ijamrsd.com JAMRSD ISSN: 2583-6404
Journal of Advanced Multidisciplinary Research Studies and Development
Peer Reviewed Referred Journal Bi-monthly Open Access

FREQUENCY-DOMAIN EEG-BASED CLASSIFICATION OF EDUCATIONAL TASK DIFFICULTY USING SUBJECT-INDEPENDENT MACHINE LEARNING

Kumud H. Salunke[1], Dhruv Sharma[2], Siddhi Choudhary[3]
[1,2,3] Bachelor of Computer Applications (Honours), Department of Computer Science and Applications, MIT World Peace University, Pune, India.

Abstract
This study investigates whether frequency-domain electroencephalography (EEG) features can offer useful information for forecasting the degree of difficulty of educational activities. In addition to classifying tasks as easy, moderate, or difficult, the objective is to evaluate how well the model generalizes to participants that were not included in the training procedure. Eight EEG channels, five spectral bands, three balanced difficulty levels, four types of educational tasks, and 48,000 records from 200 participants made up the synthetic dataset used in the research. Global spectral features and relative band-power measurements for every channel were among the 47 frequency-domain features that were examined. We used 140 participants for training and set aside 60 entirely new subjects for testing in a 70/30 split that was subject-independent. Selection and standardization of features. The purpose of this study is to determine whether frequency-domain electroencephalography (EEG) features can provide valuable information for predicting the level of difficulty of educational activities. The goal is to assess how effectively the model generalizes to individuals who were not part of the training process, in addition to categorizing tasks as simple, moderate, or tough. The synthetic dataset used in the study included eight EEG channels, five spectral bands, three balanced difficulty levels, four different kinds of educational tasks, and 48,000 records from 200 subjects. 47 frequency-domain features were analyzed, including global spectral features and relative band-power measurements for each channel. In a 70/30 split that was subject-independent, we employed 140 participants for training and reserved 60 completely new subjects for testing.

Keywords: Electroencephalography, Frequency-Domain Features, Educational Task Difficulty, Cognitive Workload, Subject-Independent Classification, Machine Learning.

► *Corresponding Author: Kumud H. Salunke*

I. Introduction

There are several levels of cognitive demand associated with educational activities. Different levels of attention, working memory, executive control, and information-processing resources are needed for tasks including reading, math, problem-solving, and multimedia learning. Determining how challenging a task is for a learner is essential in adaptive educational systems. A task that is too difficult can cause overburden and impede effective learning, while a task that is too easy may not push them sufficiently. Traditionally, learner performance, completion times, surveys, self-reports, and predetermined task parameters are used to estimate difficulty.

Although these approaches are helpful, they frequently fall short of continuously tracking the learner's physiological state while they are working on the task. Electroencephalography (EEG), which measures the electrical activity of the brain with great temporal resolution, provides a unique viewpoint in this situation. Because job demands can cause variations in spectral activity throughout a variety of ranges, including delta, theta, alpha, beta, and gamma, frequency-domain EEG analysis is particularly pertinent for researching cognitive stress. Theta, alpha, and beta spectral measurements have predictable connections with cognitive stress, however the effects can vary depending on the activity and circumstance, according to a meta-analysis by Chikhi et al. This result lends credence to the notion of modeling computational burden and complexity using spectral features as a clear representation.

EEG-based machine learning research has benefited from the growth of benchmark workload datasets. Simplified spectral representations can successfully differentiate between various workload levels, as demonstrated by Lim et al.'s introduction of the STEW dataset. Subsequent studies have explored a broader range of feature representations and classifiers. Although their methodology mostly concentrated on individual-specific modeling, Gupta et al. investigated functional-connectivity representations and deep learning approaches for identifying cognitive burden [3]. Intriguing task-independent research by Zhang et al. demonstrates how spectral and connectivity patterns can significantly aid in generalizing across various working-memory tasks. This emphasizes the necessity of evaluation approaches that consider more than just within-subject predictions.

[4]. Despite our progress, it is still challenging to establish a methodological connection between the difficulty of educational assignments and EEG-based workload estimation. When trained and evaluated on samples from the same subjects, a model may perform exceptionally well; nevertheless, when presented with different individuals, it may falter. Overly optimistic estimates may result from each subject's distinct EEG properties, particularly when random sample splitting permits data from the same participant to appear in both training and testing sets. Clear preprocessing, minimizing data leaking, and employing evaluation techniques that are consistent with the desired generalization claims have all been emphasized in recent talks over reproducibility in EEG workload research.

[5]. Therefore, a good educational application should genuinely test whether the spectrum knowledge acquired by one group of students can be applied to whole new students. The diversity of representation is another challenge. A combination of raw signals, power spectral density, relative band power, entropy, connection, multimodal physiological characteristics, and deep-learned representations have all been employed in workload studies.

While hybrid EEG–fNIRS methods can boost performance by bringing in complementary signals, they also add layers of complexity in sensing and processing [6]. Connectivity-based methods have shown impressive results, like the STEW-based framework from Safari et al., but they tend to be more complicated in terms of feature construction compared to straightforward spectral representations [7]. This leads us to an important practical question: how much predictive power can we really get from a compact, reproducible frequency-domain EEG representation on its own? This study tackles the question of classifying educational task difficulty. Instead of mixing in task identity, participant identity, behavioural outcomes, or metadata with EEG features, the proposed approach focuses solely on frequency-domain EEG characteristics for predictions. The aim is to categorise tasks into three difficulty levels: easy, moderate, or hard. While task type is included for dataset description and analysis, it's intentionally left out of the predictive feature set. This setup allows us to directly assess whether the spectral EEG data holds valuable discriminative information without depending on explicit task labels. The research utilises 47 frequency-domain features from eight EEG channels across five spectral bands, along with global spectral measures and two spectral ratios. The channel-level data includes relative power features for F3, F4, C3, C4, P3, P4, O1, and O2 in the delta, theta, alpha, beta, and gamma bands. To enhance the channel-wise representation, global relative-power measures and theta/alpha and beta/alpha ratios are also included. Feature selection is conducted solely on the training set by reducing the feature count from 47 down to 40. Standardisation is applied only to the training data as well. These precautions are crucial because performing selection or scaling before separating the train and test data could inadvertently leak information from the test subjects into the training phase. A rigorous subject-independent evaluation design is then implemented, splitting the 200 subjects into 140 for training and 60 for final testing, ensuring no overlap. This results in 33,600 samples for training and 14,400 for testing, with the final test set consisting of subjects who were completely unseen during training. Logistic Regression is chosen as the main model due to its clear linear baseline and the ability to interpret the coefficients of the selected spectral features. Additionally, Random Forest and Decision Tree classifiers are employed for lightweight comparisons under the same input representation and test set.

This research tackles a threefold gap. First, studies on EEG workload often vary in how they define features and set evaluation protocols, which complicates the task of determining the effectiveness of a consistent frequency-domain representation. Second, many of the top-performing studies depend on complex, subject-specific, within-task, multimodal, or connectivity-based methods, leaving the question of whether a simple spectral representation can effectively differentiate educational difficulty across new subjects unanswered. Third, moderate difficulty can physiologically blend with adjacent levels, yet the focus tends to be more on overall accuracy rather than the specific error patterns for each class. To address these issues, the proposed workflow integrates a fixed spectral feature space, controls for leakage in subject-independent testing, benchmarks lightweight models, conducts error analysis, and interprets features at a granular level.

What sets this work apart is its focus on weaving these methodological components around a clearly defined research question, rather than claiming to discover a new physiological mechanism. The pipeline assesses a frequency-domain-only representation for classifying educational task difficulty into three levels, deliberately omitting task and participant metadata from the model input, and tests the final model on subjects who have never been seen before. It also provides insights into model comparisons, class-specific errors, common confusion pairs, and the significance of different frequency bands. This approach clarifies and makes reproducible the analytical journey from spectral representation to predicting difficulty.

The key contributions of this work include:

1) The introduction of a leakage-controlled machine-learning pipeline that classifies educational task difficulty as easy, moderate, or hard, using only frequency-domain EEG characteristics as predictive input.

2) The execution of a rigorous 70/30 subject-independent evaluation, where 140 subjects were used for training, and 60 completely new subjects were set aside for final testing, ensuring no overlap between the two groups.

3) We reduced the original 47-feature spectral representation down to 40 EEG features chosen for training. After that, we compared Logistic Regression with lighter options like Random Forest and Decision Tree, all while keeping the evaluation conditions consistent.

4) Using Logistic Regression, we achieved a subject-independent accuracy of 75.69% and an F1-score of 75.71%. This performance exceeded the three-class chance level by an impressive 42.35 percentage points and outshone the other models we tested.
5) We performed class-level error analysis and coefficient-based feature evaluations, which pointed out that moderate difficulty was the most challenging class. We also highlighted the importance of global theta/alpha and beta/alpha ratios, along with spectral characteristics tied to alpha, theta, and beta.

II. Related Work

Zhou et al. examined machine learning and deep learning methods for identifying cognitive workload using EEG in greater detail in 2021. Numerous signal representations, preprocessing techniques, and validation procedures were discovered throughout their evaluation. They highlighted the growing promise of learning-based methods, but they also highlighted the difficulties in comparing outcomes across various experimental configurations [8]. Gupta et al. investigated the classification of cognitive workload using deep learning and EEG functional connectivity in 2021 as well. Their results showed that neural models can utilize interactions across various channels and that connection can capture workload-relevant information. However, the emphasis on specific topics makes it more difficult to make generalizations about how difficult schooling will be for certain people. [3].
In 2021, Zhang et al. used spectral and functional-connectivity fingerprints across working-memory tasks to study task-independent mental workload differentiation. Impressive cross-task discrimination was obtained by their methodology, demonstrating that some neuronal signatures are not limited to a particular task. However, their method only addressed two workload levels and incorporated many feature families, whereas educational difficulty may require a three-way discrimination with a simpler spectral representation [4].
Back in 2021, researchers delved into subject-specific workload by using stochastic configuration networks to compare different classification strategies. They found that individual differences play a significant role in how we model EEG workload. This study highlighted the importance of thoughtful evaluation design, but it also pointed out that even with a focus on subject-specific modelling, we still need to ensure rigorous validation with unseen subjects [9]. In 2022, Chikhi, Matton, and Blanchet took a closer look at EEG power spectral measures related to cognitive workload through a meta-analysis. They discovered that theta waves had the strongest connection, while alpha and beta waves were influenced too, though their relationships varied depending on the context. This research provided solid physiological backing for spectral modelling, but it also noted the inconsistencies that can arise across different experimental setups [1]. Later in 2022, studies combining EEG and fNIRS explored workload levels by examining functional brain connectivity and using multimodal machine learning. The integration of these two methods enhanced the ability to distinguish between various workload levels in several scenarios. However, adding another sensing modality and calculating connectivity does complicate things compared to a straightforward frequency-domain EEG-only approach [6]. Fast forward to 2023, where cognitive-load detection studies employed signal decomposition and metaheuristic feature selection, achieving impressively high benchmark accuracies. After going through decomposition, extracting entropy features, optimising selections, and using nonlinear classifiers, these findings showcased the promise of advanced pipelines. However, the complexity of the methodology can pose challenges for reproducibility, increase computational costs, and make direct comparisons trickier [10].

In 2024, Safari and colleagues made significant strides by combining effective brain connectivity, feature selection, and machine learning techniques on the STEW dataset, achieving impressive results in workload classification. Their findings highlight that EEG data holds valuable information for distinguishing workload levels. However, their method leans on connectivity estimation and employs a sevenfold cross-validation approach, which differs from the frequency-domain-only evaluation of unseen subjects that we utilise in this study [7]. In the same year, research on multimodal operator-workload classification merged eye tracking with EEG to enhance multiclass workload recognition. The findings advocate for the use of multimodal sensing when aiming for high predictive performance, but they fall short of isolating the specific predictive power of spectral EEG features alone [11]. Additionally, in 2024, EEG workload research with a focus on reproducibility offered guidelines for creating transparent and reproducible machine-learning workflows. This is particularly important since reported performance can be skewed by undocumented preprocessing, data leakage, and inconsistent validation practices. Our current study addresses these issues by ensuring that selection and scaling are applied solely to training data and by maintaining zero subject overlap [5].

By 2025, a comprehensive review of machine-learning performance across various workload task types highlighted the variability in EEG-based workload results depending on the context of the tasks. It underscored that model performance is influenced by factors such as experimental design, task domain, feature representation, and validation strategy. This review emphasises the importance of making clearly defined claims rather than treating workload accuracy values as interchangeable across different datasets [12].

Overall, the literature reveals a trend moving from benchmark spectral features and traditional classifiers to more nuanced approaches like subject-specific deep learning, task-independent fingerprints, multimodal sensing, connectivity representations, and a strong emphasis on reproducibility in validation. The remaining methodological opportunity isn't just about increasing model complexity, but it is about assessing a clearly defined and interpretable EEG representation within a generalisation framework that aligns with the claims being made. Consequently, this study focuses on a compact set of frequency-domain features, categorising them into three levels of educational difficulty, and tests them on completely new subjects.

III. Research Methodology and Proposed System

A. System Architecture

The arrows in Fig. 1 represent the one-directional transformation of information from dataset records to spectral representation, training-only feature selection and scaling, machine-learning classification, unseen-subject evaluation, and final difficulty labels. Critically, no information flows backwards from the test set into feature selection, scaling, or model fitting.

B. Input, Processing, and Output

Table I. Input, Processing, and Output of the Proposed Pipeline

Stage	Definition
Input	47 frequency-domain EEG characteristics. Nonfeature columns include subject identity, task type, difficulty metadata, trial identifier, synthetic flag, and recommended split metadata.
Processing	Subject-level 70/30 split; training-only feature selection; training-only standardisation; classifier fitting; prediction on unseen subjects.

Output	Predicted educational task difficulty: easy, moderate, or hard, together with evaluation metrics and interpretability analyses.

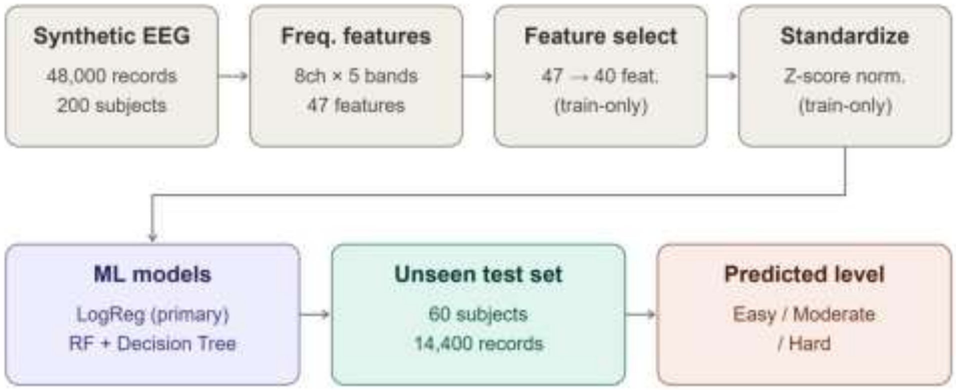

Fig. 1. Proposed subject-independent frequency-domain EEG classification pipeline.

C. Mathematical Formulation

Let's break down the EEG dataset, which we can represent as:

$D = \{(x_i, y_i, s_i)\}$ for $i = 1, 2, ..., N$.

Here, $x_i \in R^{47}$ stands for the frequency-domain EEG feature vector, $y_i \in$ {Easy, Moderate, Hard} indicates the difficulty level of the educational task, and s_i refers to the subject involved. This dataset comprises a total of $N = 48{,}000$ samples gathered from 200 different subjects.

To ensure that our evaluation is independent of the subjects, we divided the subject set S into distinct training and testing subsets:

$S\_train \cap S\_test = \emptyset$

In this setup, we have $|S\_train| = 140$ and $|S\_test| = 60$, which means that no subject in the training set overlaps with those in the testing set.

Feature selection was carried out using only the training data. Out of the original 47 frequency-domain EEG features, we narrowed it down to 40 features:

$x_i \in R^{47} \rightarrow x_i^J \in R^{40}$.

These selected features were then standardised based on the statistics from the training set:

$z_{ij} = (x_{ij} - \mu_j) / \sigma_j$

where μ_j and σ_j represent the mean and standard deviation of feature j, calculated solely from the training data.

We classified the standardised feature vector using multinomial Logistic Regression. For K = 3 difficulty classes, the probability of class k is computed as:

$P(y = k \mid z) = \exp(w_k^T z + b_k) / \Sigma_{r=1}^{K} \exp(w_r^T z + b_r)$.

The predicted educational task difficulty is determined by: $\hat{y} = \arg\max_k P(y = k \mid z)$

To evaluate the model, we looked at accuracy, precision, recall, F1-score, balanced accuracy, and a confusion matrix. For three balanced classes, the chance-level accuracy is:

1/3 = 33.33%

The framework we proposed achieved a solid subject-independent test accuracy of 75.69% and an F1-score of 75.71%. This indicates that the frequency-domain EEG characteristics are rich in predictive information, helping to differentiate between easy, moderate, and hard educational task difficulties.

D. Algorithm

1) Begin by loading the dataset and identifying the subject ID, difficulty label, task type, metadata, and frequency-domain EEG columns.
2) Exclude the subject ID, task type, difficulty-level metadata, trial ID, synthetic metadata, and recommended split metadata from the predictive features.
3) Construct your feature set X from the 47 frequency-domain EEG features, while y will be derived from the difficulty label.
4) Split the subjects into 140 for training and 60 for testing.
5) Ensure that there's no overlap in subject IDs between the training and testing sets.
6) Perform feature selection on the training data only, retaining 40 features.
7) Fit the scaler to the selected training features and then transform both the training and testing data using this scaler.
8) Train your primary model using Logistic Regression.
9) For a lightweight comparison, also train Random Forest and Decision Tree models.
10) Evaluate all models on the same unseen subject test set.
11) Compute classification reports, confusion matrices, per-class errors, and feature/band importance for your primary model.
12) Finally, interpret the results as predictive evidence within the dataset's distribution and report any limitations of the dataset.

IV. Results and Discussion

A. Experimental Setup

The experiments were conducted in a Python/Jupyter Notebook environment on a system equipped with a 12th Gen Intel® Core™ i5-1235U processor, 16 GB of RAM, and Windows 11 operating system. The experiments were performed using CPU-based computation without the use of an NVIDIA GPU. Python version 3.13.5 was used for the implementation.

The data preprocessing, frequency-domain feature analysis, model training, and evaluation were performed using standard scientific computing and machine-learning libraries. The versions used were pandas 2.2.3, NumPy 2.1.3, scikit-learn 1.6.1, and Matplotlib 3.10.0.

B. Dataset Description and Evaluation Design

The dataset comprises 48,000 records collected from 200 subjects, covering four types of educational tasks: reading, arithmetic, problem-solving, and multimedia. The target classes are categorised as easy, moderate, and hard, with each subject contributing 240 samples. The dataset features eight EEG channels (F3, F4, C3, C4, P3, P4, O1, and O2), five spectral bands (delta, theta, alpha, beta, and gamma), channel-wise relative power features, five global relative power features, and two global spectral ratios. It's crucial to highlight that the data is explicitly labelled as synthetic, which is a significant limitation that will be discussed further.

Table II. Subject-Independent Data Partitions

Partition	Subj.	Samples	Purpose
Training	140	33,600	Feature selection, scaling, model fitting

Testing	60	14,400	Final unseen-subject evaluation
Overlap	0	—	Leakage validation

C. Primary Performance

Table III. Primary Model Performance (Logistic Regression)

Metric	**Value**
Accuracy	75.69%
Precision	75.89%
Recall	75.69%
F1-score	75.71%
Macro F1	75.71%
Balanced accuracy	75.69%
Chance accuracy	33.33%
Above chance	42.35 pp

The primary model achieved a notably higher performance compared to the balanced three-class chance baseline. Because the test subjects were not involved in the model fitting, feature selection, or scaling, this result provides strong evidence that the learned spectral representation generalised well within the synthetic subject population that was held out. However, it's essential to recognise that this doesn't guarantee that the model will generalise to real learners or clinical-grade EEG data.

D. Class-Level Analysis

Table IV. Per-Class Performance

Class	**Prec.**	**Recall**	**F1**	**Support**
Easy	0.7534	0.8154	0.7832	4,800
Hard	0.8844	0.8290	0.8558	4,800
Moderate	0.6388	0.6262	0.6324	4,800

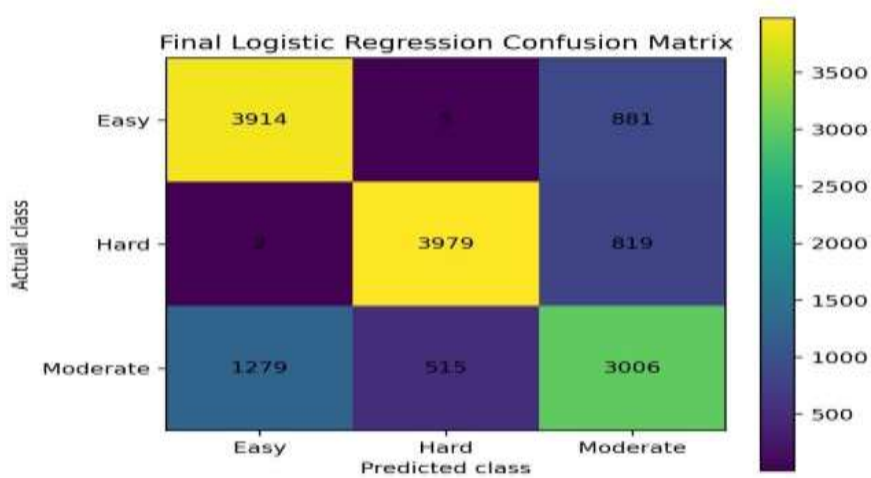

Fig. 2. Confusion matrix of the primary Logistic Regression model.

Hard difficulty was the strongest class, while moderate difficulty was the weakest. The most common confusion was moderate predicted as easy (1,279 errors), followed by easy predicted as moderate (881 errors) and hard predicted as moderate (819 errors). This pattern suggests that the moderate class overlaps more strongly with neighbouring difficulty states than the extreme classes. Accordingly, the overall 75.69% accuracy should not be interpreted as uniform performance across all difficulty levels.

E. Comparative Analysis

Table V. Comparison with Lightweight Baseline Models

Model	Acc.	Prec.	Rec.	F1	Time(s)
Logistic Regression	75.69%	0.759	0.757	0.757	2.33
Random Forest	72.57%	0.728	0.726	0.726	12.53
Decision Tree	65.66%	0.668	0.657	0.661	5.72

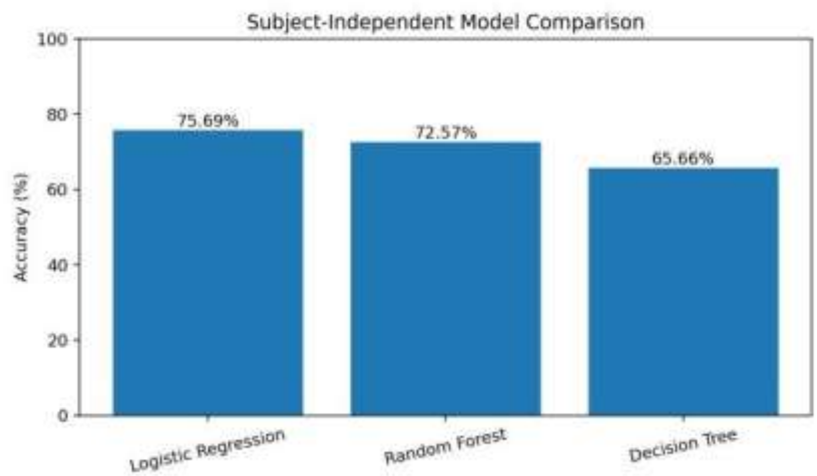

Fig. 3. Lightweight model comparison under the same subject-independent test set.

Logistic Regression stood out as the top performer, delivering the highest accuracy and F1-score of all the models we tested, and it required less training time than the Random Forest model. This outcome suggests that the spectral representation we used had a discernible structure that a relatively simple linear decision function could effectively capture. While this is advantageous for interpretability, it's crucial to note that it doesn't mean Logistic Regression is the ideal choice for every EEG difficulty dataset.

F. Feature and Frequency-Band Interpretation

Table VI. Top-10 Features by Mean Absolute Coefficient

Rank	Feature	Coeff.
1	global_theta_alpha_ratio	1.7346
2	global_beta_alpha_ratio	1.6821
3	P3_beta_relpower	0.1888

4	P4_beta_relpower	0.1838
5	global_beta_relpower	0.1668
6	O2_beta_relpower	0.1545
7	C3_beta_relpower	0.1544
8	C4_beta_relpower	0.1495
9	F4_beta_relpower	0.1447
10	F3_alpha_relpower	0.1415

Table VII. Aggregate Frequency-Band Importance

Band	**Importance**	**# Features**
Alpha	2.7604	10
Theta	2.1021	10
Beta	1.3895	9
Gamma	0.2604	9
Delta	0.0440	2

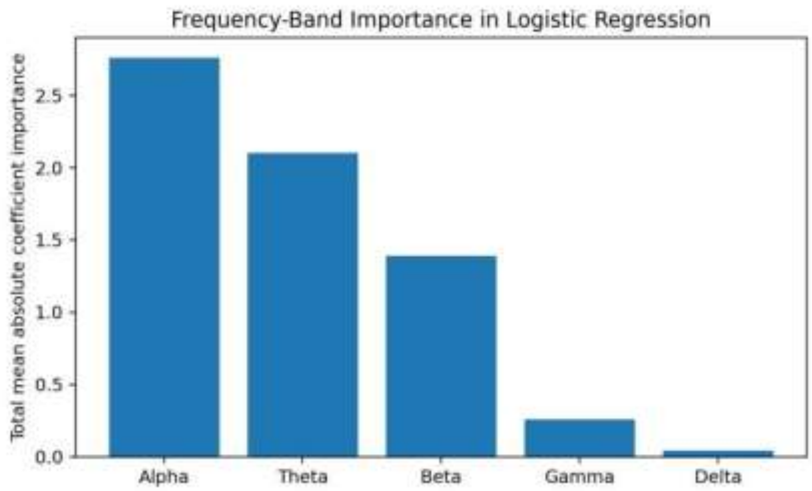

Fig. 4. Aggregate frequency-band importance based on mean absolute Logistic Regression coefficients.

The two global ratio features stood out as the most significant individual features based on their coefficient magnitudes. When we look at the band levels, alpha and theta emerged as the most important, with beta following closely behind. These results align well with existing literature that highlights the relevance of theta, alpha, and beta spectral activity in relation to cognitive workload [1]. However, it's important to note that coefficient magnitude is specific to the model used and shouldn't be seen as a direct measure of causal relationships in neural mechanisms.

G. Discussion

The main research question investigates whether frequency-domain EEG characteristics can effectively differentiate between easy, moderate, and hard educational tasks. Based on the current 70/30 subject-independent evaluation, the answer is a resounding yes in terms of prediction: the primary model achieved an impressive 75.69% accuracy and a 75.71% F1-score on subjects that were not part of the training set. This performance is significantly higher than the 33.33% chance baseline and is further validated by a comparison showing that Logistic Regression outperformed the two simpler alternatives. This finding is particularly significant given the methodological decision to rely solely on spectral EEG data. We intentionally left out task type, subject identity, difficulty-level metadata, trial identifiers, synthetic flags, and recommended split metadata from the predictive inputs. As a result, the performance reported here cannot be attributed to directly encoding the task category or participant identity as features. Additionally, the subject-level split helps mitigate a common issue of overly optimistic EEG results by ensuring that the same subject does not contribute samples to both model training and final evaluation.

Despite the challenges, classifying moderate difficulty turned out to be significantly trickier than categorising easy or hard levels. This makes sense scientifically, as an intermediate category might exhibit traits from both adjacent levels. However, this experiment doesn't pinpoint the physiological reasons behind this overlap. It's crucial to understand that the chosen spectral representation offers valuable, yet incomplete insights into the target labels. This distinction matters: just because we can predict classifications doesn't mean that EEG data alone can define educational difficulty.

The results of the comparison also suggest we should be careful when considering model complexity. Even though Random Forest and Decision Tree models can handle nonlinear patterns, they didn't outperform Logistic Regression. This might suggest that, within the feature space we selected, a linear boundary is sufficient to capture much of the class-discriminative structure. Plus, this offers a practical benefit for interpretation since we can directly examine the logistic coefficients. Future research should explore additional methods, but only under carefully controlled validation conditions and without choosing models based on the final test set.

One major caveat to keep in mind is the synthetic nature of the dataset. While synthetic data can be useful for controlled methodological experiments, the generated distributions might not capture the full range of variability found in real EEG data, including factors like electrode impedance, movement artefacts, physiological differences, task engagement, environmental noise, and label uncertainty. So, it's important to frame the current numerical results as evidence that our computational pipeline can differentiate difficulty labels within this synthetic dataset, rather than as a validated educational neurotechnology system for real-world applications. We need external validation using independently collected real EEG datasets.

V. Conclusion

This study explored whether we can use frequency-domain EEG characteristics to classify the difficulty of educational tasks as easy, moderate, or hard. We put a leakage-controlled machine-learning pipeline to the test on 48,000 records from 200 participants, following a strict 70/30 subject-independent design. Initially, we considered 47 spectral features, narrowed it down to 40 using only the training data, and employed Logistic Regression as our main model. The final model achieved an impressive 75.69% accuracy, 75.89% precision, 75.69% recall, and a 75.71% F1-score on 60 completely new subjects, surpassing the three-class chance level by 42.35 percentage points. Notably, Logistic Regression outperformed both Random Forest and Decision Tree models in the same evaluation setup.

Our error analysis revealed that the moderate difficulty class posed the biggest challenge, while our feature interpretation pointed to global theta/alpha and beta/alpha ratios, along with alpha-, theta-, and beta-related spectral features. These findings suggest that frequency-domain EEG characteristics hold valuable predictive information regarding educational task difficulty in the dataset we examined. However, the study is limited by its reliance on synthetic data, so we need to validate our results with real EEG data before applying them in practice. Future research should focus on cross-dataset transfer, real learner populations, richer temporal features, and calibration strategies, all while maintaining a strict evaluation of unseen subjects.

VI. Limitations and Future Work

• This dataset is synthetic, so keep in mind that the performance results might not directly apply to actual EEG recordings.
• The current target labels are set up as predefined categories of easy, moderate, and hard, which might not fully reflect individual experiences of difficulty.
• We only looked at frequency-domain features; we intentionally left out temporal dynamics, connectivity, and multimodal information.
• The comparison set is made up solely of lightweight models, which means it doesn't provide a complete state-of-the-art benchmark.
• The importance of features is determined by the magnitudes of Logistic Regression coefficients, but it's important not to interpret these causally.
• For future research, it's essential to validate the pipeline using real, independently collected educational EEG data, implement preregistered train/validation/test protocols, explore cross-task and cross-dataset transfer, and compare additional models while ensuring leakage control is maintained.

References

1. S. Chikhi, N. Matton, and S. Blanchet, “EEG power spectral measures of cognitive workload: A meta-analysis,” Psychophysiology, vol. 59, e14009, 2022, doi: 10.1111/psyp.14009.
2. W. L. Lim, O. Sourina, and L. P. Wang, “STEW: Simultaneous Task EEG Workload Data Set,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 26, no. 11, pp. 2106–2114, 2018, doi: 10.1109/TNSRE.2018.2872924.
3. A. Gupta, G. Siddhad, V. Pandey, P. P. Roy, and B.-G. Kim, “Subject-Specific Cognitive Workload Classification Using EEG-Based Functional Connectivity and Deep Learning,” Sensors, vol. 21, no. 20, p. 6710, 2021, doi: 10.3390/s21206710.
4. Y. Zhang et al., “EEG Fingerprints of Task-Independent Mental Workload Discrimination,” IEEE J. Biomed. Health Inform., 2021, doi: 10.1109/JBHI.2021.3085131.
5. “Reproducible machine learning research in mental workload classification using EEG,” Frontiers in Neuroergonomics, 2024.
6. Y. Zhao et al., “EEG/fNIRS Based Workload Classification Using Functional Brain Connectivity and Machine Learning,” Sensors, vol. 22, no. 19, p. 7623, 2022, doi: 10.3390/s22197623.
7. M. R. Safari, R. Shalbaf, S. Bagherzadeh, et al., “Classification of mental workload using brain connectivity and machine learning on electroencephalogram data,” Sci. Rep., vol. 14, Art. no. 9153, 2024, doi: 10.1038/s41598-024-59652-w.
8. Y. Zhou, S. Huang, Z. Xu, P. Wang, et al., “Cognitive Workload Recognition Using EEG Signals and Machine Learning: A Review,” IEEE Trans. Cogn. Develop. Syst., 2021, doi: 10.1109/TCDS.2021.3090217.
9. “Subject-specific mental workload classification using EEG and stochastic configuration network,” Biomed. Signal Process. Control, vol. 68, Art. 102711, 2021.

10. “Cognitive load detection using circulant singular spectrum analysis and Binary Harris Hawks Optimisation based feature selection,” Biomed. Signal Process. Control, vol. 79, Art. 104006, 2023.
11. M. U. Iqbal, B. Srinivasan, and R. Srinivasan, “Multi-class classification of control room operators' cognitive workload using the fusion of eye-tracking and electroencephalography,” Comput. Chem. Eng., vol. 181, Art. 108526, 2024, doi: 10.1016/j.compchemeng.2023.108526.
12. “Machine learning performance in EEG-based mental workload classification across task types: a systematic review,” 2025.
13. A. B. Gevins and M. E. Smith, “Neurophysiological measures of cognitive workload during human-computer interaction,” Theor. Issues Ergon. Sci., foundational workload EEG literature.
14. A. F. Kramer, “Physiological metrics of mental workload: A review of recent progress,” Multiple-task Performance, foundational workload literature.
15. C. Berka et al., “EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks,” Aviat. Space Environ. Med., foundational applied EEG workload literature.