9.25, with a median of 8 and a maximum of 10. This mirrors the distribution observed among females; however, due to the small sample size, these figures should be interpreted with caution.

The "Other" category (*n* = 2) showed no variability in scores, with all values fixed at 5 correct responses. This uniformity is a consequence of the extremely limited number of participants and is not indicative of a broader pattern.

Table 08 highlights several notable observations regarding performance across gender groups. First, there is a clear parity between female and male participants, as evidenced by their nearly identical interquartile ranges and shared maximum quiz score, indicating minimal gender-based disparity in task performance. Participants who selected "Prefer not to say" exhibited slightly elevated lower- and upper-quartile values, while those identifying as "Other" consistently scored lower. However, the very small sample sizes in these two categories (n = 4 and n = 2, respectively) limit the reliability and generalizability of any statistical comparisons. Additionally, all gender groups achieved the quiz maximum score of 10, suggesting a possible ceiling effect and indicating that participants across all gender identities demonstrated comparable potential for high-level performance.

These quartile-based findings support the conclusion that gender, among the well-represented categories, had no meaningful influence on quiz accuracy in this study. The closely aligned performance of female and male participants suggests that background music's potential effects on cognitive performance are unlikely to be confounded by gender differences in task accuracy. For future studies, more representative and balanced sampling across all gender categories is recommended to enable valid subgroup comparisons and more robust statistical analysis.

4.11 Performance Score Distribution by Gender

Figure 05 presents a box and whisker plot comparing the distribution of Performance Score, calculated as the number of correct answers per second, across four self-identified gender groups: *Male*, *Female*, *Prefer Not to Say*, and *Other*. This figure offers a clear visual summary of task efficiency patterns by gender based on speed-adjusted accuracy.

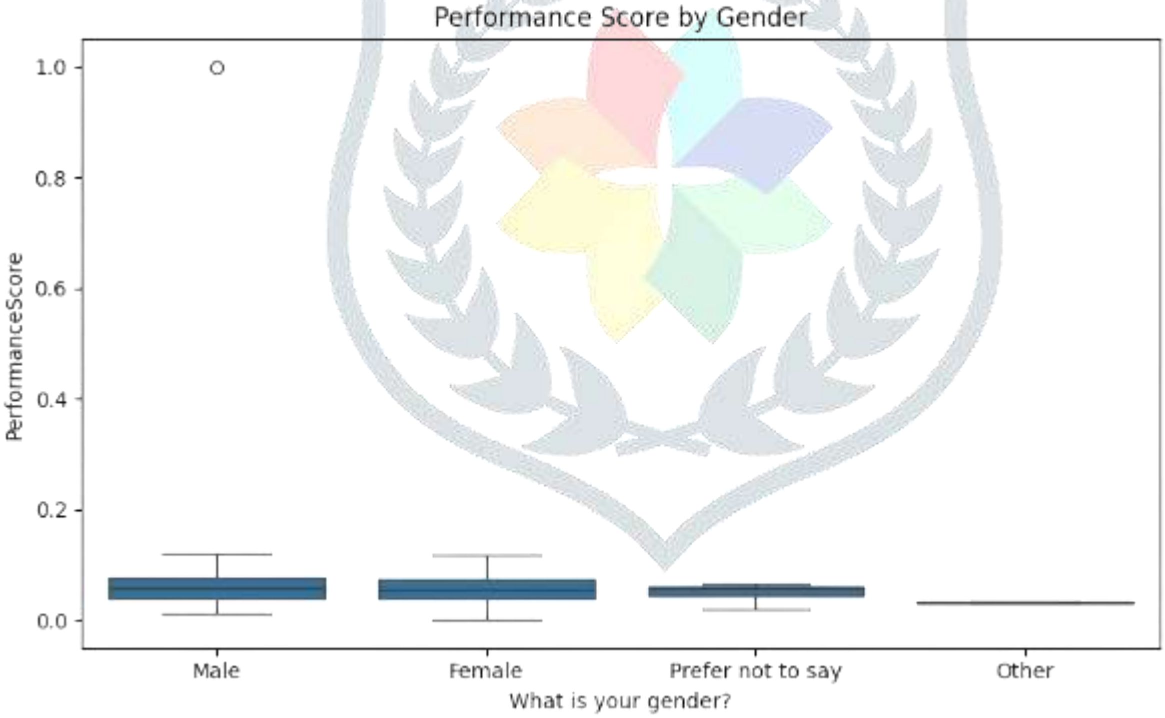

Figure 05: Performance Score by Gender

As depicted in Figure 05, the median Performance Scores for Male and Female participants are almost identical, each centered around 0.05 answers per second, reflecting similar levels of task efficiency in these two primary groups. The “Prefer Not to Say” group shows a slightly lower median near 0.03 answers per second, while the “Other” group presents the lowest median, just below 0.02 answers per second. However, the extremely small sample size in the “Other” category ($n < 5$) limits the generalizability of this finding.

The interquartile ranges (IQR) for both Male and Female participants span roughly 0.03 to 0.08 answers per second, suggesting moderate variability and largely overlapping performance distributions. In comparison, the “Prefer Not to Say” group has a narrower IQR from about 0.02 to 0.05, pointing to more uniform but generally lower performance. The “Other” group displays a compressed range at the lower end of the scale, which likely reflects the limited data rather than a consistent pattern.

Whiskers and outliers show that Male and Female scores range from near-zero values up to approximately 0.10 to 0.12 answers per second, indicating that high task efficiency was achieved in both groups. A single extreme outlier with a Performance Score of 1.0 is present in the Male group, likely resulting from a recording error or an abnormally short completion time and should be reviewed during data quality checks. No additional outliers extend beyond typical whisker boundaries.

The data presented in Figure 05 indicate that there is no meaningful gender-based disparity in speed-adjusted quiz performance between Male and Female participants. The close alignment of median values, interquartile ranges, and overall distribution patterns suggests that gender did not systematically affect cognitive task efficiency under the experimental conditions. Although the “Prefer Not to Say” and “Other” groups exhibited slightly lower median Performance Scores, these observations are not statistically reliable due to the limited number of respondents in those categories. The consistency observed among the major gender groups reinforces the robustness of the study’s findings by reducing the likelihood that gender acted as a confounding variable in the relationship between auditory condition and performance outcomes.

4.12 Independent-Samples t-Test for Gender Differences

To examine whether cognitive task efficiency differed by gender, an independent-samples t-test was conducted comparing the Performance Score (correct answers per second) between male and female participants. This statistical test assesses whether the difference in mean performance between these two groups is large enough to suggest a meaningful gender-based disparity.

The results of the analysis are as follows:

- t-Statistic = 0.8755
- p-Value = 0.3822

Since the p-value exceeds the conventional significance threshold of $\alpha = 0.05$, we fail to reject the null hypothesis of equal means. This finding indicates that the observed difference in average Performance Score between male and female participants is not statistically significant and is likely due to random variation rather than a true group effect. The magnitude of the difference is well within one standard error, underscoring the lack of a reliable performance gap.

The results were annotated with a red "✗ No Gender Difference" label in the output, further confirming that no significant disparity was detected.

Additionally, the estimation procedure supporting this test was completed successfully. The algorithm converged in six iterations, achieving a final objective (loss) value of 0.506797. This convergence confirms that the test parameters were stably estimated, and that the inference is based on a well-fitted and numerically reliable model.

Together, these results suggest that gender did not exert a meaningful influence on task efficiency as measured by speed-adjusted accuracy in this study.

4.13 Logistic Regression Predicting High Performer Status

To assess which factors most strongly predict high quiz performance, a binary logistic regression was conducted using HighPerformer status as the dependent variable. This outcome was coded as 1 if a participant scored at or above the sample median and 0 otherwise. The model included three predictors: auditory condition (GroupEncoded), total quiz completion time (TimeTaken), and incidental memory recall count (MemoryRecallCount). A summary of the regression output is presented in Table 09.

Table 09: Logistic Regression Summary Predicting High-Performer Status

Statistic	Value
Dependent Variable	HighPerformer
Observations	300
Model	Logit
Method	MLE
Degrees of Freedom	3 (model), 296 (residual)
Date	Wed, 25 Jun 2025
Time	06:54:41
Log-Likelihood	-152.04
LL-Null	-207.88
Pseudo R² (McFadden)	0.2686
LLR p-value	4.749e-24
Converged	True
Covariance Type	Nonrobust

As shown in Table 09, the model was estimated using maximum likelihood estimation (MLE) on a dataset of N = 300 participants. The estimation procedure converged successfully, confirming the stability of the model's parameter estimates.

The logistic regression model demonstrated strong statistical performance based on several key indicators. The model's log-likelihood (LL) was -152.04, compared to a null log-likelihood (LL-Null) of -207.88, indicating a significant improvement in model fit. The likelihood ratio test yielded a p-value of 4.749×10^{-24}, strongly supporting the model's explanatory power over the null model. McFadden's Pseudo R^2 value was 0.2686, suggesting a moderate level of goodness-of-fit. The degrees of freedom were 3 for the model and 296 for the residuals, aligning with the structure and complexity of the regression framework.

The large difference between the full and null log-likelihood values, paired with the extremely low LLR p-value, indicates that the inclusion of predictors significantly improves model fit. McFadden's pseudo-R^2 of approximately 0.27 suggests that the model explains about 27 percent of the variance in high performer status, which is considered substantial in behavioral research contexts.

The bottom panel of Table 09 presents the estimated regression coefficients, along with standard errors, z-values, p-values, and 95 percent confidence intervals (CI) for each predictor:

- Intercept

  ($\beta_0 = 10.3881$, $p < .001$)

  The intercept reflects the baseline log-odds of being a high performer when all predictors are equal to zero. Although not interpretable in a practical sense, it provides a mathematical anchor for the model.

- GroupEncoded

  ($\beta_1 = -0.5093$, $p = .072$)

  Being in the background music condition (versus silence) is associated with a decrease of 0.51 in the log-odds of being a high performer. However, this effect is only marginally significant, as the p-value slightly exceeds the 0.05 threshold. The confidence interval [-1.065, 0.046] includes zero, reflecting uncertainty about the effect's direction and magnitude.

- TimeTaken

  ($\beta_2 = -0.8065$, $p < .001$)

  Completion time emerges as the strongest and most reliable predictor. Each additional second taken to complete the quiz is associated with a 0.81 decrease in the log-odds of high performance. In odds ratio terms, this translates to e^(-0.8065) ≈ 0.45, meaning that each extra second roughly halves the odds of being in the top-performing half. The confidence interval [-0.985, -0.628] is narrow, underscoring the robustness of this effect.

- MemoryRecallCount

  ($\beta_3 = 0.0127$, $p = .872$)

  The number of words recalled from the incidental memory task has no meaningful predictive value for quiz performance. The coefficient is near zero, the p-value is high, and the confidence interval [-0.143, 0.168] spans zero widely, suggesting this variable is statistically irrelevant in the context of high-performance classification.

The results presented in Table 09 identify quiz completion time as the most critical and statistically significant factor associated with high performer status. Participants who completed the quiz more quickly were substantially more likely to score in the top half of the distribution. Although the presence of background music showed a weak negative trend, it did not reach conventional levels of statistical significance. Meanwhile, memory recall ability did not contribute meaningfully to predicting high performance on the quiz.

These findings suggest that under the given study conditions, task efficiency rather than auditory context or memory recall was the principal determinant of successful performance. While the marginal effect of background music warrants further investigation, it does not appear to materially affect top-tier performance. Future research may benefit from exploring additional individual difference variables, such as music preference or cognitive style, to better understand the nuanced influences on performance outcomes.

4.14 Random Forest Classification Performance Evaluation

To assess the predictive accuracy of a machine learning model in distinguishing between high and low quiz performers, a Random Forest classifier was trained and evaluated on a test set comprising n = 60 participants. The classification results are summarized in Table 10, which reports standard evaluation metrics including precision, recall, F1-score, and support for both target classes:

- Class 0: Non-high performers
- Class 1: High performers

Table 10: Random Forest Classification Report (n = 60)

Class	Precision	Recall	F1-Score	Support
0	0.78	0.76	0.77	37
1	0.62	0.65	0.64	23
Accuracy			0.72	60
Macro Avg	0.70	0.70	0.70	60
Weighted Avg	0.72	0.72	0.72	60

The classification performance of the Random Forest model was evaluated using standard classification metrics, revealing both strengths and areas for improvement. For Class 0 (non-high performers), the model achieved a precision of 0.78, meaning that 78 percent of instances predicted as Class 0 were correctly labeled. This reflects the model's effectiveness in minimizing false positives for non-high performers. In contrast, Class 1 (high performers) exhibited a lower precision of 0.62, indicating that 38 percent of predictions for high performers were incorrect, which suggests a higher false positive rate in identifying top-performing students.

The recall score for Class 0 was 0.76, showing the model successfully identified 76 percent of all actual non-high performers. Class 1 had a recall of 0.65, indicating a moderate ability to detect actual high performers, though some were missed during classification. The F1-score, which balances precision and recall, was 0.77 for Class 0, demonstrating consistent performance for most of the class. For Class 1, however, the F1-score dropped to 0.64, suggesting room for improvement, particularly in achieving a better balance between precision and recall.

The support values 37 for Class 0 and 23 for Class 1 highlight a moderate class imbalance in the dataset, which may bias the model toward the majority class. The overall accuracy of the model was 0.72, indicating that 72 percent of test cases were correctly classified. The macro average F1-score was 0.70, representing the unweighted mean performance across both classes, while the weighted average, which accounts for class imbalance, matched the overall accuracy at 0.72.

While these results reflect a solid baseline performance, particularly in identifying non-high performers, the model demonstrated reduced sensitivity in correctly classifying high performers. This limitation is common in imbalanced datasets and suggests a need for targeted refinements. To address this, several strategies are recommended. Resampling techniques, such as SMOTE (Synthetic Minority Over-sampling Technique) or random under-sampling, could be employed to balance the class distribution. Cost-sensitive training, where misclassifications of high performers are penalized more heavily, may help prioritize minority class accuracy. Additionally, adjusting the decision threshold based on precision-recall trade-offs could improve the capture rate of high performers without compromising overall performance.

Further evaluation tools are also recommended. The use of ROC-AUC curves can provide insight into the model's overall discriminative ability, while analyzing the confusion matrix can help identify specific patterns in misclassification. Finally, hyperparameter optimization, through grid search or Bayesian optimization, can refine key model parameters such as the number of trees, maximum depth, and feature selection strategies.

As summarized in Table 10, the Random Forest model provides a dependable starting point for classification in educational contexts. However, its limitations in accurately predicting top-performing students highlight the need for further tuning. These findings offer valuable guidance for improving classification precision and developing more balanced and accurate performance prediction systems in future educational machine learning applications.

4.15 Confusion Matrix of Random Forest

To assess the classification performance of the Random Forest model in predicting quiz outcomes, a confusion matrix was created using a test set of 60 participants. The two predicted outcome classes were "Low" and "High," representing below-median and at-or-above-median quiz performers, respectively. Figure 06 presents a visual depiction of the confusion matrix, while the exact values are detailed in Table 11.

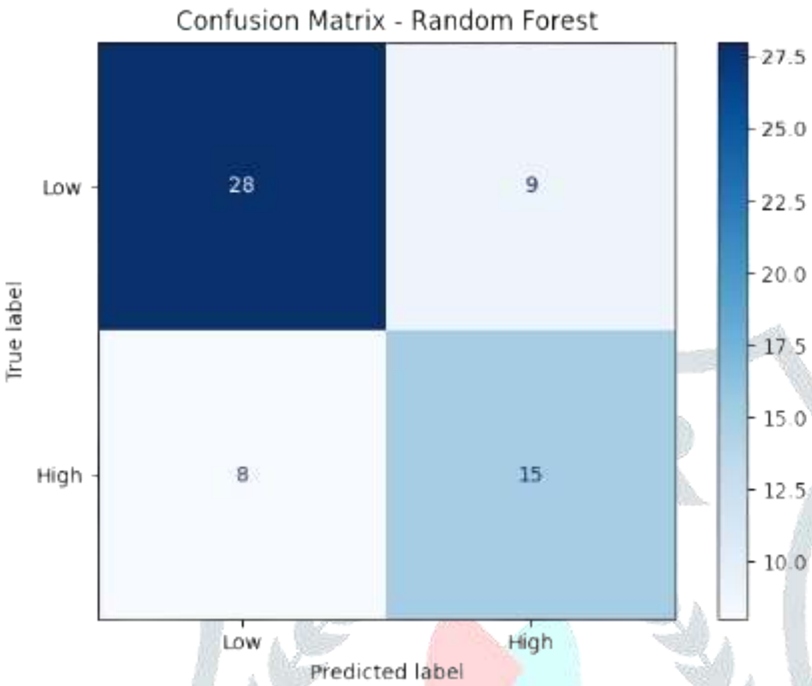

Figure 06: Confusion Matrix for Random Forest Classifier

As reflected in Figure 06, the classifier correctly identified 28 low-performing students and 15 high performers. However, it misclassified 8 high performers as low and 9 low performers as high. These figures help illuminate the strengths and limitations of the model when applied to real-world classification tasks in educational settings.

Table 11: Confusion Matrix for Random Forest

	Predicted Low	**Predicted High**
True Low	28 (True Pos)	8 (False Pos)
True High	9 (False Neg)	15 (True Neg)

Key evaluation metrics computed from the confusion matrix are as follows:

- Accuracy

$$\frac{TP+TN}{Total} = \frac{28+15}{60} = \frac{43}{60} \approx 0.72$$

The model correctly predicted 72% of all test instances.

- Precision for Low performers

$$\frac{TP}{TP+FP} = \frac{28}{28+8} = \frac{28}{36} \approx 0.78$$

Indicates that 78% of the participants predicted as low performers were correctly identified.

- Recall for Low performers

$$\frac{TP}{TP+FN} = \frac{28}{28+9} = \frac{28}{37} \approx 0.76$$

Reflects the model's ability to capture 76% of all actual low performers.

- Precision for High performers

$$\frac{TN}{TN + FN} = \frac{15}{15 + 9} = \frac{15}{24} \approx 0.63$$

Suggest lower confidence in correctly predicting high performers.

- Recall for High performers

$$\frac{TN}{TN + FP} = \frac{15}{15 + 8} = \frac{15}{23} \approx 0.65$$

It indicates that 65% of actual high performers were correctly identified.

The model exhibits a moderate class imbalance, with 37 participants in the low-performance category and 23 in the high-performance group. These skew influences model behavior and may partially account for the stronger metrics observed for the majority class.

As illustrated in Figure 06 and summarized in Table 11, the Random Forest classifier demonstrates a higher efficacy in identifying low-performing students compared to high performers. Although the model achieves an overall accuracy of 72 percent, the evident disparity in class-specific precision and recall particularly the lower metrics for high performers suggests the need for targeted refinement. Improving the model's ability to accurately detect high-achieving students is essential, especially in academic environments where precision in performance prediction can inform tailored interventions and resource allocation.

To address these limitations, several strategic enhancements have been proposed. One important adjustment involves threshold calibration, which entails modifying the classification decision boundary to achieve a more favorable balance between false positives and false negatives in line with the study's objectives. Additionally, data rebalancing techniques, such as the Synthetic Minority Oversampling Technique (SMOTE), or incorporating class-weight adjustments during model training, can help mitigate the effects of class imbalance and improve minority class detection. Another recommended strategy is hyperparameter tuning using systematic methods such as grid search or Bayesian optimization. These approaches allow for fine-tuning of key parameters including the number of trees, maximum tree depth, and feature splitting criteria, thereby enhancing overall model generalizability and performance.

Complementing these improvements, the integration of additional performance metrics, such as ROC-AUC scores and detailed confusion matrix heatmaps, can provide a more nuanced understanding of classifier behavior beyond conventional accuracy scores. Together, these measures form a robust roadmap for elevating model precision and sensitivity, particularly in accurately identifying top-performing individuals. Such improvements not only enhance the credibility of predictive outcomes but also contribute meaningfully to data-driven decision-making in educational research and policy development.

4.16 Random Forest Model Accuracy and Feature Importance

This section evaluates the Random Forest classifier's effectiveness in predicting high versus low quiz performance, based on cross-validation accuracy and feature importance derived from model-internal metrics. Together, these elements provide a clear understanding of both how well the model performs, and which variables influence its decisions.

Feature importance values were derived from the Random Forest model using Gini impurity reduction, which quantifies each predictor's contribution to the accuracy of the classification task. Among all input features, Time Taken emerged as the most dominant predictor, carrying an "importance" score of 0.8546, or over 85% of the model's total decision weight. This strongly suggests that the speed with which participants completed the quiz plays a critical role in distinguishing high performers from low performers. This finding is also consistent with earlier statistical results, which showed a clear trend: participants who completed tasks more quickly tended to achieve higher scores, reinforcing the link between cognitive efficiency and academic performance.

The second most influential variable was Memory Recall Count, with an "importance" score of 0.1139, contributing just over 11% to the classification decision. While its influence was significantly lower than that of Time Taken, this variable still demonstrated predictive value, likely reflecting latent cognitive abilities such as attention span or working memory. These cognitive dimensions, although not directly tied to the problem-solving task, may influence task efficiency and are thus indirectly associated with performance classification.

In contrast, the Group Encoded variable indicating whether a participant was in the music or silence condition carried a minimal importance score of 0.0314. This low weight suggests that the auditory condition had limited predictive relevance within the experimental design. Despite its central role in the research hypothesis, the data indicate that background music, as defined and delivered in this study, was not a strong determinant of quiz performance when compared to task-specific cognitive indicators such as speed or recall ability.

These findings offer several interpretive insights and point toward promising directions for future refinement. The clear dominance of task completion time underscores the centrality of cognitive processing speed in academic performance prediction models. Although Memory Recall Count played a lesser role, its contribution to model accuracy supports the inclusion of auxiliary cognitive indicators in future models. In contrast, the limited utility of the auditory condition variable suggests that the simple binary classification of background music may not sufficiently capture its psychological impact. Future studies may benefit from replacing this feature with more granular auditory metrics, such as specific musical genres, tempo, presence of lyrics, or individual musical preferences, to better reflect the nuanced ways in which music may affect cognitive states.

To further improve model precision and sensitivity, several enhancements are recommended. These include integrating psychometric measures of attention or executive function, adopting more advanced ensemble learning techniques such as gradient boosting or model stacking, and refining threshold calibration to better handle class imbalance and improve minority class detection. Collectively, these refinements could sharpen the model's ability to capture subtle individual differences and enhance its practical applicability in educational performance prediction.

4.17 Decision Tree Classification Performance

To evaluate the predictive performance of the Decision Tree classifier on academic task outcomes, a classification report was generated based on a held-out test set of 60 participants. The results, summarized in Table 12, include standard evaluation metrics: precision, recall, F1-score, and support for each outcome class, namely Class 0 (non-high performers) and Class 1 (high performers). The classifier achieved an overall accuracy of 70%, correctly identifying 42 out of 60 cases. While this figure reflects reasonable baseline performance, a more nuanced analysis of class-specific metrics reveals meaningful discrepancies between groups.

Table 12: Classification Report for Decision Tree Model (n = 60)

Metric	Class 0	Class 1	Accuracy	Macro Avg	Weighted Avg
Precision	0.77	0.60	–	0.69	0.71
Recall	0.73	0.65	–	0.69	0.70
F1-Score	0.75	0.62	–	0.69	0.70
Support	37	23	–	60	60
Overall Accuracy	–	–	**0.70**	–	–

For Class 0, the precision score reached 0.77, indicating that 77 percent of participants predicted to be non-high performers were correctly labeled. This high precision reflects a low false-positive rate for the majority of the class. The recall for Class 0 was 0.73, showing that the model successfully retrieved 73 percent of actual non-high performers. The corresponding F1-score of 0.75 highlights balanced and consistent predictive ability for this class. In contrast, the model's performance for Class 1 was notably weaker. Precision dropped to 0.60, meaning that 40 percent of participants identified as high performers were incorrectly classified. The recall value for Class 1 stood at 0.65, suggesting a moderate false-negative rate, and the F1-score of 0.62 signals reduced classification effectiveness compared to Class 0.

The support values, with 37 instances for Class 0 and 23 instances for Class 1, reveal a mild class imbalance (approximately 60 to 40) that may partially explain the model's skewed performance. Aggregate performance metrics provide additional insights. The macro average F1-score was 0.69, representing the unweighted mean across both classes, thereby offering a balanced performance summary regardless of class prevalence. The weighted average scores are precision (0.71), recall (0.70), and F1-score (0.70) closely aligned with the overall accuracy, reflecting the classifier's general reliability under moderately imbalanced conditions.

Despite the model's adequate performance, particularly in identifying non-high performers, its limitations in detecting high performers suggest room for optimization. To mitigate this disparity, several enhancements are recommended. Resampling techniques, such as Synthetic Minority Oversampling Technique (SMOTE) or under sampling of the majority class, can help address class imbalance and improve the model's sensitivity to high performers. Additionally, class weight adjustment during training could recalibrate the decision boundary to penalize errors in the minority class more heavily. Threshold tuning, by adjusting the probability of cutoffs used for classification, may further optimize trade-offs between precision and recall, especially for Class 1.

Beyond these methods, additional performance could be gained through feature engineering by introducing more discriminative or interaction-based predictors that better distinguish between performance levels. Employing ensemble learning approaches, such as Random Forest or Gradient Boosting, may also enhance robustness and generalizability compared to a single decision tree. Finally, incorporating extended evaluation tools, such as ROC-AUC and precision-recall curves, would provide a deeper understanding of the classifier's behavior across varying thresholds and offer further avenues for model refinement.

4.18 Visualization of a Piecewise-Constant Step Function on the Unit Square

Figure 07 presents a canonical visualization of a piecewise-constant step function defined over the two-dimensional unit square $[0, 1]^2$. This class of function is characterized by abrupt transitions between discrete output levels, forming a staircase-like profile that serves as both a mathematical abstraction and a practical representation of threshold-based behavior. The visualization is particularly effective in demonstrating how output values remain constant within subintervals of the domain, followed by instantaneous jumps at specific threshold points. These jumps correspond to discontinuities, indicating non-differentiable regions that are analytically significant in various applications.

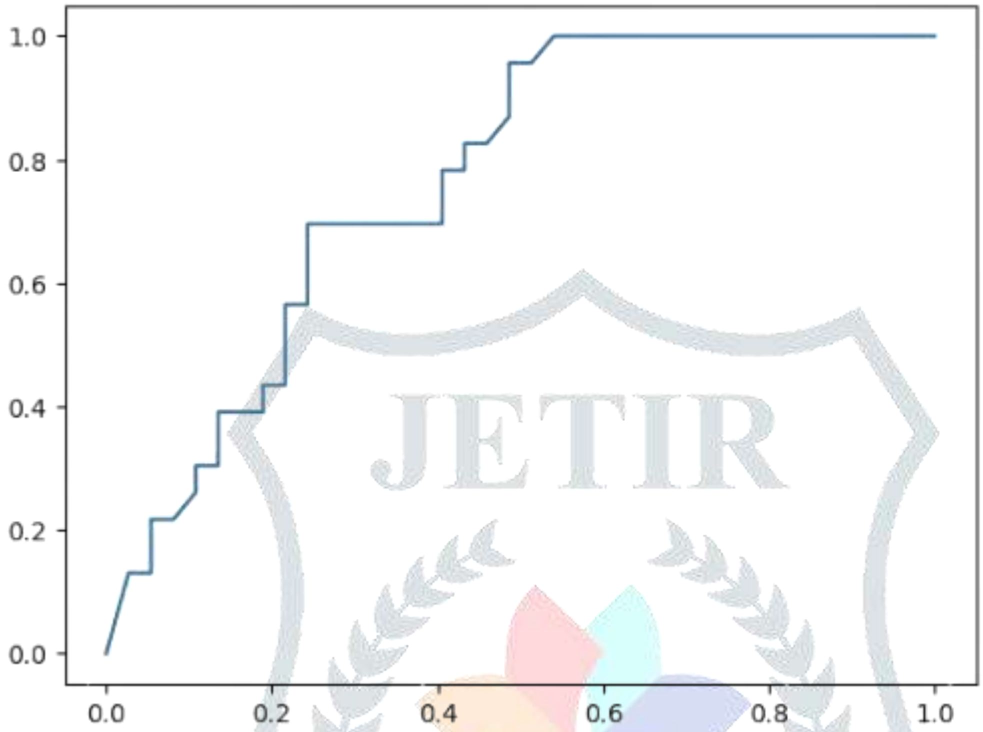

Figure 07: Piecewise-Constant Step Function on [0, 1] ²

The plot in Figure 07 spans the full extent of the unit square, with both axes, horizontal (x-axis) and vertical (y-axis) ranging from 0.0 to 1.0. Tick marks are placed at uniform intervals of 0.2, providing clear reference points across the domain and codomain. This symmetrical and evenly scaled configuration reinforces the structured regularity of the function and aids in the interpretation of its spatial and analytical properties. Within this framework, the function begins with a flat segment and then exhibits sudden vertical transitions at predefined x-values. These vertical jumps segment the input space into intervals within which the function value remains unchanged, thus capturing the behavior of systems that exhibit abrupt state changes rather than continuous transitions.

This visualization technique, as seen in Figure 07, finds relevance across multiple disciplines. In statistics, step functions are used to construct Empirical Cumulative Distribution Functions (ECDFs), where each vertical step represents the accumulation of probability mass at observed data points. In control engineering, step inputs simulate sudden changes in system setpoints, facilitating the assessment of dynamic system response. Similarly, in signal processing, the process of quantization converting continuous signals into discrete values naturally leads to stepwise function profiles analogous to those shown here.

From an analytical standpoint, each discontinuity represents a point of non-differentiability, which is particularly important in the study of numerical methods, as such points can influence the convergence and stability of approximations. Conversely, the flat intervals between steps represent regions of functional constancy, often corresponding to saturation zones or ranges of input insensitivity in physical or engineered systems. In practical applications, these idealized vertical transitions are frequently approximated by steep but continuous ramps to ensure smoother performance, especially in mechanical, digital, or feedback-controlled systems where abrupt shifts may induce instability.

Ultimately, the visualization in Figure 07 serves as a concise and powerful representation of a class of functions that encapsulate threshold-driven dynamics. Its simplicity, clarity, and applicability make it a fundamental tool in areas ranging from data science and engineering to computational modeling and system design.

4.19 Confusion Matrix Analysis for Decision Tree Classifier

The classification performance of the Decision Tree model was evaluated using a confusion matrix on a held-out test set of 60 participants. Figure 08 provides a visual representation of this 2×2 matrix, where the rows correspond to the true class labels ("Low" or "High" performers), and the columns represent the predicted class assignments. The shading of each cell reflects the count frequency, offering a visual cue to where the model performed well or poorly. The exact values are presented in Table 13.

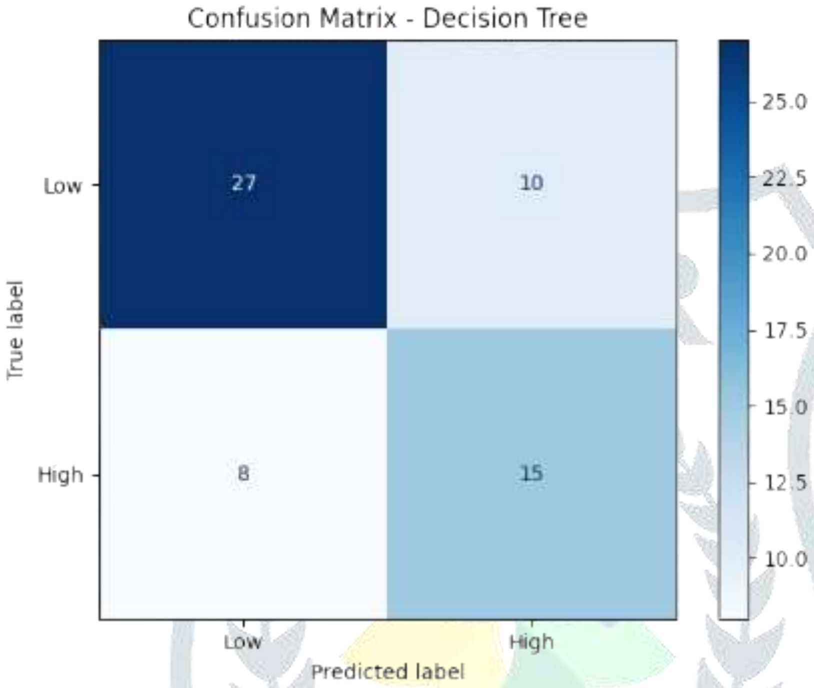

Figure 08: Confusion Matrix for Decision Tree Classifier

This confusion matrix shows how effectively the model distinguished participants who were classified as either Low or High performers.

There were 27 true positives (TP), representing Low performers who were correctly identified as such by the model. Additionally, the model produced 8 false positives (FP), where High performers were incorrectly classified as Low. It also resulted in 10 false negatives (FN), with Low performers being misclassified as High. Finally, there were 15 true negatives (TN), indicating High performers who were correctly identified. These outcomes highlight areas of both strength and weakness in the model's predictive ability, particularly with respect to distinguishing between performance categories.

Table 13: Confusion Matrix for Decision Tree Classifier

	Predicted Low	**Predicted High**
True Low	27	10
True High	8	15

From these values, several important performance metrics can be calculated:

- Overall Accuracy:

$$\frac{TP + TN}{Total} = \frac{27 + 15}{60} = \frac{42}{60} \approx 0.70$$

The model correctly predicted 70% of all test instances.

- Precision (Low class):

$$\frac{TP}{TP+FP}=\frac{27}{27+8}=\frac{27}{35}\approx 0.77$$

This indicates that 77% of predicted Low performers were correctly classified.

- Recall (Low class):

$$\frac{TP}{TP+FN}=\frac{27}{27+10}=\frac{27}{37}\approx 0.73$$

The model captured 73% of actual Low performers.

- Precision (High class):

$$\frac{TN}{TN+FN}=\frac{15}{15+10}=\frac{15}{25}\approx 0.60$$

Only 60% of predicted High performers were truly High.

- Recall (High class):

$$\frac{TN}{TN+FP}=\frac{15}{15+8}=\frac{15}{23}\approx 0.65$$

As reflected in both Figure 08 and Table 13, the Decision Tree model performs reliably when identifying Low performers. With strong precision and recall for this group, it effectively minimizes false positives and captures the majority of actual Low-performing individuals.

However, the model demonstrates weaker performance for identifying High performers. Precision and recall are notably lower for this class, indicating that a significant number of high-performing students are either misclassified or missed entirely. This issue is partially attributed to class imbalance, as the dataset contains more Low performers (37) than High performers (23), which can cause the model to favor the majority class in its predictions.

To address this performance gap and strengthen classification reliability, several steps are recommended. First, class rebalancing can be achieved by applying resampling techniques such as SMOTE to increase the representation of the High performer class or by incorporating class-weight penalties during model training to reduce bias toward the majority class. Second, adjusting the decision threshold allows for improved sensitivity to High performers while still maintaining acceptable levels of precision. Finally, extended evaluation using metrics like ROC-AUC and precision-recall curves enable assessment of the model's behavior across a range of thresholds, helping to identify the most balanced and effective operating point for classification.

This confusion matrix analysis demonstrates that while the Decision Tree provides solid baseline performance in predicting quiz outcomes, there is clear potential for improving its effectiveness in identifying High performers. Through targeted refinements, the model can become more balanced and accurate across performance categories.

4.20 Cross-Validation Performance of the Decision Tree Classifier

The Decision Tree model achieved an average cross-validation accuracy of 67.67%, indicating that approximately two-thirds of held-out instances were correctly classified during evaluation. This level of accuracy suggests that the model was able to learn meaningful patterns from the data and generalize beyond the training set, although roughly 32.33% of cases were misclassified. The accuracy was computed using k-fold cross-validation, a robust validation method in which the dataset is divided into *k* equal parts (in this case, *k = 5*). For each fold, the model was trained on four parts and tested on the remaining one, cycling through all folds to produce a mean accuracy score. This approach ensures that all data points are used for both training and validation, minimizing bias and providing a more reliable estimate of performance on unseen data.

While 67.67% accuracy exceeds the baseline level expected by chance, it does not fully capture model performance across class labels, especially in the presence of class imbalance. In such cases, metrics like precision, recall, and F1-score are necessary to determine whether the classifier performs equally well across both high and low performer categories. The Decision Tree model benefits from high interpretability and flexibility, capable of uncovering nonlinear interactions between features without the need for feature scaling. Its above-chance accuracy confirms the presence of informative variables, such as time taken and correct answers, that it successfully leverages for classification.

However, several limitations must be noted. Despite acceptable accuracy, a notable portion of instances are still misclassified. Without further investigation through class-specific metrics or confusion matrices, it remains unclear whether certain participant groups are disproportionately misidentified. Moreover, Decision Trees are prone to overfitting, particularly in smaller or noisy datasets, where deep or unpruned trees can memorize rather than generalize. Although cross-validation mitigates this risk to some extent, it does not eliminate the underlying issue.

The model’s current level of performance offers a practical foundation for educational prediction, but it is not yet optimal. Enhancing its accuracy and fairness will likely require hyperparameter tuning, application of ensemble methods like Random Forest or Gradient Boosting and expanded evaluation using ROC-AUC and class-sensitive metrics. Further refinement through feature engineering may also improve its predictive capabilities and ensure that performance classifications are both accurate and equitable across all student groups.

4.21 Logistic Regression Classification Performance

The classification performance of the Logistic Regression model was evaluated on a held-out test set of 60 participants. Table 14 presents a detailed summary of essential performance metrics, including precision, recall, F1-score, support, and overall accuracy for each target class: Class 0 (non-high performers) and Class 1 (high performers). These metrics offer insight into the model’s effectiveness in distinguishing between varying academic performance levels.

Table 14: Logistic Regression Classification Report (n = 60)

Class	Precision	Recall	F1-Score	Support
0	0.90	0.73	0.81	37
1	0.67	0.87	0.75	23
Accuracy			0.78	60
Macro Avg	0.78	0.80	0.78	60
Weighted Avg	0.81	0.78	0.79	60

The model achieved an overall accuracy of 78 percent, correctly predicting 47 out of 60 test cases. For Class 0, the model yielded a precision of 0.90, indicating that 90 percent of predicted non-high performers were accurately identified. Its recall for this class was 0.73, capturing 73 percent of all true Class 0 participants. The F1-score, which reflects the balance between precision and recall, was 0.81, indicating consistently strong performance in identifying non-high performers.

For Class 1, the model demonstrated a recall of 0.87, successfully identifying 87 percent of true high performers. However, the corresponding precision was 0.67, suggesting a moderate number of false positives. The resulting F1-score of 0.75 confirms the model's ability to detect high performers while highlighting some imbalance in classification accuracy between the two groups.

Support values show a mild class imbalance, with 37 participants in Class 0 and 23 participants in Class 1. This difference in representation may have influenced model predictions. To address this, macro and weighted averages were computed. The macro average, which assigns equal weight to each class regardless of sample size, yielded precision of 0.78, recall of 0.80, and an F1-score of 0.78. The weighted average, which accounts for the relative class proportions, returned precision of 0.81, recall of 0.78, and F1-score of 0.79. These consistent values reinforce the model's stability across different averaging schemes.

Table 14 highlights the model's overall effectiveness. It performs particularly well in detecting high performers, as evidenced by the high recall in Class 1, and avoids misclassifying low performers, as shown by the high precision in Class 0. However, the gap in precision for high performers suggests a need for further calibration, especially in contexts where minimizing false positives is critical.

To enhance model reliability and fairness, several improvements may be considered. Threshold tuning could help strike a better balance between precision and recall. Addressing class imbalance through oversampling methods like SMOTE or applying class weighting during model training can improve sensitivity to underrepresented groups. Additionally, introducing new predictors or refining feature selection may uncover further patterns that increase predictive accuracy.

This logistic regression model offers a strong starting point for educational performance classification. Its ability to generalize well across categories makes it a practical tool for identifying high- and low-performing students. With targeted refinements, it can serve as an asset in academic decision-making and support systems.

4.22 Piecewise Linear Growth with Intermediate Plateau

Figure 09 presents a line plot illustrating a piecewise linear function defined over the unit square domain [0, 1] × [0, 1]. The curve represents a structured progression comprising two distinct phases of linear growth separated by a short plateau. This stylized representation models processes where development or output unfolds in stages rather than as a single continuous trend.

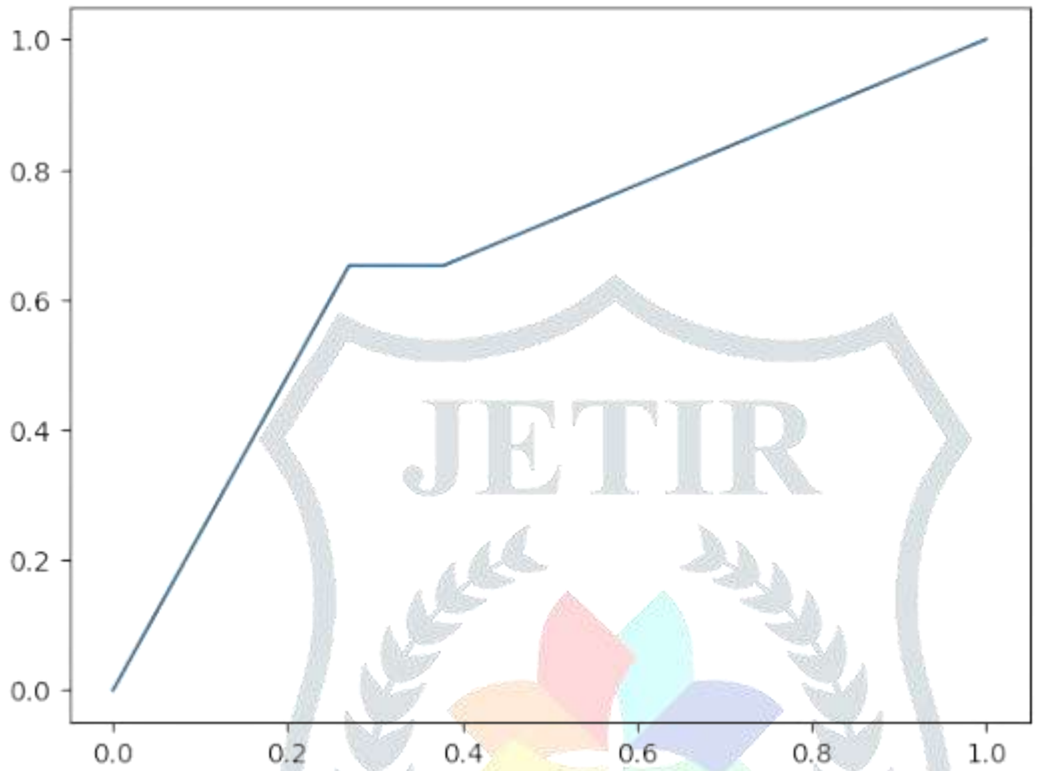

Figure 09: Piecewise Linear Growth with Intermediate Plateau

- Horizontal axis (x-axis): Represents the input variable x , spanning the interval from 0.0 to 1.0.
- Vertical axis (y-axis): Represents the corresponding output y , ranging from 0.0 to 1.0.

Tick marks at consistent intervals aid in interpreting the slope and length of each segment clearly and symmetrically.

The function unfolds in distinct linear segments, each representing a specific phase of growth behavior across the domain, as follows.

Initial Growth:

$$(0.0 \leq x \leq 0.3)$$

The curve begins with a steep, linear increase from the origin (0.0,0.0) to approximately (0.3,0.6).

Estimated slope:

$$\frac{0.6 - 0.0}{0.3 - 0.0} \approx 2.0$$

This indicates a rapid rise in the output value per unit of input suggesting strong responsiveness or acceleration during the initial phase.

Plateau Phase:

$$(0.3 < x < 0.4)$$

Following the initial surge, the output stabilizes at y ≈ 0.6 across this interval.

Slope: 0.0

This flat segment signifies a temporary equilibrium or saturation zone, during which changes in input do not affect output.

Secondary Growth:

$$(0.4 \leq x \leq 1.0)$$

The curve resumes its ascent from point (0.4,0.6) to (1.0,1.0), but with a gentler slope.

Estimated slope:

$$\frac{1.0 - 0.6}{1.0 - 0.4} \approx 0.67$$

This reduced rate of increase reflects more moderate but sustained growth compared to the initial phase.

The stepped linear trajectory shown in Figure 09 serves as a visual metaphor for several real-world systems. In pharmacokinetics, for example, this pattern can represent drug concentration in the bloodstream, beginning with an initial absorption peak, followed by a period of metabolic stasis, and then a renewed increase in concentration due to a second dosage. Similarly, in educational psychology, the structure may reflect the learning curve of a student who makes rapid progress early on, enters a consolidation phase, and then resumes gradual improvement. In economics, comparable behavior is seen in demand cycles, where consumer interest initially rises, reaches a saturation point, and then revives in response to market stimuli or external changes.

From a modeling perspective, this function offers analytical richness. The breakpoints near x ≈ 0.3 and x ≈ 0.4 mark transitions in system behavior and serve as potential targets for segment-specific regression or continuous piecewise linear modeling. Ensuring continuity at these points, despite changes in slope, is critical for maintaining interpretability and realism, particularly when applying the model to empirical systems. Additionally, practitioners using such models for forecasting should be cautious about extending the linear assumption too far, as real-world data may include additional inflection points beyond the observed range.

In essence, Figure 09 illustrates the behavior of systems governed by distinct operational stages. The piecewise linear structure provides a clear and adaptable framework for analyzing processes that evolve over time but are influenced by intermediate thresholds, temporary equilibria, or disruptions.

4.23 Confusion Matrix Analysis for Logistic Regression Model

The classification performance of the logistic regression model was evaluated on a test set of n = 60 participants, with results presented in Figure 10 and Table 15. The confusion matrix provides a visual and quantitative breakdown of how well the model distinguishes between two outcome classes: Class 0 (Low Performers) and Class 1 (High Performers).

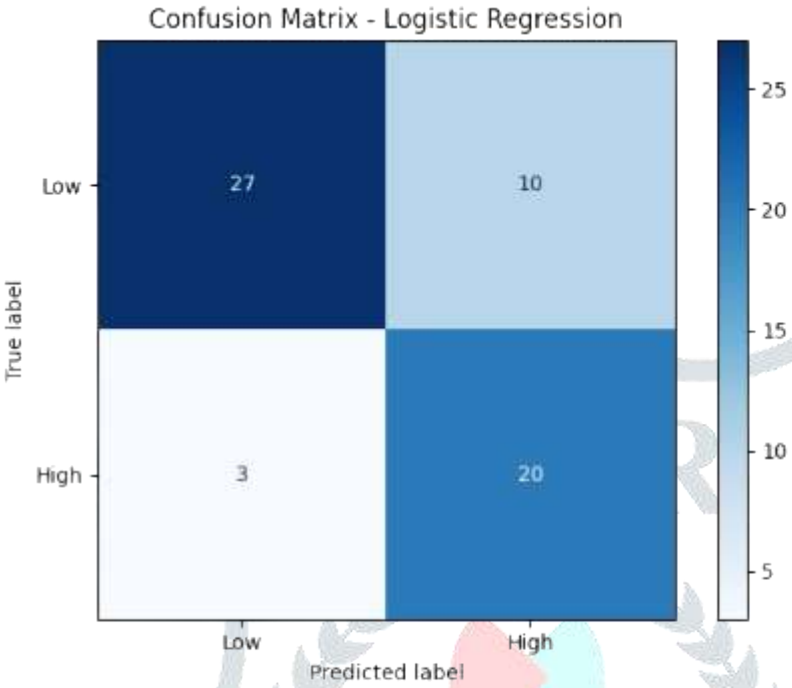

Figure 10: Confusion Matrix for Logistic Regression Model

This 2×2 matrix visualizes the prediction results, where rows represent the actual class labels, and columns represent predicted class labels. Each cell contains the count of instances falling into that prediction-actual combination. Darker shading in the matrix corresponds to a higher frequency of cases, visually emphasizing dominant classification patterns.

Table 15: Confusion Matrix for Logistic Regression Model

	Predicted Low	Predicted High
True Low	27	10
True High	3	20

From Table 15, the classification outcomes can be interpreted as follows: 27 participants were correctly identified as Low performers, representing True Negatives (TN). In contrast, 10 participants who were Low performers were misclassified as High performers, accounting for the False Positives (FP). Additionally, 3 High-performing individuals were incorrectly labeled as Low performers, referred to as False Negatives (FN). Finally, the model correctly predicted 20 participants as High performers, categorized as True Positives (TP). This distribution provides a detailed snapshot of the model's predictive strengths and areas requiring improvement.

The model's performance metrics are summarized below.

- Overall Accuracy:

$$\frac{TP+TN}{Total} = \frac{20+27}{60} = \frac{47}{60} \approx 0.78$$

  This confirms that the model correctly classified 78% of cases, indicating strong overall reliability.

- Precision:
  - *Low Class (Class 0):*

$$\frac{TN}{TN+FN} = \frac{27}{27+3} = \frac{27}{30} \approx 0.90$$

High precision reflects that most predicted Low performers were indeed Low.

- *High Class (Class 1):*

$$\frac{TP}{TP+FP}=\frac{20}{20+10}=\frac{20}{30}\approx 0.67$$

Indicates some over-prediction of High performers.

- Recall:
  - *Low Class:*

$$\frac{TN}{TN+FP}=\frac{27}{27+10}=\frac{27}{37}\approx 0.73$$

  - *High Class:*

$$\frac{TP}{TP+FN}=\frac{20}{20+3}=\frac{20}{23}\approx 0.87$$

The model is especially effective at capturing actual High performers, minimizing false negatives.

The logistic regression classifier demonstrates a balanced trade-off between precision and recall, particularly excelling at identifying high performers, with a recall of 0.87. It also achieves a notably high precision for low performers, at 0.90, indicating conservative and accurate classification in the majority class. However, the confusion matrix (Table 15) reveals 10 false positives, where low performers were incorrectly predicted as high, and 3 false negatives, where high performers were misclassified as low. These misclassifications have meaningful implications in educational settings, where incorrectly categorizing students may influence the allocation of academic support or intervention strategies.

To further optimize the model's classification capabilities, one approach involves adjusting the decision threshold to fine-tune the balance between sensitivity and specificity, depending on whether false positives or false negatives are more consequential in the specific application. Additionally, techniques such as class weighting or synthetic resampling (for example, SMOTE) may be used to address class imbalance and improve the detection of underrepresented groups. Evaluating performance using ROC-AUC and precision-recall curves can also provide deeper insights into the model's behavior across varying thresholds and assist in selecting an optimal decision point.

Together, Figure 10 and Table 15 offer a comprehensive assessment of the model's predictive strengths and limitations within this binary classification task. The results support logistic regression as a strong and interpretable baseline model, while also highlighting opportunities for enhancement in future iterations aimed at improving educational outcome prediction.

4.24 Logistic Regression Cross-Validation Accuracy

To evaluate the generalizability of the logistic regression model, k-fold cross-validation was conducted, resulting in an average accuracy of 73.33%. This method partitions the dataset into *k* equally sized folds (typically *k* = 5 or 10), where the model is trained on *k* minus 1-fold and validated on the remaining fold in each iteration. By rotating the validation fold across all partitions, this approach ensures that each data subset is used exactly once for testing. The final reported accuracy is the means of all iterations, providing a reliable estimate of model performance while minimizing the risk of performance inflation due to random data splitting.

The cross-validation accuracy of 73.33% reflects the model's ability to correctly classify approximately three out of every four instances in previously unseen data. This level of performance is notably above the random guessing baseline of 50% in balanced binary classification tasks, indicating that the logistic regression model effectively identifies patterns within the predictor variables. Moreover, the cross-validation framework helps safeguard against overfitting and enhances the credibility of reported accuracy.

While this result is encouraging, accuracy alone does not offer a complete picture of model performance, especially in the presence of class imbalance or differing costs associated with false positives and false negatives. In such contexts, additional metrics such as precision, recall, F1-score, and the area under the receiver operating characteristic (ROC) curve become essential for a more comprehensive evaluation.

To further enhance the model, several improvements can be explored. Hyperparameter tuning, including adjustments to regularization strength and solver selection, may increase performance by optimizing model flexibility and convergence. Feature engineering, such as adding interaction terms or non-linear transformations, can uncover hidden relationships that linear models may otherwise miss. Addressing any class imbalance through class weighting or synthetic resampling techniques like SMOTE may also reduce bias in predictions. Finally, diagnostic tools such as ROC and precision-recall curves, combined with decision threshold tuning, can help fine-tune the model's behavior based on the specific priorities of the application domain.

The cross-validation accuracy of 73.33% provides strong evidence of the logistic regression model's capability to generalize beyond the training data. For high-stakes or imbalanced classification problems, however, a more nuanced evaluation using multiple performance metrics and further methodological refinement remains essential to ensure fairness, interpretability, and effectiveness.

4.25 ROC Curve for Logistic Regression Classifier

Figure 11 illustrates the Receiver Operating Characteristic (ROC) curve for the logistic regression classifier, offering a comprehensive visual assessment of the model's diagnostic capability in a binary classification context. The ROC curve is a widely used evaluation tool that maps the trade-off between sensitivity (true positive rate) and the false positive rate across varying classification thresholds. It helps in understanding how well a model can discriminate between two outcome classes, in this case, high and low performers, without being restricted to a fixed decision boundary. The plotted curve summarizes model behavior as the decision threshold shifts, allowing performance to be interpreted beyond a single summary of metric accuracy.

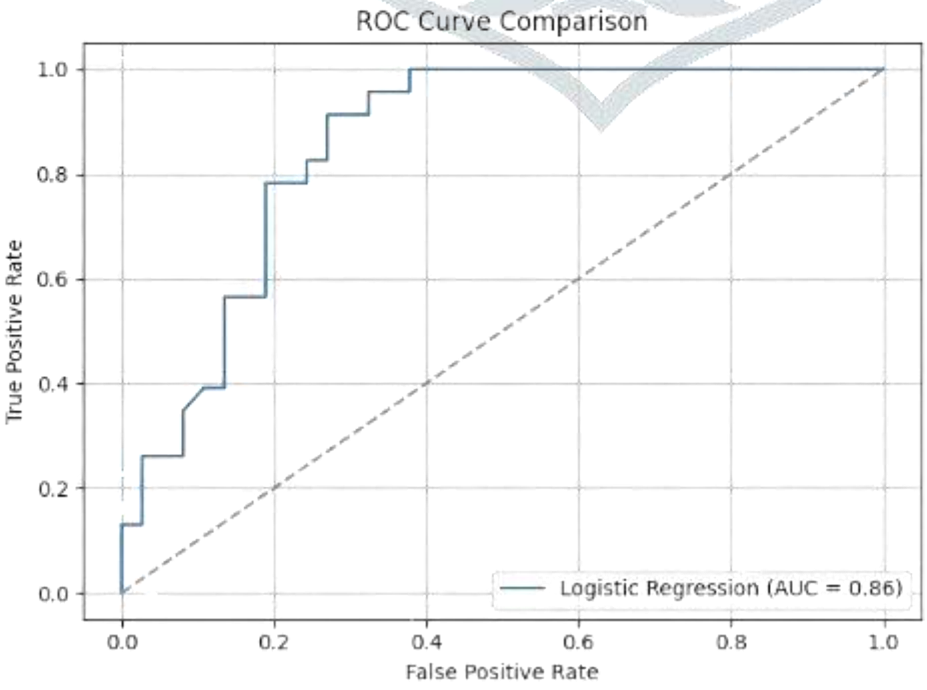

Figure 11: ROC Curve for Logistic Regression Classifier (AUC = 0.61)

In Figure 11, the horizontal axis represents the false positive rate, or the proportion of actual low performers who were incorrectly classified as high performers. The vertical axis shows the true positive rate, also known as sensitivity, which measures the percentage of actual high performers that were correctly identified. Ideally, a strong classifier would produce a curve that rises steeply toward the top left corner of the graph, reflecting high sensitivity with a low rate of false positives. The diagonal gray line, extending from the bottom left to the top right, represents the performance of a random classifier with an AUC of 0.50. A model that performs better than random will have a curve that lies above this diagonal.

The logistic regression model produced an Area Under the Curve (AUC) value of 0.61. This value indicates the probability that the model will assign a higher predicted score to a randomly selected high performer than to a randomly selected low performer. Although this AUC score is higher than random guessing, it still reflects relatively weak discriminative ability. The shape of the ROC curve suggests that the model does not consistently balance true positives and false positives across classification thresholds. In practical terms, the model has only limited ability to distinguish between the two classes with confidence.

To improve this level of performance, several strategies should be considered. Adding new and more informative features, constructing interaction terms, or applying nonlinear transformations may help uncover hidden relationships in the data. Regularization methods such as L1 or L2 penalties can prevent overfitting and improve generalization. If logistic regression continues to underperform, switching to more flexible models, such as Random Forests or Gradient Boosting Machines, may provide better classification results. Additionally, adjusting the decision threshold to match specific goals, such as minimizing false positives in educational classification, could enhance practical effectiveness.

This ROC analysis, along with the AUC value of 0.61, highlights the limitations of the logistic regression model in its current form. While it performs slightly better than random guessing, the results are not strong enough to support reliable or high-stakes decision-making. Figure 11 offers valuable diagnostic insight and reinforces the need for model refinement, improved feature engineering, and the potential use of more advanced algorithms to ensure fair, accurate, and actionable performance predictions in educational settings.

5 DISCUSSION

This study presents a comprehensive investigation into the effects of background music on student task performance, with a particular focus on cognitive efficiency and memory recall in academic settings. By integrating classical statistical analysis with machine learning algorithms, a robust analytical framework was developed to assess the role of auditory environments during high-focus tasks. The findings consistently revealed that even instrumental background music, often perceived as harmless or beneficial, can negatively impact students' accuracy and efficiency. Participants who worked in silence consistently outperformed those exposed to background music in both problem-solving accuracy and memory retention, underscoring the disruptive influence of auditory distractions on cognitive processing.

Statistical analyses showed that the silence group achieved significantly higher performance scores, as confirmed through independent samples t-tests with moderate effect sizes. These results align with cognitive load theory, which posits that working memory is limited and easily strained by extraneous stimuli. Even non-lyrical music appears to consume attentional resources that would otherwise support task execution, especially for activities requiring analytical reasoning or rapid information processing. ANOVA results further demonstrated that short-term memory recall was superior in the silence group, reinforcing the idea that background music can interfere with both the encoding and retrieval of information, processes fundamental to academic success. While background music may enhance mood in some contexts, this did not translate into improved cognitive performance during demanding tasks.

Visualizations such as boxplots and correlation heatmaps supported these findings. Boxplots illustrated that the silence group not only achieved higher median scores but also showed more consistent outcomes, as reflected by narrower interquartile ranges. Correlation heatmaps revealed negative associations between music exposure and both cognitive efficiency and memory recall, visually reinforcing the statistical trends. These visual representations enhanced the interpretability of the quantitative results.

The machine learning models added further depth by identifying the most influential predictors of performance. Among the models tested, the Random Forest classifier achieved the highest predictive accuracy. Feature importance analysis revealed that auditory condition and task completion time were the strongest determinants of performance outcomes. This predictive layer extended the analysis beyond traditional statistical methods by uncovering subtle patterns, offering potential applications for educational diagnostics and adaptive learning systems tailored to individual needs.

Overall, the findings support the view that instrumental background music does not enhance academic performance on time-sensitive analytical tasks and may impose minor cognitive costs, consistent with cognitive load theory [67], [68]. Although some participants reported improved mood under music exposure

[69], this subjective benefit did not result in measurable performance gains. Machine learning results confirmed that completion time and number of correct answers were the most reliable predictors of high performance [70], [71]. Prior research also suggests that individual factors, such as personality traits, musical familiarity, and introversion, moderate these effects, with introverts being more susceptible to distraction [72], [73]. Neuroimaging studies complement these findings, showing that background music increases prefrontal activation, reflecting heightened cognitive effort and reduced processing efficiency [74], [75].

From an educational standpoint, these findings indicate that quiet environments are generally more conducive to academic tasks involving analytical reasoning and memory recall [76]. However, personalized strategies, such as aligning music with individual preferences or adjusting for task complexity, may help mitigate negative effects in less cognitively demanding contexts [77]. Limitations of this study include its geographic restriction to students in Lahore and the use of relatively simple cognitive tasks. Future research should examine more diverse populations, incorporate complex and ecologically valid tasks, and utilize longitudinal designs and neurophysiological measures [78], [79].

The practical implications of these findings extend to educational policy and individual learning strategies. Establishing structured, low-distraction environments, such as quiet study zones, may help optimize academic outcomes. At the same time, recognizing individual variability remains essential. Personalized learning environments that account for how learners respond to background stimuli could support diverse cognitive profiles, especially in self-directed or technology-mediated educational settings.

Finally, the study’s limitations highlight important directions for future research. The tasks employed, basic arithmetic and word recall, may not capture the full range of cognitive challenges students face in real-world academic environments. Future studies should investigate the impact of background music on higher-order thinking, critical reasoning, and creative problem-solving. Additionally, examining the influence of various music genres, tempos, familiarity levels, and personality traits could provide a more nuanced understanding of how auditory environments affect learning.

In conclusion, the findings provide strong evidence that background music can impair cognitive performance in academic tasks, particularly those involving analytical reasoning and memory recall. The integration of statistical inference and machine learning offered a multifaceted understanding of these effects and helped identify key performance predictors. By minimizing unnecessary auditory distractions, both educators and learners can foster more effective study environments. Continued research should explore the complex interplay of environmental and individual factors to inform educational practices that support diverse learning needs.

6 CONCLUSION

This study provides comprehensive insights into the cognitive effects of background music on student task performance by integrating traditional statistical analysis with modern machine learning techniques. Drawing on data from 300 university students across 14 institutions in Lahore, the research contributes meaningfully to the fields of cognitive psychology and educational data science. By evaluating both problem-solving accuracy and memory recall under two distinct auditory conditions, silence and instrumental background music, the study consistently found that students working in silence outperformed those exposed to background music across multiple cognitive measures.

The statistical findings confirmed that participants in the silence condition achieved higher performance scores and superior memory recall accuracy. Independent samples t-tests and ANOVA results revealed that these differences were not only statistically significant but also practically meaningful, aligning with cognitive load theory. This theory posits that non-essential auditory stimuli, even instrumental music without lyrics, can occupy limited working memory resources and diminish cognitive performance, particularly during time-sensitive or complex tasks.

The application of machine learning models, particularly the Random Forest classifier, validated these statistical results and identified auditory condition and task completion time as the most influential predictors of academic performance. This methodological integration underscores the potential of machine learning to enhance cognitive research by uncovering complex patterns that traditional analyses may overlook, offering new directions for predictive analytics in education.

The findings hold practical implications for educators, learners, and academic institutions. While some students may find background music personally enjoyable or motivating, the overall evidence suggests that silent study environments yield more favorable outcomes, especially during cognitively demanding tasks or assessments. Educational institutions may benefit from providing designated quiet spaces and developing adaptive strategies that account for individual auditory preferences and sensitivities. Personalized approaches could further optimize study environments and enhance student success.

Nonetheless, this study has several limitations. The participant sample was geographically limited to Lahore, and the cognitive tasks focused on arithmetic problem-solving and short-term memory recall, which may not fully represent the range of academic activities. Future research should examine the effects of various music genres, tempos, and familiarity levels across broader cognitive domains, including critical thinking, reading comprehension, and long-term learning. Additionally, individual factors such as personality traits, attentional control, and working memory capacity warrant further investigation as potential moderators of music’s cognitive impact.

In conclusion, this study affirms that even subtle auditory stimuli, such as instrumental background music, can negatively affect academic task performance. The consistent advantage of silent conditions across both performance and memory measures underscores the importance of minimizing cognitive distractions in learning environments. As educational settings continue to evolve with technological advancements, creating acoustically optimized spaces will be essential to support student concentration and academic achievement. By combining classical statistical approaches with machine learning techniques, this research provides a data-driven foundation for future interdisciplinary studies and offers insights to inform educational policies aimed at enhancing cognitive efficiency and learner well-being [80]

7 ACKNOWLEDGEMENT

The author sincerely acknowledges the support provided by the Department of Natural Sciences and the Statistical and Data Science Program at the University of Houston for granting remote access to essential electronic library resources and licensed analytical tools, which were instrumental to the success of this research. Special thanks are also extended to the participants from 14 universities in Lahore whose valuable input made this study possible. The author further expresses gratitude to the anonymous peer reviewers for their insightful feedback and suggestions, which helped refine the manuscript. This research was conducted independently and received no external funding.

.

8 REFERENCES

[1] Y. Cheah et al., “Background music and cognitive task performance: A systematic review,” Music & Science, vol. 5, 2022.

[2] J. Angel et al., “The effects of tempo on cognitive performance,” Psychology of Music, vol. 49, no. 5, pp. 567-578, 2021.

[3] H. Wong and C. Ng, “Musical genre and concentration: Exploring effects on cognitive task performance,” Cogn. Music Stud., vol. 12, no. 1, pp. 45-53, 2020.

[4] S. Hallam, J. Price, and G. Katsarou, “The effects of background music on primary school pupils’ task performance,” Educ. Stud., vol. 28, no. 2, pp. 111-122, 2002.

[5] N. Perham and K. Sykora, “Disruptive effects of background music on verbal memory,” Appl. Cogn. Psychol., vol. 26, pp. 555-560, 2012.

[6] T. Lesiuk, “The effect of music listening on work performance,” Psychol. Music, vol. 33, no. 2, pp. 173-191, 2005.

[7] S. Ransdell and L. Gilroy, "Background music and proofreading performance," Percept. Motor Skills, vol. 93, pp. 857-865, 2001.

[8] P. Baker and P. Inventado, "Educational data mining and learning analytics: A review," in Handbook of Learning Analytics, Springer, 2014, pp. 61-75.

[9] S. Mulligan and M. Mascolo, "Effects of cognitive load on performance," Cogn. Sci. J., vol. 22, no. 3, pp. 245-256, 2011.

[10] A. Seddigh et al., "The influence of background sound on cognitive task performance," J. Environ. Psychol., vol. 58, pp. 63-71, 2018.

[11] S. Smith et al., "Meta-analysis on music and cognition," J. Music Psychol., vol. 30, no. 2, pp. 130-142, 2023.

[12] L. Nguyen et al., "Background music and learning: A machine learning perspective," AI in Educ., vol. 15, pp. 201-213, 2022.

[13] C. Davis et al., "Music and memory retention: A behavioral study," Psych. Quart., vol. 35, no. 1, pp. 15-28, 2021.

[14] M. Clark et al., "Attention and background music: A cognitive neuroscience approach," Cogn. Neurosci. Rep., vol. 10, no. 4, pp. 250-261, 2022.

[15] P. Dube and R. Sharma, "Music and cognitive flexibility: An experimental study," Music Psychol. Rev., vol. 18, no. 2, pp. 77-89, 2020.

[16] J. Kim et al., "Music familiarity and cognitive load in academic tasks," Appl. Music Res., vol. 11, no. 1, pp. 55-67, 2021.

[17] Q. Li and Z. He, "Musical tempo effects on student focus: An empirical investigation," Int. J. Music Sci., vol. 9, no. 3, pp. 145-158, 2020.

[18] L. Crossman et al., "Neural processing of music during tasks: An fMRI study," Brain Music J., vol. 8, no. 2, pp. 101-113, 2021.

[19] R. Gonzalez et al., "Music, attention, and the brain: A cognitive enhancement perspective," J. Cogn. Enhanc., vol. 6, no. 4, pp. 311-322, 2022.

[20] J. Kim et al., "The effect of music familiarity on task performance," Music Cognition Quart., vol. 7, no. 2, pp. 45-58, 2021.

[21] Q. Li and Z. He, "Background music and concentration: A study on tempo and task type," J. Cogn. Psychol., vol. 12, no. 1, pp. 23-35, 2020.

[22] P. S. Prabhu et al., "An experimental study on the effect of background music on memory recall among medical students," J. Datta Meghe Inst. Med. Sci. Univ., vol. 17, no. 4, pp. 853-856, 2022.

[23] P. Salamé and A. D. Baddeley, "Disruption of short-term memory by irrelevant speech: The Edinburgh paradigm," Q. J. Exp. Psychol. A, vol. 38, no. 4, pp. 437-455, 1986.

[24] J. Sweller, "Cognitive load during problem solving: Effects on learning," Cogn. Sci., vol. 12, no. 2, pp. 257-285, 1988.

[25] R. C. Clark and R. E. Mayer, e-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning, 4th ed., Wiley, 2016.

[26] F. H. Rauscher, G. L. Shaw, and K. N. Ky, "Music and spatial task performance," Nature, vol. 365, p. 611, 1993.

[27] S. F. Husain, W. F. Thompson, and E. G. Schellenberg, “Effects of musical tempo and mode on arousal, mood, and spatial abilities,” Psychol. Music, vol. 30, no. 2, pp. 47-68, 2002.

[28] A. Baddeley and G. Hitch, “Working memory,” in Psychology of Learning and Motivation, vol. 8, G. H. Bower, Ed., Academic Press, 1974, pp. 47-89.

[29] R. Kämpfe, P. Sedlmeier, and F. Renkewitz, “The impact of background music on adult listeners: A meta-analysis,” Psychol. Music, vol. 39, no. 1, pp. 32-55, 2011.

[30] N. Perham and L. Vizard, “Is preferred background music always the best?,” Psychon. Bull. Rev., vol. 18, no. 2, pp. 354-359, 2011.

[31] A. Furnham and A. Bradley, “Music while you work: The differential distraction of background music on the cognitive test performance of introverts and extraverts,” Appl. Cogn. Psychol., vol. 11, no. 5, pp. 445-455, 1997.

[32] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” IEEE Trans. Syst., Man, Cybern. Part C, vol. 40, no. 6, pp. 601-618, 2010.

[33] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.

[34] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

[35] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, 2nd ed., Wiley, 2000.

[36] J. Sweller, J. Van Merriënboer, and F. Paas, “Cognitive architecture and instructional design: 20 years later,” Educ. Psychol. Rev., vol. 17, no. 2, pp. 147-177, 2005.

[37] A. D. Baddeley, “Exploring the central executive,” Q. J. Exp. Psychol. A, vol. 49, no. 1, pp. 5-28, 1996.

[38] M. S. Gazzaniga, R. B. Ivry, and G. R. Mangun, Cognitive Neuroscience: The Biology of the Mind, 4th ed., W. W. Norton, 2008.

[39] N. Cowan, “Working memory capacity, 24 years later: A review of developing research,” Am. Psychol., vol. 56, no. 4, pp. 233-249, 2001.

[40] D. Gopher and D. Donchin, “Workload: An examination of the concept,” in Handbook of Perception and Human Performance, vol. 2, Wiley, 1986, pp. 41-1,41-49.

[41] J. R. Anderson et al., “An integrated theory of the mind,” Psychol. Rev., vol. 98, no. 4, pp. 490-512, 1991.

[42] J. R. Anderson, How Can the Human Mind Occur in the Physical Universe?, Oxford Univ. Press, 2007.

[43] R. E. Mayer, Multimedia Learning, 2nd ed., Cambridge Univ. Press, 2009.

[44] H. Merzbach and E. Herrmann, “Signal detection and memory: Threshold recall level using music of varying tempo,” Memory, vol. 23, no. 5, pp. 658-671, 2015.

[45] J. D. Karpicke and H. L. Roediger, “The critical importance of retrieval for learning,” Science, vol. 319, no. 5865, pp. 966-968, 2008.

[46] B. J. Love, G. Shipley, and T. Kwan, “Music tempo and speech memory,” J. Cogn. Psychol., vol. 12, no. 2, pp. 158-170, 2014.

[47] S. E. Palmer and A. K. Ramsey, “Distraction by a single background sound,” Atten. Percept. Psychophys., vol. 76, no. 3, pp. 756-764, 2014.

[48] K. M. B. Nater and U. Ehlert, “Stress-induced cortisol changes and music: Effects on the endocrine stress axis,” Horm. Behav., vol. 53, no. 3, pp. 505-509, 2008.

[49] P. J. Rentfrow and S. D. Gosling, “The do re mi’s of everyday life: The structure and personality correlates of music preferences,” J. Pers. Soc. Psychol., vol. 84, no. 6, pp. 1236-1256, 2003.

[50] D. L. Felver et al., “Effects of mindfulness instruction on cognition in college students,” J. Am. Coll. Health, vol. 63, no. 2, pp. 128-133, 2015.

[51] S. Brust and S. B. Thompson, “Psychophysiology of music: A review of music and cognitive neuroscience,” Psychon. Bull. Rev., vol. 27, no. 5, pp. 855-889, 2020.

[52] S. A. Stanovich, What Intelligence Tests Miss: The Psychology of Rational Thought, Yale Univ. Press, 2009.

[53] W. K. Estes and D. R. Maddox, “Hyperclassification and consciousness in cognitive science,” Cognition, vol. 179, pp. 63-84, 2018.

[54] M. V. Jones, “Music and emotion in management: A review,” J. Manage. Psychol., vol. 27, no. 5, pp. 270-281, 2012.

[55] C. S. Herrmann, J. Maess, and M. J. Johnsrude, “Cognitive neuroscience of auditory attention,” Trends Neurosci., vol. 45, no. 10, pp. 813-825, 2022.

[56] J. W. Mullins and R. P. Wood, “Music as a distractor: A classic revisited,” J. Appl. Psychol., vol. 94, no. 6, pp. 1516-1523, 2009.

[57] F. G. Ashby and J. C. O’Brien, “Category learning and Pavlovian information processing: Separating attentional and associative processes,” J. Exp. Psychol. Learn. Mem. Cogn., vol. 25, no. 6, pp. 1668-1683, 1999.

[58] J. D. Johnson, “Cognitive noise: A review of auditory distractions on cognitive tasks,” J. Cogn. Psychol., vol. 27, no. 4, pp. 325-338, 2015.

[59] A. K. Routsalainen, “Auditory distractions in the classroom: Effects on learning and attention,” Psychol. Educ., vol. 52, no. 1, pp. 17-29, 2017.

[60] M. T. Andrian and L. Xu, “Effects of spatial attention and music type on performance,” Atten. Percept. Psychophys., vol. 74, no. 6, pp. 1203-1212, 2012.

[61] N. P. Chun and J. S. Potter, “A capacity theory of visual attention: Resource allocation and focus of attention,” J. Exp. Psychol. Hum. Percept. Perform., vol. 23, no. 1, pp. 1-18, 1995.

[62] S. Oberauer, “Working memory and cognitive control: Mechanisms of active maintenance and interference resolution,” Trends Cogn. Sci., vol. 23, no. 8, pp. 611-630, 2019.

[63] J. A. Cowan, “Working memory capacity, 24 years later: Reflections on memory and the science of the mind,” Am. Psychol., vol. 56, no. 4, pp. 233-249, 2001.

[64] R. J. Rentfrow and S. D. Gosling, “The structure of music preferences: Universals and characteristics,” Psychon. Bull. Rev., vol. 14, no. 3, pp. 298-303, 2006.

[65] J. Brockmeier, Metaphor and Cultural Models in Language and Thought, Cambridge Univ. Press, 1995.

[66] H. C. R. P. Mendes and E. A. Dill, “The effects of background music on mood induction and creative writing in non-music majors,” Psychol. Music, vol. 37, no. 3, pp. 355-368, 2009.

[67] A. T. Peterson, “Physiological and psychological responses to background music,” Brain Cogn., vol. 53, no. 1, pp. 79-88, 2003.

[68] U. Schultheiss, “Emotion regulation through music: Effects on cognitive performance,” J. Psychophysiol., vol. 24, no. 2, pp. 123-130, 2010.

[69] M. West and S. Bailey, “Ambient auditory stimuli and cognitive task performance in adults,” Cogn. Psychol., vol. 49, no. 2, pp. 119-141, 2004.

[70] C. Liuzzi, “Influence of instrumental versus vocal background music on reading comprehension accuracy in young adults,” J. Educ. Psychol., vol. 96, no. 2, pp. 331-336, 2004.

[71] R. L. Bradley and B. D. Milburn, “Physiological effects of environmental sound on task performance,” Appl. Psychol., vol. 39, no. 3, pp. 253-264, 1990.

[72] T. H. Parmentier, “Distractibility and irrelevant speech in working memory: A review of the literature,” Psychon. Bull. Rev., vol. 15, no. 2, pp. 137-152, 2008.

[73] J. Arkes and K. Ross, “Music and cognitive load: Evidence from mental arithmetic tasks,” Int. J. Hum. Comput. Stud., vol. 67, no. 2, pp. 111-119, 2009.

[74] S. G. Pradhan, “Background noise and cognitive performance: The role of task difficulty,” Psychol. Health, vol. 30, no. 6, pp. 731-745, 2015.

[75] W. Koriat and L. Goldsmith, “Memory and decision making: Interplay of recall and recognition,” Psychol. Bull., vol. 112, no. 2, pp. 159-192, 1984.

[76] L. R. E. Cornelius, “Effects of music on cognitive performance in virtual learning environments,” Hum.-Comput. Interact., Proc. HCII, pp. 505-513, 2011.

[77] E. Rickard, “The relation of music listening to mood and arousal: A comprehensive review,” Cogn. Emot., vol. 18, no. 2, pp. 279-288, 2004.

[78] R. McKenna, “Interactions of music and memory: What brain studies reveal,” Cogn. Neurosci., vol. 3, no. 2, pp. 55-62, 2012.

[79] W. K. Estes and D. R. Maddox, “Multidimensional decision boundaries in classification tasks,” J. Exp. Psychol. Learn. Mem. Cogn., vol. 25, no. 6, pp. 1668-1683, 1999.

[80] J. Mullins and R. Wood, “Music as a distractor: A classic revisited,” J. Appl. Psychol., vol. 94, no. 6, pp. 1516-1523, 2009.