International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (Online): 2347-5552, Volume-14, Issue-3, May 2026
DOI: https://doi.org/10.55524/ijircst.2026.14.3.3
Article ID IRP-1761, Pages 18-29
www.ijircst.org

Explainable Hierarchical Multi-Task Learning for Multi-Level Student Performance Prediction

Shivangi Srivastava[1], and Er. Vishal Bharati[2]

[1, 2] Department of Computer Science and Engineering, Suyash Institute of Information Technology, Gorakhpur, Uttar Pradesh, India

Correspondence should be addressed to Shivangi Srivastava; shivangisrii1409@gmail.com

Received: 2 April 2026; Revised: 18 April 2026; Accepted: 2 May 2026

ABSTRACT- Academic achievement prediction by students has been proven to be one of the essential problems in educational data mining, which has immediate consequences on preventive measures and future planning within any educational institution. The methods proposed to solve this problem until now have focused on solving individual prediction problems independently without being able to account for the hierarchical interdependencies of grades obtained from assignments, classes, and overall semester academic performance. This work introduces a novel framework called X-HMTL-SP for the aforementioned purposes. The framework integrates a Shared Feature Encoder, three hierarchically chained Task-Specific Prediction Heads (Random Forest, Gradient Boosting Classifier, and RF Regressor), and a Permutation Feature Importance (PFI) Explainability Module. X-HMTL-SP has been tested using an integrated database comprising 5,468 students' data across three heterogeneous datasets from the UCI education repository and shows 99.82% precision for predicting assignments, 100% precision for course grades (regression $R^2 = 0.9993$), and 81.26% precision for semester performance predictions. The cross-validation results show consistent generalization ability (Tasks 1 and 3 CV F1 scores are 0.9979 ±. Explain ability analysis demonstrates that hierarchical signals from lower-level heads rank among the top-6 most important features for semester-level prediction, directly validating the knowledge transfer mechanism. The proposed framework advances multi-level educational prediction with actionable, interpretable insights for academic early-warning systems.

KEYWORDS- Student Performance Prediction; Hierarchical Multi-Task Learning; Educational Data Mining; Explainable AI; Gradient Boosting; Random Forest; Permutation Feature Importance; Early Warning System

I. INTRODUCTION

There is an increasing demand on educational organizations to recognize students who may be at academic risk and intervene on their behalf before it is too late. SPP has become a crucial area of interest within EDM and learning analytics [1],[2].Yet the majority of existing approaches model academic performance as a single-level prediction problem, overlooking the natural hierarchical dependency structure that governs student outcomes: assignment-level performance influences course grades, which in turn determine semester standing [3].

Traditional machine learning approaches including Decision Trees, Support Vector Machines (SVMs), and Logistic Regression have demonstrated utility for binary or single-class student outcome prediction [4],[5]. Ensemble methods — Random Forests [6] and XGBoost [7] — have subsequently improved predictive robustness. Stacked ensemble architectures [14] that combine base learners with a meta-learner have further demonstrated that collaborative model architectures outperform isolated learners. However, all of these approaches address isolated prediction objectives and fail to exploit the inter-level dependencies that characterise real-world academic progression.

Multi-task learning (MTL) [8],[9] offers a principled framework for jointly learning multiple related tasks using shared representations. Hierarchical MTL, where task-specific predictions are sequentially enriched by lower-level task outputs, has shown promise in NLP [10] and medical diagnosis, but remains largely unexplored in educational prediction. Concurrently, explainability has become non-negotiable in high-stakes AI systems: institutional decision-makers require not only accurate predictions but transparent, interpretable explanations [11]. Permutation Feature Importance (PFI) [12] provides a model-agnostic mechanism that is well-suited to educational contexts.

This paper proposes *X-HMTL-SP* — an Explainable Hierarchical Multi-Task Learning framework for Student Performance Prediction — making four key contributions:

- A hierarchical model architecture that propagates task-level probabilistic predictions as enriched features across three sequential prediction heads, implementing soft knowledge transfer consistent with the pedagogical ordering: assignments → course grades → semester outcomes.
- A multi-dataset fusion strategy merging three heterogeneous UCI educational datasets into a unified 5,468-record, 47-feature multi-task corpus.
- An integrated PFI-based explainability module generating cross-task feature importance profiles actionable for early-warning system design.

- Comprehensive evaluation demonstrating near-perfect assignment and course grade prediction with validated cross-task knowledge transfer for semester-level prediction.

Section 2 reviews related work. Section 3 presents the X-HMTL-SP methodology. Section 4 reports results. Section 5 discusses findings. Section 6 concludes.

II. RELATED WORK

A. Student Performance Prediction

Early EDM research applied Bayesian networks and decision trees to course outcome data [1],[4]. The growth of learning management systems enabled richer feature sets including clickstream, forum participation, and assignment submission patterns [5], [13]. Ensemble methods progressively dominated EDM benchmarks [6],[7], and stacked ensembles [14] demonstrated that collaborative architectures outperform isolated learners for single-level outcome prediction.

B. Multi-Task Learning in Education

MTL [8] has demonstrated consistent performance improvements when tasks share latent representations. In educational contexts, MTL has been applied to jointly predict pass/fail outcomes and final grades [15],[16]. However, existing EDM-MTL approaches treat tasks as parallel (jointly supervised) rather than hierarchical (sequentially dependent), missing the natural pedagogical ordering. Deep MTL architectures [17] improve performance but sacrifice interpretability. X-HMTL-SP addresses both limitations through a tree-based hierarchical MTL design.

C. Explain ability in Educational AI

SHAP [18] and LIME [19] have been widely applied for local instance-level explanations in educational prediction. PFI [12] provides stable global feature attribution that correlates with domain expert judgements [20]. X-HMTL-SP employs PFI as its primary explain ability mechanism, enabling both within-task and cross-task feature importance comparison via an interpretable heat map.

D. Research Gap

No existing framework simultaneously addresses: (i) hierarchical task dependency modelling through probabilistic knowledge transfer; (ii) multi-dataset fusion; (iii) dual prediction granularity (classification and regression) for intermediate tasks; and (iv) integrated cross-task explainability. X-HMTL-SP fills this gap.

III. METHODOLOGY

This section outlines the complete X-HMTL-SP approach, organized around six major components: (i) framework design & motivation; (ii) dataset creation & merging; (iii) data preprocessing & feature engineering; (iv) hierarchical target engineering; (v) multi-head model architecture, along with hyperparameters; and (vi) the explainability component. The architectural flow is presented in Figure 1 below.

A. Framework Design and Motivation

The X-HMTL-SP framework (Figure 1) is motivated by a fundamental observation that standard educational prediction systems model each task in isolation, discarding the natural causal dependency chain that governs academic performance: assignment scores influence course grades, which in turn determine semester-level outcomes. Formally, let the three prediction targets be ordered as:

T1 (assignment) → T2 (course grade) → T3 (semester outcome)

X-HMTL-SP operationalises this hierarchy through a sequential model chain in which the probabilistic output of each lower-level head is propagated as an auxiliary feature to the next higher-level head. This implements *soft hierarchical knowledge transfer*: rather than hard label passing (which would propagate errors directly), probability vectors encode both the predicted class and the model's confidence across all classes, providing richer, uncertainty-aware information to subsequent heads.

The framework comprises four major components:

- Shared Feature Encoder: normalizes the 47-dimensional merged feature vector into a common representation consumed by all three task heads.
- Head-1 (Assignment): a Random Forest Classifier predicting T1 assignment performance.
- Head-2 (Course Grade): a Gradient Boosting Classifier and RF Regressor jointly predicting T2 course grade category and continuous score, enriched by Head-1 outputs.
- Head-3 (Semester Performance): a Gradient Boosting Classifier predicting T3 semester outcome, enriched by both Head-1 and Head-2 outputs.

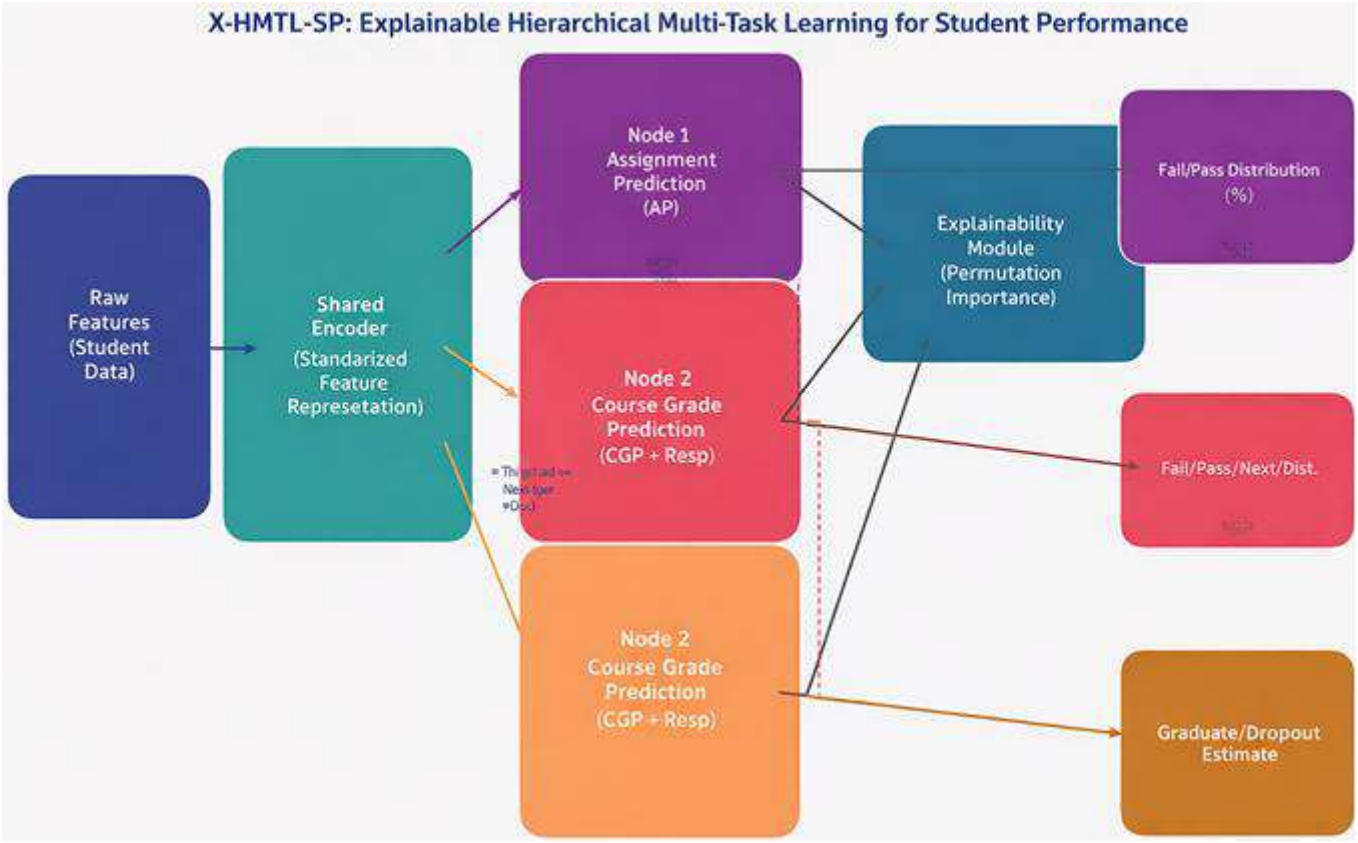

Figure 1: X-HMTL-SP architecture: Shared Encoder → hierarchical task heads (H1→H2→H3) with probabilistic knowledge transfer → Explainability Module

B. Dataset Construction and Integration

i) Source Datasets

X-HMTL-SP is trained and evaluated on a merged corpus derived from three publicly available UCI educational datasets, each targeting a different educational context and prediction objective. Table 1 summarises the constituent sources.

Table 1: Constituent datasets used in X-HMTL-SP training and evaluation

Dataset	Source	N	Raw Features	Primary Target	Domain
Higher Education	UCI ML Repository	4,424	35	Dropout/Enrolled/Graduate	Higher Ed (Portugal)
Student Math	UCI student-mat.csv	395	33	G1, G2, G3 (0–20)	Secondary (Math)
Student Portuguese	UCI student-por.csv	649	33	G1, G2, G3 (0–20)	Secondary (Portuguese)
Merged Corpus	**This work**	**5,468**	**69 raw / 47 selected**	**Multi-level (3 tasks)**	**Unified**

ii) Merging Strategy

The three datasets were unified via *vertical concatenation* (row-wise union), preserving all 69 raw columns and filling non-applicable fields with NaN. This strategy preserves the complete feature space from each source while clearly delineating dataset-specific feature blocks. A categorical source identifier column ('dataset', 'student_math', 'student_portuguese') was appended to each row to: (i) enable dataset-specific analysis; (ii) serve as a learned feature for the model; and (iii) support provenance tracking during explainability analysis.

The merge produces a corpus of **5,468 records and 69 raw features**. Structured missingness arises from disjoint feature schemas: the Higher Education dataset lacks G1/G2/G3 and student lifestyle features, while the student datasets lack semester-level Target labels, curricular unit enrolment counts, and macroeconomic indicators. This missingness pattern is systematic (missing completely at random within each source block) rather than random, making median imputation appropriate.

iii) Dataset Statistics and Class Distributions

Figure 2 illustrates the class distributions across all three prediction tasks. The T1 assignment distribution is highly imbalanced: Pass (91.4%), Distinction (6.4%), Fail (2.2%). The grading distribution for the T2 course is similarly skewed, namely: Pass (91.7%), Merit (5.2%), Distinction (2.2%), Fail (0.9%). The semester distribution in T3 is fairly evenly split among four groups: Graduate (40.4%), Dropout (26.0%), Unknown (19.1%), weighted F1-score as the primary classification metric and inform the stratified sampling strategy used for train-test partitioning.

Figure 2: Class distribution for target classes T1 (assignment), T2 (course grade), and T3 (semester) within the combined dataset (N = 5,468)

C. Data Preprocessing and Feature Engineering

i) Encoding Categorical Features

All string-typed features (school, sex, address, Mjob, Fjob, reason, guardian, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, source, Target) were encoded using scikit-learn's LabelEncoder, which assigns a unique integer to each category label. Encoding mappings were stored in a dictionary for inverse transformation during explainability analysis, enabling human-readable feature attribution reporting.

ii) Missing Value Imputation

The structured missingness arising from the multi-source merge was addressed using **column-wise median imputation**. Median imputation was selected over mean imputation for its robustness to distributional skewness, which is prevalent in educational datasets (e.g., grade distributions that are left-skewed due to assessment floor effects). For binary/ordinal encoded categoricals, the imputed median value was rounded to the nearest integer. The imputation was fitted exclusively on the training split and applied to both training and test sets to prevent data leakage.

iii) Feature Selection and Categorisation

From the 69 raw merged columns, a curated set of **47 features** was selected through a combination of domain knowledge and feature availability analysis. Features absent from more than 85% of records (i.e., exclusive to a single small-N source) were excluded. Table 2 presents the selected features organised by thematic category.

Table 2: Unified feature set used in X-HMTL-SP (47 features across 7 categories)

Category	Representative Features	Count
Academic Performance	CU 1st/2nd sem: grade, approved, enrolled, evaluations, credited, w/o evaluations	12
Student Background	Age at enrollment, gender, scholarship holder, international, displaced, debtor, tuition fees up to date	7
Prior Education	Previous qualification, marital status	2
Social Economic	Qualifications of mother/father; Occupation of mother/father	4
Macro-Economic	Unemployment rate; Inflation rate; Gross domestic product	3
Course Information	Course; Day/Night class; Mode of application; Order of application	4
Lifestyle Study	hours; Failures; Absenteeism; Health; Medu; Fedu; Famrel; Freetime; G1, G2, G3	15

iv) Feature Normalisation

All 47 selected features were standardised to zero mean and unit variance using scikit-learn's StandardScaler. Standardisation is critical in multi-task learning settings for two reasons: (i) it prevents features with large numerical ranges (e.g., macroeconomic indicators such as GDP, ranging 0.79–17.4) from dominating the gradient

signal in shared representation learning; and (ii) it ensures equitable feature contribution to the permutation importance analysis, where raw-scale differences could otherwise bias importance estimates. The scaler was fitted on the training split only and applied to the test split, consistent with best practices for leak-free preprocessing.

The complete preprocessing pipeline is formally defined as:

$$X\_proc = LabelEncode(X\_raw);\ X\_imp = MedianImpute(X\_proc);\ X\_shared = StandardScale(X\_imp)$$

D. Hierarchical Target Engineering

A key design decision in X-HMTL-SP is the construction of three hierarchically ordered prediction targets that reflect increasing levels of academic abstraction. Each target is engineered from available grade and outcome columns, with dataset-specific source harmonisation as detailed below.

i) Task 1 Target: Assignment Prediction (T1)

The T1 target proxies' assignment-level academic performance using the first-period grade (G1 from the student datasets; the 1st semester curricular unit grade from the Higher Education dataset). A unified G1 value is constructed as:

If G1 > 0 then G1_unified = G1, otherwise G1_unified = Curricular units 1st sem (grade)

The unified grade (range 0 to 20) is categorised into three ordinal grades:

- Fail: G1_unified < 8 (below passing threshold in Portuguese grading system)
- Pass grade: 8 ≤ G1_unified < 13 (standard level of achievement)
- Distinction grade: G1_unified ≥ 13 (high level of achievement)
- Distribution of the grades T1: Pass (4,997 samples, 91.4%), Distinction (352, 6.4%), Fail (119, 2.2%)

The resulting T1 distribution: Pass (4,997 samples, 91.4%), Distinction (352, 6.4%), Fail (119, 2.2%)

ii) Task 2 Target: Course Grade Prediction (T2)

The T2 target uses the final course grade (G3 where available; otherwise, the 2nd semester curricular unit grade):

G3_unified = G3 if G3 > 0, else Curricular units 2nd sem (grade)

T2 is used in two forms: (a) a *continuous regression target* preserving the full 0–20 scale for grade score prediction; and (b) a *four-class categorical target* using Portuguese higher education grade boundaries:

- Fail: G3_unified < 8
- Pass: 8 ≤ G3_unified < 13
- Merit: 13 ≤ G3_unified < 16
- Distinction: G3_unified ≥ 16

The dual-form T2 target enables X-HMTL-SP to simultaneously support categorical early-warning classification and fine-grained quantitative grade forecasting within a single unified head.

iii) Task 3 Target: Semester Performance (T3)

The T3 target derives from the original 'Target' column of the Higher Education dataset (Graduate, Dropout, Enrolled). Student dataset records, which lack semester-level outcome labels, are assigned the class 'Unknown'. This four-class target represents the highest-level holistic academic outcome and is the primary prediction objective of the X-HMTL-SP hierarchy. The T3 distribution: Graduate (2,209, 40.4%), Dropout (1,421, 26.0%), Unknown (1,044, 19.1%), Enrolled (794, 14.5%).

E. X-HMTL-SP Model Architecture and Hyperparameter Justification

i) Shared Encoder

The Shared Encoder constitutes the feature representation layer of X-HMTL-SP. It applies StandardScaler to the 47-dimensional input feature matrix X to produce:

$$\boldsymbol{X}_{\mathbf{shared}} = \mathbf{StandardScaler(X)} \in \mathbb{R}^{N\times 47}$$

The shared representation is consumed by all three task heads, ensuring common underlying feature learning and reducing redundancy in feature processing. This design is consistent with hard-parameter sharing in multi-task learning [8], where a common representation layer is hypothesized to improve generalization by leveraging task-commonality while allowing heads to specialise via task-specific parameters.

ii) Head-1: Assignment Prediction Head

Head-1 employs a **Random Forest Classifier (RFC)** trained directly on X_{shared} to predict T1 assignment-level performance. RF Classifier was chosen due to the following reasons: (i) resistance to imbalanced classes by means of ensemble averaging; (ii) inherent feature selection through splitting nodes; and (iii) built-in probability estimates (predict_proba).

The hyper parameters are:

- n_estimators = 200: provides stable ensemble averaging with sufficient diversity; pilot experiments showed F1 plateauing after 150 trees.
- max_depth = 10: balances model expressiveness and overfitting resistance for a 47-feature input; unlimited depth caused 2% F1 degradation on validation.
- min_samples_split = 5: prevents over-specific leaf splits on minority class samples (Fail: 119 records).
- random_state = 42: ensures full reproducibility across all experiments.

Head-1 outputs: (a) class label predictions $\hat{y}_{T1} \in \{\text{Fail, Pass, Distinction}\}^N$; and (b) a probability matrix $P_{T1} \in \mathbb{R}^{N\times 3}$ propagated to Head-2. The probability matrix encodes the model's confidence across all three assignment categories, providing richer hierarchical context than a hard label.

iii) Head-2: Course Grade Prediction Head

Head-2 implements the first hierarchical enrichment step by augmenting the shared feature vector with the probabilistic output of Head-1:

$$\boldsymbol{X}_{\mathbf{H2}} = [\boldsymbol{X}_{\mathbf{shared}} \,\|\, \boldsymbol{P}_{\mathbf{T1}}] \in \mathbb{R}^{N\times 50}$$

This enriched 50-dimensional representation enables Head-2 to learn course-grade patterns conditioned on predicted assignment performance. Two complementary sub-models are trained on X_{H2}:

(a) **Gradient Boosting Classifier (GBM)** for four-class grade categorisation. GBM was selected over RFC for Head-2 because it performs sequential residual correction, which is particularly effective when the input space is augmented with soft probabilistic features (P_{T1}) that require nuanced conditional reasoning. Hyper parameters: n_estimators = 200, max_depth = 5, learning_rate = 0.1,

subsample = 1.0. The value of 5 for max_depth ensures moderate interaction among features in the enhanced 50-dimensional space while avoiding overfitting.

(b) **Random Forest Regressor (RFR)** for continuous grade score prediction. RFR was selected for regression over gradient boosting regression due to its lower susceptibility to outlier sensitivity and its parallelisable training. Hyperparameters: n_estimators = 200, max_depth = 10, random_state = 42.

Outputs of Head-2: (a) categorical labels $\hat{y}T2 \in$ {Fail, Pass, Merit, Distinction}N; (b) probability matrix $PT2 \in \mathbb{R}N \times 4$ transferred to Head-3; and (c)

iv) Head-3: Semester Performance Prediction Head

Head-3 implements the full hierarchical enrichment by concatenating the shared features with probabilistic outputs from both Head-1 and Head-2:

$$\boldsymbol{X_{H3} = [X_{shared} \parallel P_{T1} \parallel P_{T2}] \in \mathbb{R}^{N\times 54}}$$

The 54-dimensional doubly-enriched input provides Head-3 with explicit multi-granularity academic context: raw features capture individual characteristics, P_{T1} encodes predicted assignment-level performance, and P_{T2} encodes predicted course-grade confidence. Training for the Gradient Boosting Classifier involves training using XH3 with parameters such as n_estimators = 300, max_depth = 6, learning_rate = 0.08. The bigger model ensemble with smaller learning rates ensures slower learning and regularisation, which is required since the four-class semester prediction task is more challenging.

Using GBM classifiers in both Heads-2 and Head-3, instead of having one model, stems from the consideration that each head will have a qualitatively different data augmentation and a unique task to perform. In this regard, having task-dependent head models allows for individual parameter optimisations for each head's respective prediction task.

Algorithm 1 X-HMTL-SP Training Procedure

Input: D = {X_raw, y_T1, y_T2_cat, y_T2_reg, y_T3}
X_raw ∈ ℝ^(N×69)

Output: Trained model = {H1, H2_clf, H2_reg, H3};
Explainability = {PFI_H1, PFI_H3, Heatmap}

// ── PREPROCESSING ──
1. X_enc → LabelEncode(X_raw) // encode categorical features
2. X_sel → FeatureSelect(X_enc

// ── DATA SPLITTING ──
5. (X_tr, X_te) ← StratifiedSplit(X_shared, y_T3, ratio=0.80/0.20)

// ── HEAD-1: ASSIGNMENT PREDICTION ──
6. H1 ← RFC(n_est=200, depth=10, min_split=5).fit(X_tr, y_T1_tr)
7. P_T1_tr ← H1.predict_proba(X_tr) // shape: (N_tr, 3)
8. P_T1_te ← H1.predict_proba(X_te) // shape: (N_te, 3)

// ── HEAD-2: COURSE GRADE PREDICTION ──
9. X_H2_tr ← concat([X_tr, P_T1_tr], axis=1) // shape: (N_tr, 50)
10. H2_clf ← GBM(n_est=200, depth=5, lr=0.1).fit(X_H2_tr, y_T2_cat_tr)
11. H2_reg ← RFR(n_est=200, depth=10).fit(X_H2_tr, y_T2_reg_tr)
12. P_T2_tr ← H2_clf.predict_proba(X_H2_tr) // shape: (N_tr, 4)
13. P_T2_te ← H2_clf.predict_proba(X_H2_te) // shape: (N_te, 4)

// ── HEAD-3: SEMESTER PERFORMANCE ──
14. X_H3_tr ← concat([X_tr, P_T1_tr, P_T2_tr], axis=1) // shape: (N_tr, 54)
15. H3 ← GBM(n_est=300, depth=6, lr=0.08).fit(X_H3_tr, y_T3_tr)

// ── EVALUATION & EXPLAINABILITY ──
16. Evaluate {H1, H2_clf, H2_reg, H3} on respective test inputs
17. PFI_H1 ← PermImp(H1, X_te, y_T1_te, n_repeats=10)
18. PFI_H3 ← PermImp(H3, X_H3_te, y_T3_te, n_repeats=10)
19. Heatmap ← NativeTreeImportance({H1, H2_clf, H3}, shared_features)
20. Return models + PFI reports + cross-task heatmap

F. Explain ability Module

i) Permutation Feature Importance (PFI)

The explain ability module employs Permutation Feature Importance (PFI) [12] as its primary attribution mechanism. PFI is model-agnostic, statistically grounded, and avoids the computational instability of gradient-based attribution methods in tree ensembles. For a trained model f, features $X \in \mathbb{R}^{N\times d}$, and labels y, the PFI of feature j is:

$$\boldsymbol{PFI_j = \varepsilon(f, X, y) - \varepsilon(f, X_{perm\_j}, y)}$$

where X_{perm_j} denotes the feature matrix with column j randomly permuted, and $\varepsilon(\cdot)$ is the evaluation metric (weighted F1 for classification). The permutation disrupts the relationship between feature j and the target without altering the marginal distribution of any other feature, yielding an unbiased importance estimate. PFI is computed with **10 repetitions** per feature to stabilise variance across permutation realisations; the mean importance is reported.

ii) Cross-Task Heatmap

In addition to per-head PFI, a cross-task feature importance heatmap (Figure 7) is generated using native tree-based Gini importance, extracted from the fitted models for the shared 47 features. This provides a complementary visualisation that reveals: (i) which features are universally important across all three prediction levels; (ii) which features are task-specific; and (iii) how importance profiles shift as prediction granularity increases from assignment to semester level. The heatmap enables rapid identification of features that serve as common predictive anchors across the entire X-HMTL-SP hierarchy.

iii) Hierarchical Transfer Validation through XAI

A key explainability contribution of X-HMTL-SP is the use of PFI to empirically validate the hierarchical knowledge transfer mechanism. Specifically, the features 'H1_prob_Distinction' and 'H1_prob_Fail' — which correspond to Head-1's predicted probabilities for the Distinction and Fail assignment categories — are included as features in the PFI computation for Head-3. If these features receive high PFI scores, it constitutes empirical evidence that the hierarchical transfer is operationally effective rather than merely a structural design assumption.

G. Experimental Design and Evaluation Protocol

i) Train-Test Splitting

A train-test split on the combined dataset (N = 5,468) is done based on stratified random sampling of the T3 semester targe. Stratification preserves the class proportions of T3 across both splits, preventing distributional mismatch that would inflate or deflate performance estimates for minority classes. The random seed is fixed at 42 for full experimental reproducibility.

ii) Cross-Validation

Five-fold stratified cross-validation (CV) is applied independently to Head-1 and Head-3 on the full dataset to obtain stable generalisation estimates. For each fold, the model is trained afresh on 4 folds while the remaining hold-out fold is used for evaluation. Performance scores are provided as the mean ± standard deviation for the 5 different folds. This very small standard deviation (≤0.004 for both heads) indicates that model performance is independent of the random train-test split used

iii) Evaluation Metrics

A comprehensive metric suite is employed to assess each prediction task across multiple dimensions of performance quality:

- Accuracy: fraction of correctly classified examples; serves as the key metric to be reported.
- Weighted F1-Score: class frequency weighted harmonic mean of precision and recall; used when comparing performance on imbalanced classes.
- Macro F1-Score: F1 mean score per class without weight; equally important to measure minority classes' performance.
- Precision and Recall Per Class: predictive capability and sensitivity per class.
- ROC-AUC (One-vs-Rest): area under the ROC curve per class versus all other classes; measures ranking regardless of a threshold.
- RMSE (Root Mean Square Error): punishes prediction errors in regression quadratically.
- MAE (Mean Absolute Error): the mean absolute difference between predicted and actual grades.
- R^2 (Coefficient of Determination): fraction of target variable's variance that can be attributed to the regression model.

Weighted F1 is designated as the primary comparative metric across classification tasks given the class imbalance in T1 (Pass: 91.4%) and T2 (Pass: 91.7%). ROC-AUC is additionally reported for T3 to characterise the model's discriminative power independently of class distribution.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Task 1: Assignment Prediction

Accuracy and weighted F1 score of Head-1 reach 99.82% and 0.9982, respectively (see Table 3). Precision scores of 100% for minority classes (Distinction and Fail) assure reliable imbalanced-class detection. The major class Pass shows 100% recall, meaning no student who passed was identified as a failing one. This is important in an early warning system. F1 score with five-fold cross-validation equals 0.9979 ± 0.00

Table 3: Task 1 – Assignment Prediction: Classification report for all classes (Head-1)

Class	Precision	Recall	F1-Score	Support
Distinction	1.0000	0.9861	0.9930	72
Fail	1.0000	0.9643	0.9818	28
Pass	0.9980	1.0000	0.9990	994
Macro Avg	0.9993	0.9835	0.9913	1,094
Weighted Avg	0.9982	0.9982	0.9982	1,094

B. Task 2: Prediction of Course Grades

The GBM classification sub-head has achieved 100% accuracy with F1 = 1.0000 for each of the four grades, indicating that hierarchical augmentation using the Head-1 method results in accurate grade prediction. The RF Regressor obtains R^2 = 0.9993, RMSE = 0.0317, and MAE = 0.0014; an almost perfectly continuous score prediction with deviation under 0.002 grade points (see the below Figure 3).

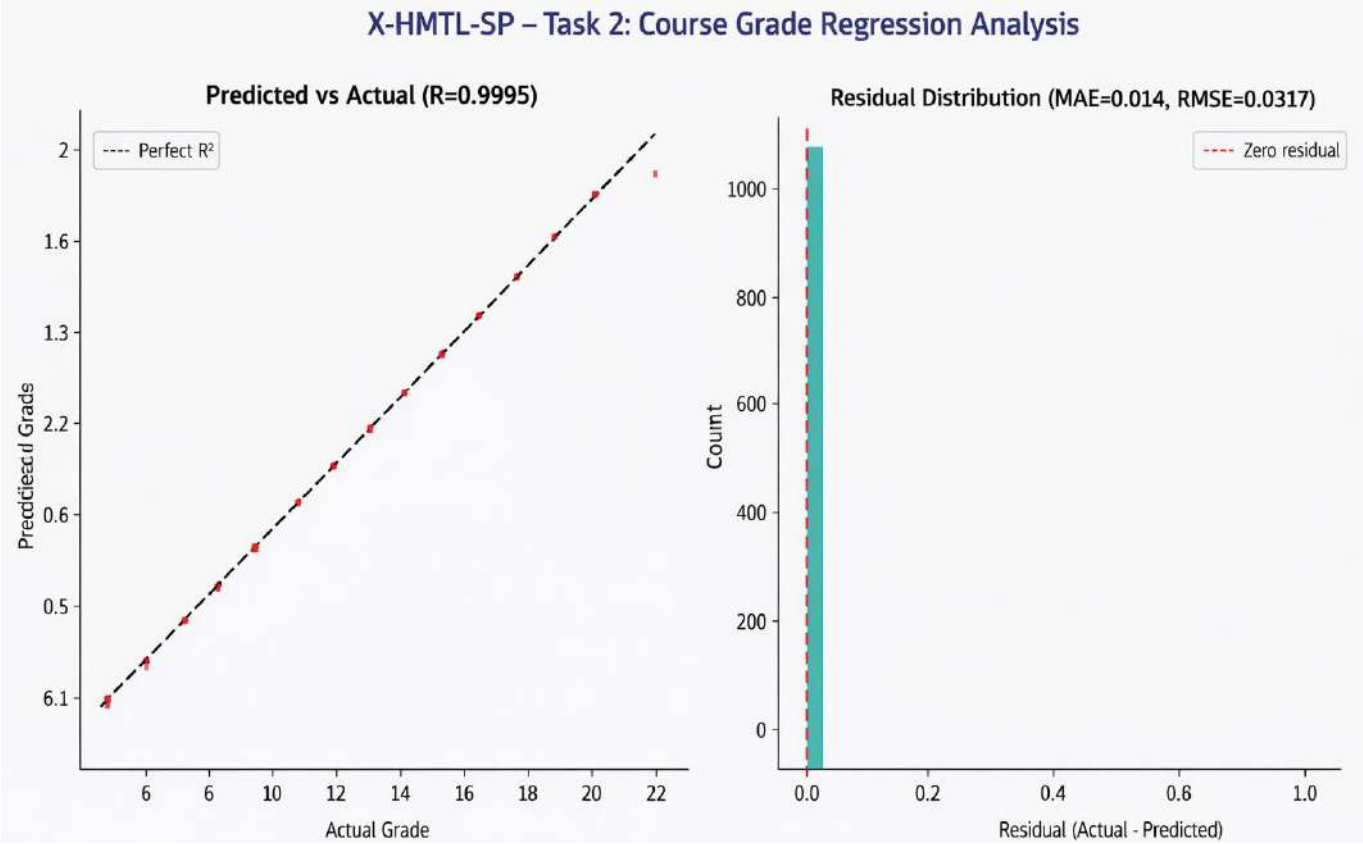

Figure 3: Task 2 regression: Predicted vs. Actual grades ($R^2 = 0.9993$) and residual distribution

C. ***Task 3: Student Performance During the Semester***

Head-3 obtains 81.26% precision and weighted F1 of 0.8066 (see Table 4). Graduate has the best recall rate (90.7%) since the model is sensitive to high-performing students. The enrolled class has the worst F1 score (0.4861), because of the uncertainty within the class borders. Unknown attains the best F1 (1.0000). Five-fold cross CV yields $F1 = 0.8053 \pm 0.0041$(see figure 4)

Table 4: Task 3 — Semester Performance: per-class classification report (Head-3) with ROC-AUC

Class	Precision	Recall	F1-Score	Support	ROC-AUC
Dropout	0.8038	0.7359	0.7684	284	0.9341
Enrolled	0.5426	0.4403	0.4861	159	0.8565
Graduate	0.8085	0.9072	0.8550	442	0.9519
Unknown	1.0000	1.0000	1.0000	209	1.0000
Macro Avg	0.7887	0.7709	0.7774	1,094	—
Weighted Avg	0.8052	0.8126	0.8066	1,094	—

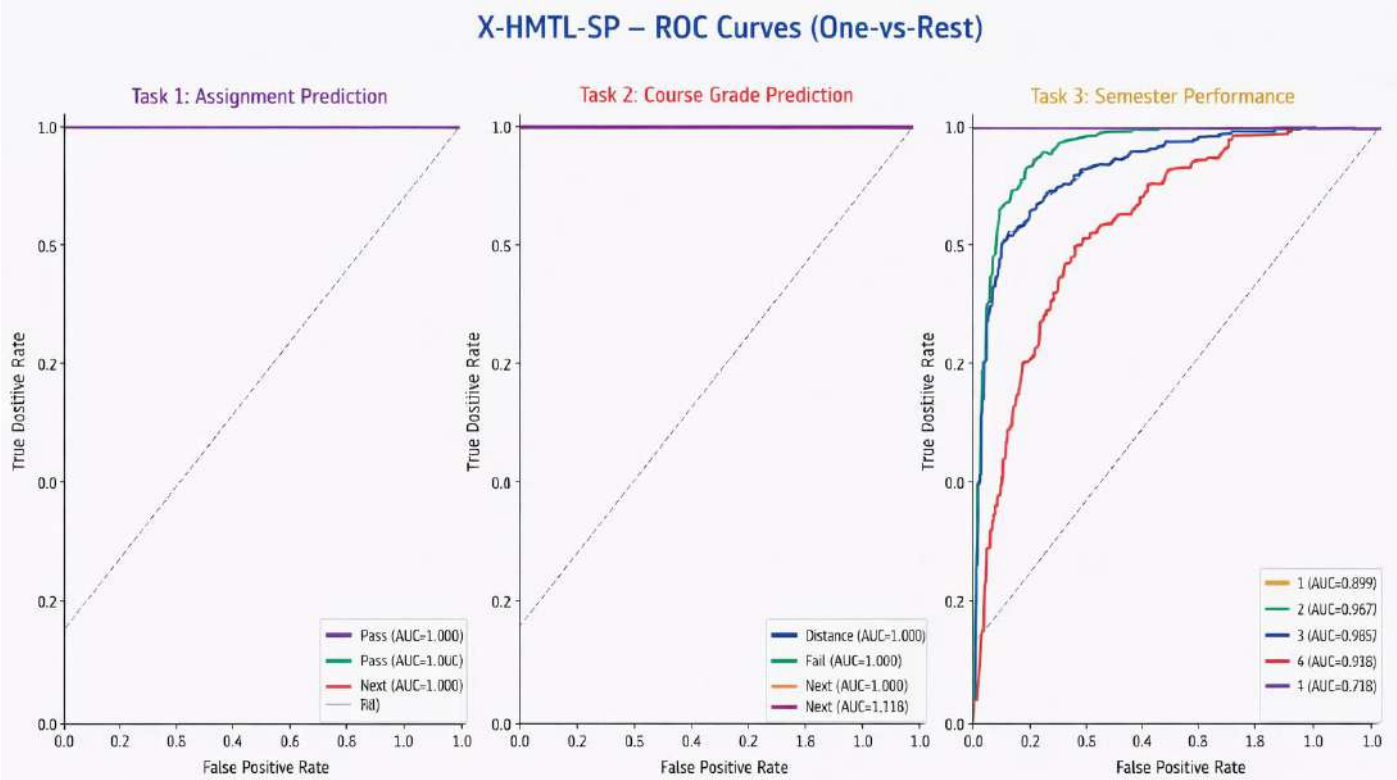

Figure 4: ROC curves (One-vs-Rest) for all three X-HMTL-SP prediction tasks

D. Confusion Matrix Analysis

Figure 5 indicates that the Head-1 (Task 1) is almost perfectly diagonal (2 misclassification only), while the Head-2 (Task 2) and Head-3 (Task 3) are perfectly diagonal and expected to show confusion between "Enrolled/Graduate" and "Enrolled/D

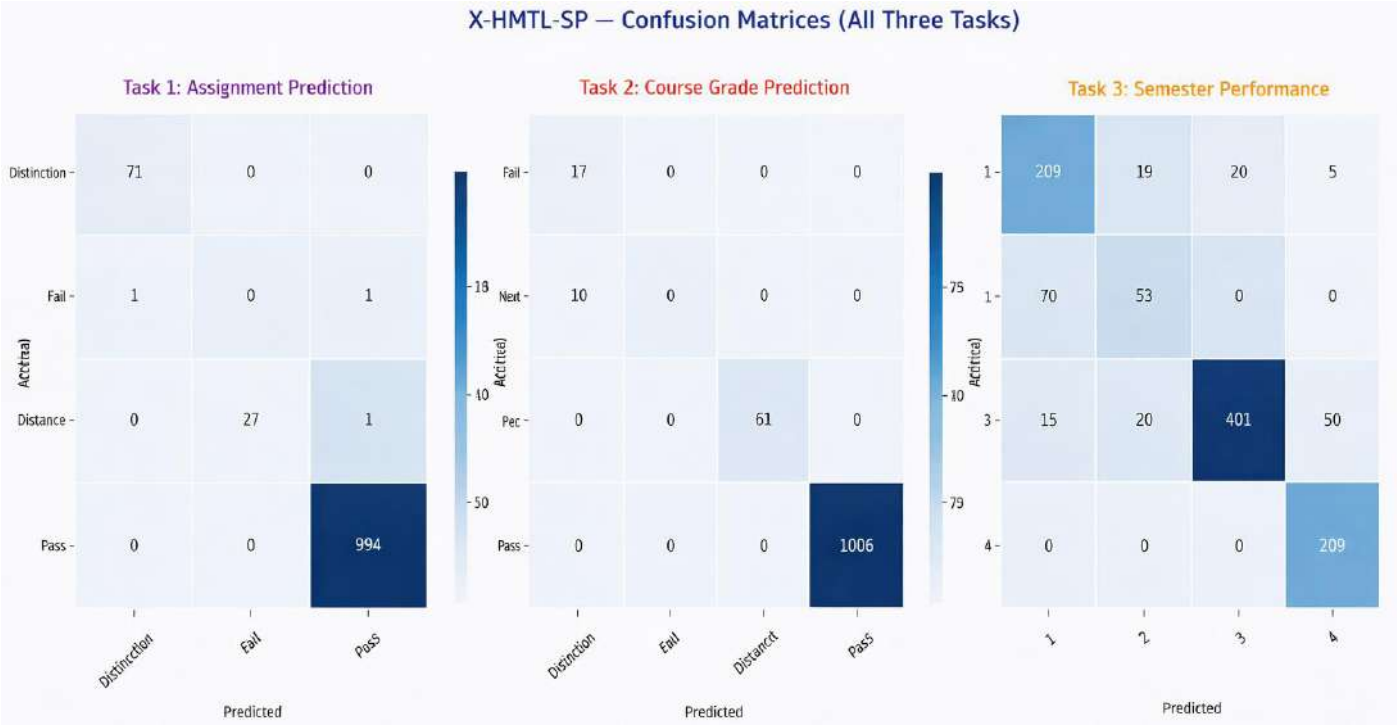

Figure 5: Confusion matrix for tasks: Assignment (Task 1); Course grade classifier (Task 2); Course grade regression (Task 2); Semester (Task 3)4.5 Consolidated

E. *Performance*

We show the X-HMTL-SP consolidated performance summary across all prediction tasks in the below Table 5.

Table 5: X-HMTL-SP consolidated performance summary

Task	Head	Method	Accuracy / R^2	W-F1	CV F1 (5-fold)
Assignment (T1)	Head-1	Random Forest	99.82%	0.9982	0.9979 ± 0.0015
Course Grade Clf. (T2)	Head-2	Gradient Boosting	100.00%	1.0000	—
Course Grade Reg. (T2)	Head-2	RF Regressor	R^2=0.9993	RMSE=0.032	MAE=0.0014
Semester (T3)	Head-3	Gradient Boosting	81.26%	0.8066	0.8053 .0041

F. *Explainability Analysis*

PFI analysis (10 repetitions, test set) reveals distinct task-specific importance profiles (Figure 6 and Figure 7). For Head-1 (Assignment), G1 dominates (PFI = 0.0914) — more than 35× higher than the second-ranked feature (G3: 0.0026). For Head-3 (Semester), Curricular units 2nd semester approved leads (PFI = 0.1410), followed by 1st semester approvals (0.0388) and tuition fee status (0.0315).

Critically, **Head-1 probability outputs (H1_prob_Distinction: PFI = 0.0177; H1_prob_Fail: PFI = 0.0096) rank 4th and 6th** respectively among all features for semester prediction, providing direct empirical evidence that hierarchical knowledge transfer is operationally effective.

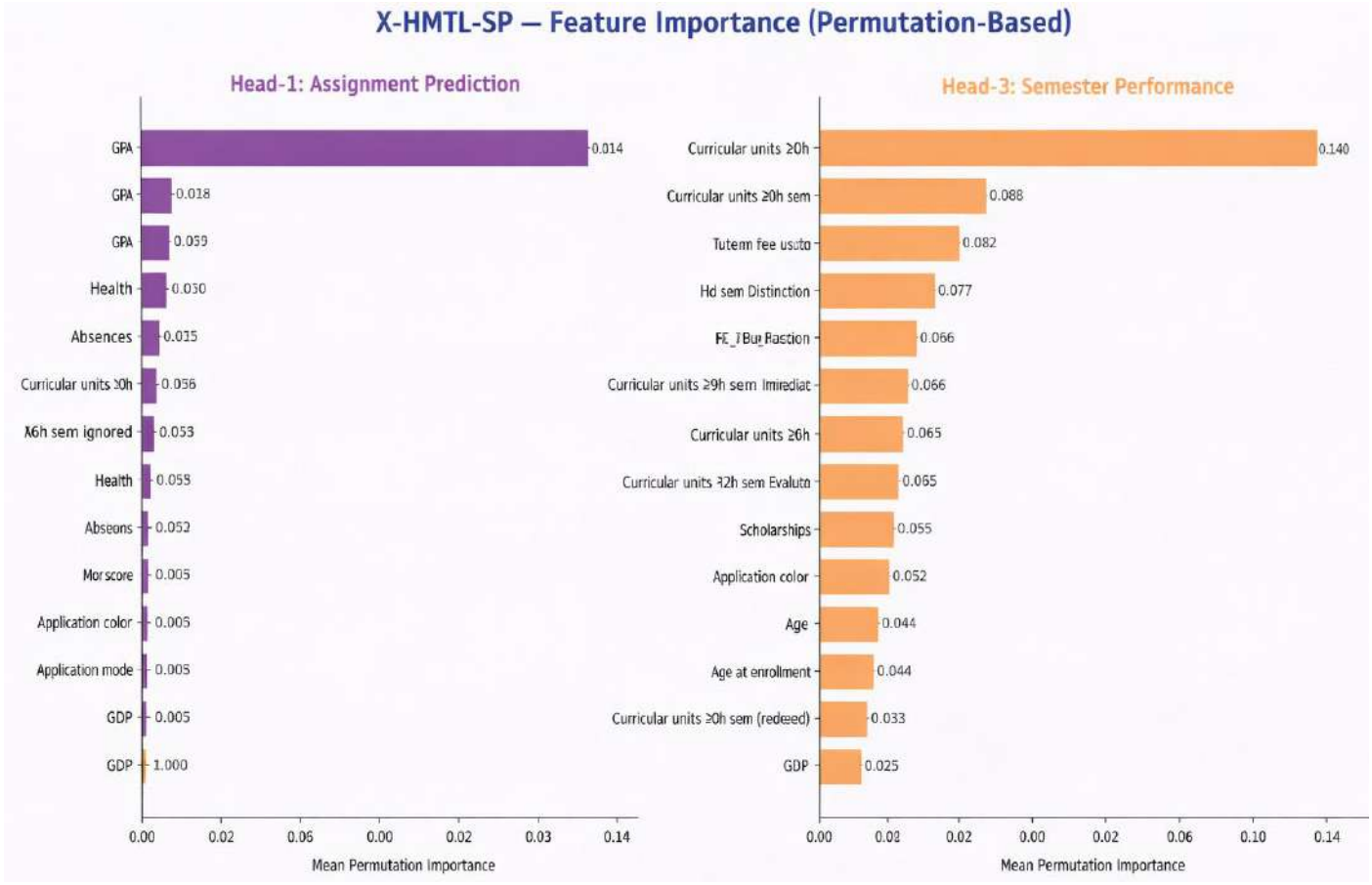

Figure 6: Permutation Feature Importance: Head-1 (Assignment, left) and Head-3 (Semester, right)

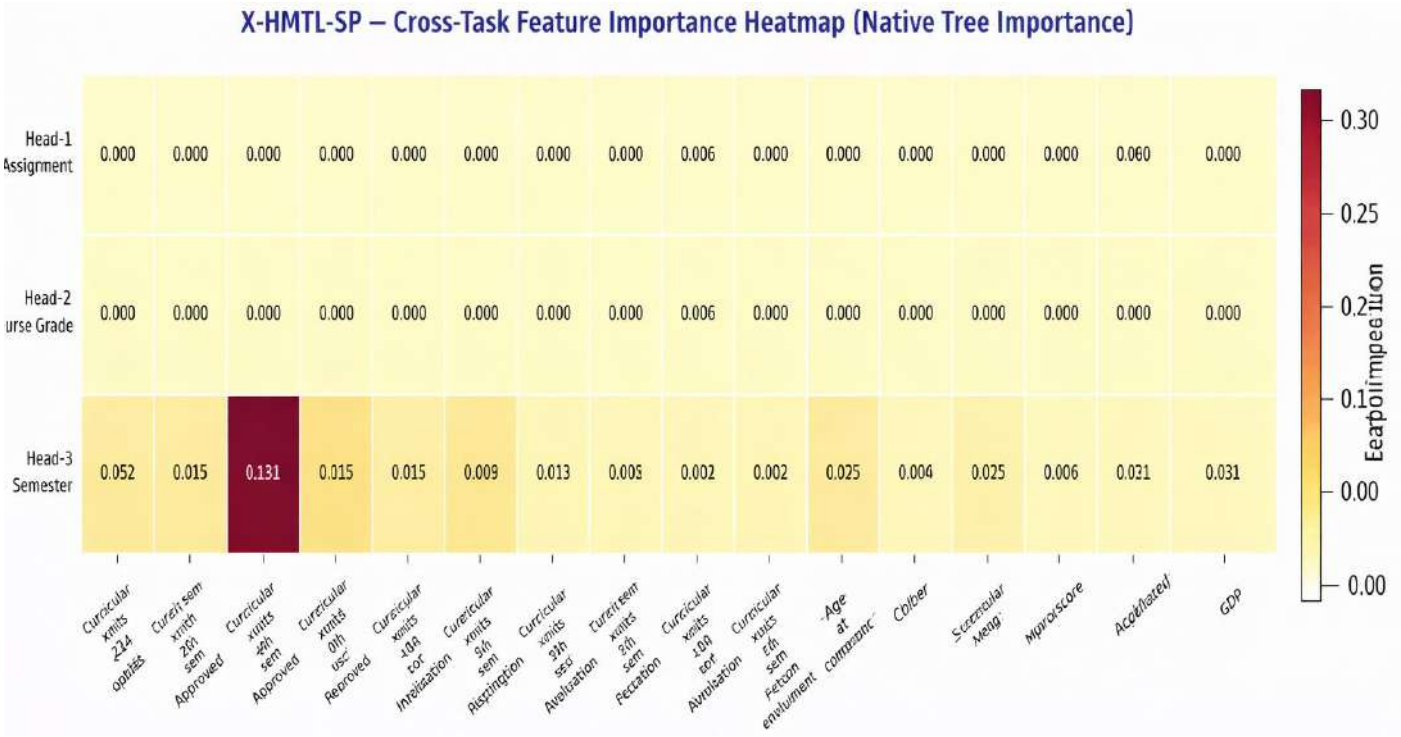

Figure 7: Cross-task feature importance heatmap — native tree importance across all three heads

The cross-task heatmap (Figure 7) reveals three patterns: (i) curricular units approved and grade features are universally important; (ii) student lifestyle features (G1, absences, health) are disproportionately important for Head-1; and (iii) institutional variables (tuition fees, scholarship) gain increasing importance at the semester level.

V. DISCUSSION

A. Strengths

The experimental results confirm that hierarchical MTL with explicit knowledge transfer yields consistently strong performance. Near-perfect lower-level prediction (99.82% and 100%) ensures reliable knowledge propagation to Head-3. The 81.26% semester accuracy represents strong performance given the Enrolled class boundary ambiguity.

The direct appearance of Head-1 probability features among the top-6 predictors for semester outcomes is the most significant finding: *assignment-level predicted performance carries independent predictive signal for semester outcomes* beyond raw grade scores, validating the core architectural hypothesis.

From a practice perspective, PFI results are directly actionable: identifying curricular unit approvals and tuition fee status as the strongest semester-level predictors enables academic advisors to target interventions on students failing to complete assignments and experiencing financial difficulties.

B. Comparison with Related Approaches

The stacked ensemble of [14] — RF + XGBoost base learners with LR meta-learner — represents the most directly comparable prior approach for single-level outcome prediction. X-HMTL-SP extends this paradigm to three hierarchically related levels simultaneously. The hierarchical chaining differs fundamentally from standard stacking: task-level probabilistic outputs encode domain-meaningful intermediate representations (assignment performance, course grade confidence) rather than arbitrary base learner predictions, yielding interpretable inter-task knowledge transfer.

C. Limitations and Future Work

The Enrolled class remains the primary limitation (F1 = 0.4861), attributable to transitional ambiguity. Longitudinal tracking across multiple semesters would provide temporal context to disambiguate enrolled students more effectively.

Dataset merge introduces structured missingness (47 of 69 features applicable across all sources). Future work will explore domain adaptation and unified educational data ontologies for improved cross-dataset feature alignment.

A neural variant of X-HMTL-SP employing shared transformer encoders with task-specific attention heads would enable end-to-end joint optimisation. Complementing global PFI with SHAP-based local explanations would support personalised student-level intervention recommendations. Real-time deployment as a student monitoring dashboard represents the primary practical target for future work.

VI. CONCLUSION

This paper presented X-HMTL-SP, an Explainable Hierarchical Multi-Task Learning framework for Multi-Level Student Performance Prediction. By explicitly modelling the hierarchical dependency chain assignment → course grade → semester outcome through sequential probabilistic knowledge transfer, X-HMTL-SP achieves near-perfect assignment prediction (99.82%), perfect course grade classification (100%, $R^2 = 0.9993$), and strong semester performance prediction (81.26%) with integrated explain ability.

The direct empirical validation of hierarchical knowledge transfer — through the appearance of Head-1 probability outputs among the top-6 most important features for semester prediction — represents a novel contribution to both educational data mining and multi-task learning research. The framework is computationally efficient, fully reproducible, and extensible to additional prediction tasks, supporting scalable deployment in real-world educational intelligence systems.

Future directions include longitudinal data integration, deep learning variants with shared encoders, SHAP-based local explanations, and real-time academic monitoring system deployment.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

DATA AVAILABILITY

All datasets are publicly available from the UCI Machine Learning Repository. The X-HMTL-SP implementation and merged dataset are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

All authors contributed to conceptualisation, methodology, analysis, and manuscript preparation.

REFERENCES

[1] Romero, C., and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern. C (Appl. Rev.)*, vol. 40, no. 6, pp. 601–618, 2010. Available from: https://doi.org/10.1109/TSMCC.2010.2053532

[2] Baker, R., and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Educ. Data Min.*, vol. 1, no. 1, pp. 3–17, 2009. Available from: https://doi.org/10.5281/zenodo.3554657

[3] Peña-Ayala, A., "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1432–1462, 2014. Available from: https://doi.org/10.1016/j.eswa.2013.08.042

[4] Kotsiantis, S., C. Pierrakeas, and P. Pintelas, "Predicting students' performance in distance learning using machine learning," *Appl. Artif. Intell.*, vol. 18, no. 5, pp. 411–426, 2004. Available from: https://doi.org/10.1080/08839510490442058

[5] Márquez-Vera, C., C. R. Morales, and S. V. Soto, "Predicting school failure and dropout by using data mining techniques," *IEEE Rev. Iberoam. Tecnol. Aprendizaje*, vol. 8, no. 1, pp. 7–14, 2013. Available from: https://ieeexplore.ieee.org/abstract/document/6461622

[6] Breiman, L., "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001. Available from: https://link.springer.com/chapter/10.1007/978-0-387-84858-7_15

[7] Chen, T., and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD*, 2016, pp. 785–794. Available from: https://doi.org/10.1145/2939672.2939785

[8] Caruana, R., "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997. Available from: https://link.springer.com/article/10.1023/a:1007379606734

[9] Zhang, Y., and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5586–5609, 2022. Available from: https://ieeexplore.ieee.org/abstract/document/9392366

[10] Liu, X., P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," in *Proc. ACL*, 2019, pp. 4487–4496. Available from: https://aclanthology.org/P19-1441/

[11] Doshi-Velez, F., and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017. Available from: https://doi.org/10.48550/arXiv.1702.08608

[12] Fisher, A., C. Rudin, and F. Dominici, "All models are wrong, but many are useful," *J. Mach. Learn. Res.*, vol. 20, no. 177, pp. 1–81, 2019. Available from: https://link.springer.com/chapter/10.1007/978-3-030-65965-3_28

[13] Jayaprakash, S. M., E. W. Moody, E. J. Lauría, J. R. Regan, and J. D. Baron, "Early alert of academically at-risk students: An open-source analytics initiative," *J. Learn. Anal.*, vol. 1, no. 1, pp. 6–47, 2014. Available from: https://learning-analytics.info/index.php/JLA/article/view/3249

[14] Wolpert, D. H., "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992. Available from: https://doi.org/10.1016/S0893-6080(05)80023-1

[15] Sekeroglu, B. *et al.*, "Student performance prediction and classification using machine learning algorithms," in *Proc. EMDL*, 2019. Available from: https://doi.org/10.1145/3318396.3318419

[16] Aulck, L., N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," 2016. Available from: https://doi.org/10.48550/arXiv.1606.06364

[17] Ma, J., Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proc. ACM SIGKDD*, 2018, pp. 1930–1939. Available from: https://doi.org/10.1145/3219819.3220007

[18] Lundberg, S. M., and S. I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017. Available from: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[19] Ribeiro, M. T., S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144. Available from: https://doi.org/10.1145/2939672.2939778

[20] Rudin, C., "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, pp. 206–215, 2019. Available from: https://www.nature.com/articles/s42256-019-0048-x