Journal of Advanced College of Engineering and Management

When Should Educational AI Stay Silent? Abstention-Aware Student Risk Prediction Using Learning Analytics

[1*]Gautam Sitam, [2]Katuwal Chhetri Anusha
[1,2]*Master of Information Technology (Professional), Deakin University, Melbourne, Australia*
Email: gautamsitam5@gmail.com, anusha.katuwal.7@gmail.com

DOI: 10.3126/jacem.v12i01.93932

Abstract
Universities increasingly rely on machine learning systems to identify students at risk of academic failure or withdrawal, where such predictions directly influence intervention decisions, resource allocation, and students' academic trajectories. However, in these high-stakes educational settings, conventional models produce predictions for all students regardless of confidence, which can result in incorrect risk identification, unnecessary student anxiety, and misallocation of academic support. This study introduces an abstention-aware prediction approach that allows the model to withhold decisions when confidence is insufficient. The approach is evaluated using the Open University Learning Analytics Dataset (OULAD). Logistic regression is adopted due to its well-calibrated probabilistic outputs, achieving a ROC-AUC of 0.889. Comparisons with Random Forest and XGBoost show that while these models achieve slightly higher discrimination, logistic regression provides more consistent probability estimates for uncertainty-aware prediction. The results demonstrate a clear risk–coverage trade-off, where uncertain cases are deferred while confident predictions remain accurate. The findings support the use of selective prediction as a practical mechanism for improving the reliability of AI-assisted decision-making in education.

Keywords*—Learning Analytics, Student Risk Prediction, Abstention-Aware Classification, Trustworthy AI, Educational Data Mining, Probability Calibration, Human-in-the-Loop Systems, Predictive Analytics in Education*

1. INTRODUCTION

Predictive analytics is widely used in education to identify students at risk of poor performance or withdrawal. With the growth of online and blended learning, large volumes of data such as assessment results and interaction records are available through learning systems. Analysing this data supports early identification of at-risk students and enables timely intervention [1].

The Open University Learning Analytics Dataset (OULAD) has been widely used for this task, with a range of machine learning models applied to predict student outcomes [2][3]. Most existing work focuses on improving predictive accuracy. However, in practical

settings, the reliability of individual predictions is equally important. Predictions made with low confidence can lead to inappropriate interventions or missed support.

Conventional models produce predictions for all students, even when the model is uncertain. This highlights the need for approaches that can account for uncertainty and avoid unreliable decisions.

This study proposes an abstention-aware approach for student risk prediction, where uncertain predictions are deferred rather than forced into a decision. By using predicted probabilities to separate confident and uncertain cases, the approach supports more reliable decision-making.

This study makes the following contribution: it shows that prediction uncertainty can be used to improve decision reliability in student risk prediction, without requiring more complex models**.**

2. PROBLEM STATEMENT

Learning analytics has become a significant area of research, particularly with the aim of using learning data to improve student success and decision-making for learning institutions [17]. Predictive models are increasingly used to identify students at risk of academic failure or dropout, supporting early intervention strategies.

However, most existing models produce predictions for all students without considering prediction uncertainty. This can lead to unreliable decisions, especially for cases near the decision boundary where the model has low confidence.

Although abstaining from low-confidence predictions can reduce error [10], this idea is not commonly applied in student risk prediction.

There is, therefore, a need for an approach that not only predicts student risk but also identifies when predictions should be deferred. This enables more reliable decisions by avoiding uncertain predictions.

3. LITERATURE REVIEW

Prior research in learning analytics has largely focused on improving predictive accuracy using traditional machine learning, ensemble methods, and deep learning approaches. Logistic regression remains widely used due to interpretability, while tree-based ensembles often achieve higher raw accuracy. Parallel work in trustworthy AI highlights that modern models are frequently overconfident and poorly calibrated. Selective prediction and reject-option classification provide theoretical mechanisms for abstention, yet their application in practical student risk prediction systems using real educational data remains limited.

The literature therefore reveals a clear gap: accuracy-centric models dominate learning analytics, while uncertainty-aware decision frameworks remain underexplored in real educational datasets.

A. Learning Analytics and Student Risk Identification

Learning analytics has become an important research area that focuses on examining educational data to improve learning processes and decision-making at institutions. Systematic reviews show the rapid growth of predictive modeling techniques used in digital learning environments [1]. The availability of structured educational datasets, like the Open University Learning Analytics Dataset (OULAD), has further supported reproducible research and comparisons in predicting student performance [2].

A significant amount of research has shown that behavioral indicators, assessment results, and engagement metrics from virtual learning environments can help identify students at academic risk [3]. These predictive systems aim to support early intervention strategies by recognizing patterns linked to dropout or failure. However, along with technical advancements, ethical and governance issues have arisen. Concerns about transparency, fairness, informed consent, and the effects of automated labeling have been widely discussed in the learning analytics field [4]. Misclassification can negatively impact student confidence, institutional trust, and resource allocation. Despite these concerns, many predictive systems assume that every student must receive a risk label, regardless of the prediction confidence. So, although predictive performance has improved, the reliability of individual predictions has not been fully explored. In particular, existing systems rarely distinguish between confident and uncertain predictions, treating all outputs as equally reliable.

B. Machine Learning Methods for Student Performance Prediction

Different supervised learning techniques have been used to model student risk. Logistic regression is commonly applied due to its ease of understanding and probabilistic output, making it suitable for decision-making at institutions [3]. Ensemble methods like Random Forest [5] and gradient boosting techniques such as XGBoost [6] have shown enhanced predictive accuracy by modeling nonlinear relationships and interactions among features. Deep learning methods have also been introduced to capture learning dynamics over time. Knowledge Tracing models, especially those based on neural networks, can model changing mastery patterns [7]. These methods have demonstrated strong performance in large online learning environments.

Despite methodological differences, these models generally focus on improving overall classification accuracy and make predictions for all cases. These models typically lack mechanisms to defer predictions in cases of low confidence, resulting in forced

decisions even when uncertainty is high. In addition, complex models may be overly confident or hard to interpret, which can hinder responsible use in education.

C. Uncertainty and Selective Prediction in Machine Learning

In the broader realm of machine learning, estimating uncertainty has become essential for creating reliable AI systems. Bayesian approximation methods, like Monte Carlo dropout, have been suggested to measure predictive uncertainty in neural networks [8]. Furthermore, calibration studies have shown that modern neural networks often make overly confident predictions, requiring techniques to improve probability reliability [9].

Selective prediction, also known as classification with a reject option, provides a structured way for models to abstain when confidence is low. Chow [10] introduced the theoretical basis for reject-option classification, which formalizes the balance between prediction coverage and error. More recent research has applied selective classification principles to deep neural networks, showing that abstaining can significantly lower risk among accepted predictions [11].

Conformal prediction offers another uncertainty-aware approach by creating prediction sets with guaranteed statistical validity under minimal assumptions about distributions [12]. While these uncertainty-aware methods have been studied in high-stakes fields like healthcare and autonomous systems, their systematic use in learning analytics is still limited. However, many of these approaches focus on complex models or theoretical guarantees and are not directly designed for simple, interpretable student risk prediction systems used in practice.

D. Research Gap

Current research in learning analytics has mainly focused on improving predictive accuracy, with less attention given to how uncertainty is handled in individual predictions. In practice, most models produce predictions for all students without considering prediction confidence, which can lead to unreliable decisions in uncertain cases. This is particularly important in educational settings, where incorrect predictions may result in inappropriate interventions.

There is a need for simple and practical approaches that can distinguish between confident and uncertain predictions and allow uncertain cases to be deferred. This study addresses this by introducing an abstention-aware approach that improves the reliability of student risk prediction.

4. METHODOLOGY

This study proposes a trustworthy student risk prediction approach with an abstention mechanism. The objective is not only to classify students as at-risk or safe, but also to allow the model to abstain from prediction when confidence is uncertain, improving the reliability of educational decision support. This approach is suitable for educational decision support, where uncertain predictions should not be treated as definitive decisions and may require human review.

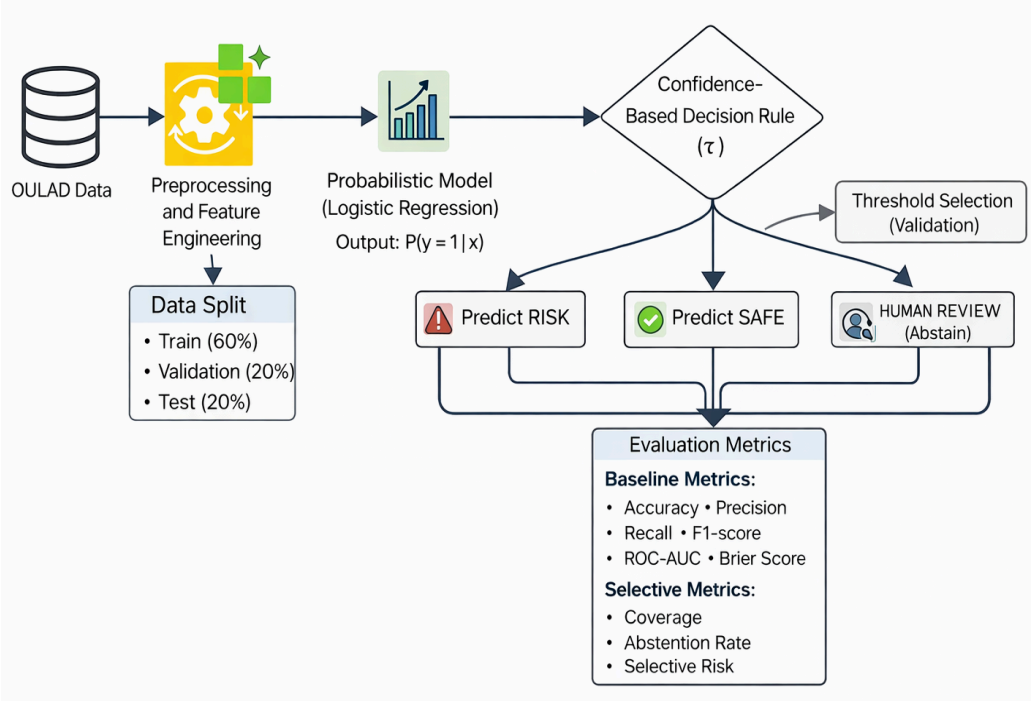

Figure 1: Overall architecture of the abstention-aware student risk prediction framework

Figure 1 illustrates the overall architecture of the proposed abstention-aware student risk prediction framework. The process begins with OULAD data input, followed by preprocessing and feature engineering. The proposed pipeline consists of data integration, preprocessing, probabilistic modelling, and confidence-based selective decision making.

The system outputs three possible decisions: RISK, SAFE, or HUMAN REVIEW. A logistic regression model then estimates the probability of student risk, after which a confidence-based abstention rule determines whether to predict risk, predict no-risk, or abstain for human review. Finally, model performance is evaluated using both standard and abstention-aware metrics to ensure reliable and trustworthy decision support. The proposed abstention-aware student risk prediction framework follows a structured pipeline comprising the following stages.

A. Dataset Input

This research uses the Open University Learning Analytics Dataset (OULAD), which contains data for 32,593 students across multiple module presentations. Instead of using a single file, several structured tables were combined to create the dataset used for modelling. The main files used are studentInfo.csv, studentAssessment.csv, assessments.csv, and studentVle.csv [2].

The studentInfo.csv file provides demographic information and final academic outcomes. The studentAssessment.csv and assessments.csv files provide assessment results along with their corresponding weights. The studentVle.csv file captures student interaction data within the Virtual Learning Environment, reflecting engagement behaviour [2].

These tables were merged using common identifiers such as student ID, module code, and presentation code. From this, a student-level dataset was constructed that combines demographic, performance, and engagement features.

For modelling, the original four-class outcome (Pass, Fail, Withdraw, Distinction) was converted into a binary variable. Students with outcomes ‘Fail’ or ‘Withdraw’ were labelled as At-Risk (1), while those with ‘Pass’ or ‘Distinction’ were labelled as Not At-Risk (0). This formulation aligns with the goal of identifying students who may require early intervention

B. Data Preprocessing and Feature Engineering

To ensure reliable probabilistic predictions, the student data obtained from the OULAD dataset [2] was carefully prepared before model training. Preprocessing was performed to improve data quality and ensure consistency across features, which is essential for stable probability estimation.

Categorical variables such as gender, region, and highest education level were converted into numerical form using one-hot encoding, enabling the model to process them effectively [18]. Continuous features, including assessment scores and VLE interaction counts, were normalised to a common scale so that variables with larger numerical ranges would not disproportionately influence the model [19]. Missing values were replaced with zero values. This is appropriate for activity-based features, where missing entries can indicate no interaction and ensures consistency across observations [20].

These preprocessing steps ensured that the model received structured and consistent inputs, allowing stable and meaningful probability estimates, which are critical for confidence-based abstention decisions.

To minimise data leakage and enhance generalisation performance, the dataset was split into training, validation, and testing subsets using stratified sampling to preserve class distribution. Specifically, 60% of the data was used for training, 20% for validation,

and 20% for testing. The validation set was used for threshold selection, while the test set was reserved for final evaluation. Random seeds were fixed to maintain reproducibility of results.

Model hyperparameters were selected through cross-validation conducted solely on the training data. Abstention thresholds were determined using validation data to prevent bias in the final evaluation. All experiments were implemented in Python using scikit-learn, pandas, NumPy, and Matplotlib, and the analysis was performed in a Jupyter Notebook environment.

C. Probabilistic Classification Model

For risk estimation, we use logistic regression, a well-established probabilistic classifier for binary outcomes. The model estimates the probability that a student is at risk based on input features, rather than directly producing a class label.

The model estimates:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

where:

- x represents student features
- w denotes learned coefficients
- b is the bias term
- σ is the logistic sigmoid function

Here, x is vector of student features, w is learned weight vector and b is bias. The model produces probability estimates for each student, which are used to assess prediction confidence and support the abstention decision process.

The sigmoid function maps the model output to a value between 0 and 1. Values close to 1 indicate a high likelihood of risk, while values close to 0 indicate a low likelihood. Values around 0.5 indicate uncertainty, where the model is not clearly confident.

In standard classification, this probability is converted into a class label using a fixed threshold, which forces the model to make a decision for every student. However, predictions near this boundary are less reliable. In this study, the probability is used to identify such uncertain cases, which are handled differently rather than being forced into a decision [15] [16].

D. Confidence-Based Abstention Rule and Threshold Selection

To improve the reliability of student risk prediction, this study introduces a confidence-based abstention rule that allows the model to remain silent when it is not

sufficiently certain about its decision. Instead of forcing a prediction for every student, the model examines the predicted probability of risk.

A confidence threshold τ (tau) is introduced to control when predictions should be made or deferred. The rule is:

$$\hat{y} = \begin{cases} 1, & P(y = 1 \mid x) > 1 - \tau \\ 0, & P(y = 1 \mid x) < \tau \\ \text{abstain}, & \text{otherwise} \end{cases}$$

Here,

- x is student features (scores, activity, demographics)
- y is student outcome (1 = at risk, 0 = not at risk)
- P (y = 1 ∣ x) is predicted probability student is at risk
- τ is confidence boundary controlling abstention
- y^ is final decision (risk / safe / abstain)

This decision rule is based on the idea of selective prediction, where the model is allowed to abstain when prediction confidence is low [15][16]. The decision process is defined as follows:

- If $P(y = 1|x) \geq 1 - \tau$, the model predicts RISK
- If $P(y = 1|x) \leq \tau$, the model predicts SAFE
- If $\tau < P(y = 1|x) < 1 - \tau$, the model returns HUMAN REVIEW instead of making a prediction.

This rule divides predictions into three zones:

1. Predict Risk (1): If the probability of risk is very high, the model is confident that students are at risk.
2. Predict no-risk (0): If the probability of risk is very low, the model is confident that the student is safe.
3. Abstain (stay silent): If probability lies between these two confidence boundaries or thresholds, the model abstains from making an automatic decision and the case can be reviewed by a human advisor.

The threshold is restricted to the range $0 < \tau < 0.5$, which creates two symmetric decision boundaries around the midpoint. This ensures consistent treatment of both high-risk and low-risk predictions.

The region between τ and $1 - \tau$ represents uncertainty, where predictions are close to the decision boundary (around 0.5). In this region, small changes in input can change the predicted class, making these predictions less reliable.

By excluding predictions in this uncertain region, the model reduces the likelihood of incorrect classifications. As a result, the predictions that are accepted are more accurate, improving overall reliability.

The threshold value controls how strict the model is. Lower values allow more predictions but include uncertain cases, while higher values result in fewer but more reliable predictions.

In this study, τ is selected using validation data by evaluating multiple values and selecting a balance between coverage and selective risk. This indicates a trade-off between coverage and reliability, where higher reliability is achieved by allowing the model to abstain in low-confidence cases.

E. Evaluation using abstention and standard metrics

Model performance is evaluated in two stages: first using standard classification metrics without abstention and then using selective metrics after applying the abstention rule.

For the baseline model, performance is assessed using accuracy, precision, recall, F1-score, ROC-AUC, and Brier score. These metrics capture both classification performance and the quality of predicted probabilities.

After applying the abstention rule, evaluation is performed only on accepted predictions, excluding cases assigned to HUMAN REVIEW. Accuracy, precision, recall, and F1-score are computed on this subset to measure the quality of decisions actually made by the model. Selective accuracy is used to reflect performance only on non-abstained cases. To quantify the effect of abstention, coverage and abstention rate are used to measure how many predictions are made versus deferred. Selective risk, defined as the error ($1 -$ accuracy) on accepted predictions, is used to capture the remaining error among accepted cases.

The effect of the threshold τ is analysed by examining how selective risk changes as coverage varies across different values of τ. A decrease in selective risk with reduced coverage indicates that uncertain predictions are being effectively removed.

In addition to logistic regression, Random Forest and XGBoost are evaluated to compare predictive performance, and the behaviour of probability estimates across models. Logistic regression is used for the abstention mechanism due to its stable probability outputs, while the other models are included for comparison.

Since the abstention mechanism depends on predicted probabilities, calibration is also evaluated. Brier score and reliability diagrams are used to assess how well predicted

probabilities match observed outcomes. Well-calibrated probabilities are important to ensure that the threshold-based decision rule correctly distinguishes between confident and uncertain predictions.

Experiments are conducted on the OULAD dataset, which contains over 32,000 student records, providing a realistic and sufficiently large-scale evaluation setting. All data are anonymized and used in accordance with OULAD data usage policies [2][4].

5. RESULTS AND ANALYSIS

A. Baseline Classification Performance

The baseline(No Abstention**)** performance is evaluated without applying the abstention mechanism to establish the predictive capability of each model.

Table 1: Baseline Classification Performance of Student Risk Prediction Models

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**	**Brier**
XGB	0.813	0.835	0.805	0.82	0.899	0.129
LR	0.801	0.825	0.791	0.808	0.889	0.134
RF	0.798	0.817	0.795	0.806	0.885	0.139

Table 1 presents a comparison of classification performance across logistic regression, Random Forest, and XGBoost models. XGBoost achieves the highest classification performance across most metrics, including accuracy (0.813) and F1-score (0.820). Logistic regression performs slightly lower, while Random Forest shows comparable but slightly weaker results. However, the differences in classification performance are relatively small. This indicates that all models are capable of learning meaningful patterns from the data, and that further improvements should focus on decision reliability rather than accuracy alone.

From a probability perspective, the Brier score provides insight into calibration quality. XGBoost achieves the lowest Brier score, followed by logistic regression and Random Forest. Despite this, calibration behaviour must also be considered alongside numerical scores. Logistic regression produces more stable and monotonic probability estimates, while ensemble models tend to exhibit overconfidence in certain regions. This makes logistic regression more suitable for confidence-based abstention, where reliable probability estimates are required.

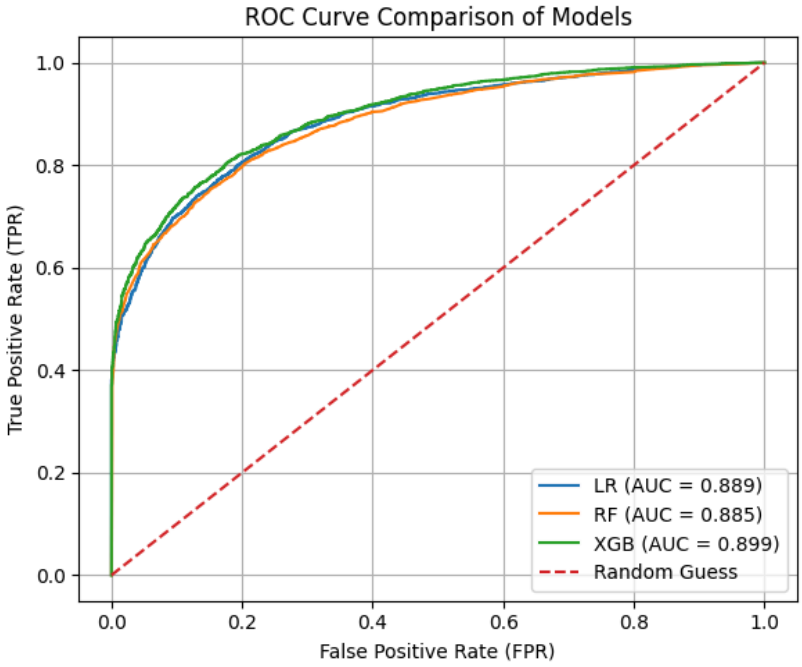

Figure 2: ROC Curves for Baseline Student Risk Prediction Models

Figure 2 presents the ROC curves of the three models. The ROC curves show that all models achieve strong discrimination performance, with curves closely aligned across the full range of false positive rates. XGBoost attains the highest ROC-AUC (0.899), followed by logistic regression (0.889) and Random Forest.

The similarity of the curves indicates that all models have comparable ability to rank at-risk and non-risk students. The performance differences, while measurable, are relatively small and do not significantly alter the decision boundary.

This observation is important because it suggests that improvements in classification accuracy alone are limited. Instead, the primary challenge lies in handling uncertainty in predictions, particularly near the decision boundary.

As a result, the focus shifts from improving predictive performance to improving decision reliability, which is addressed in the following sections through the introduction of the abstention mechanism. However, the abstention mechanism depends on reliable probability estimates rather than only classification accuracy. Logistic regression provides more stable and well-calibrated probabilities, making it more suitable for threshold-based decision-making.

B. Probability Behaviour and Decision Uncertainty

To understand how predictions are formed, the relationship between model outputs and predicted probabilities is examined**.**

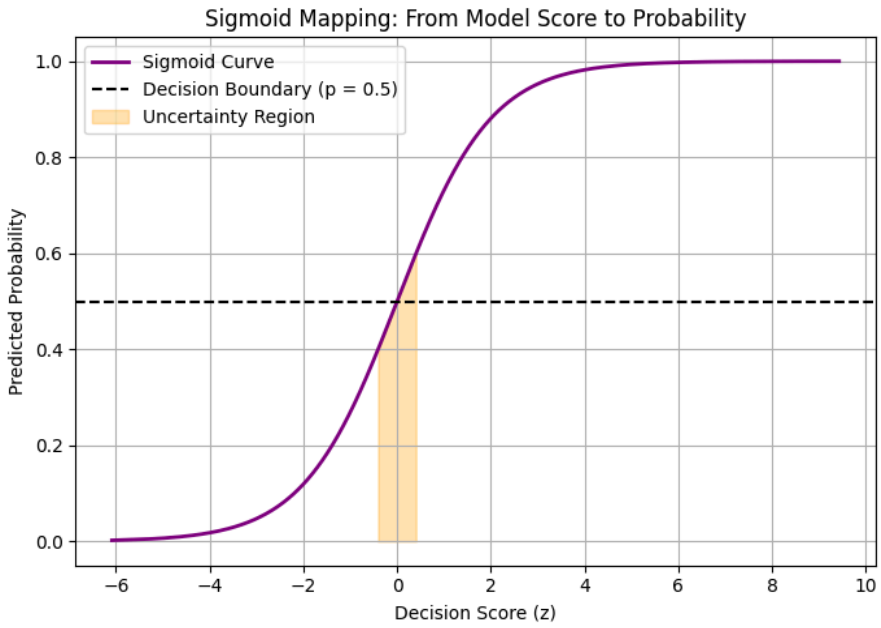

Figure 3: Sigmoid Mapping from Model Score to Predicted Probability

Figure 3 illustrates how decision scores are transformed into probabilities using the sigmoid function. The central region around probability 0.5 corresponds to the decision boundary. Predictions in this region are unstable, as small changes in input features can shift the predicted class.

From a decision-theoretic perspective, these predictions correspond to low-margin regions, where the model has limited separation between classes. Such low-margin predictions are inherently uncertain and contribute disproportionately to classification errors.

This behaviour is critical for the proposed approach, as it identifies the region where predictions are least reliable. The abstention mechanism is designed to operate on this region by deferring decisions when the predicted probability lies close to the decision boundary.

C. Calibration and Reliability of Probability Estimates

Since the proposed abstention mechanism depends directly on predicted probabilities, it is necessary to assess how well these probabilities reflect actual outcomes.

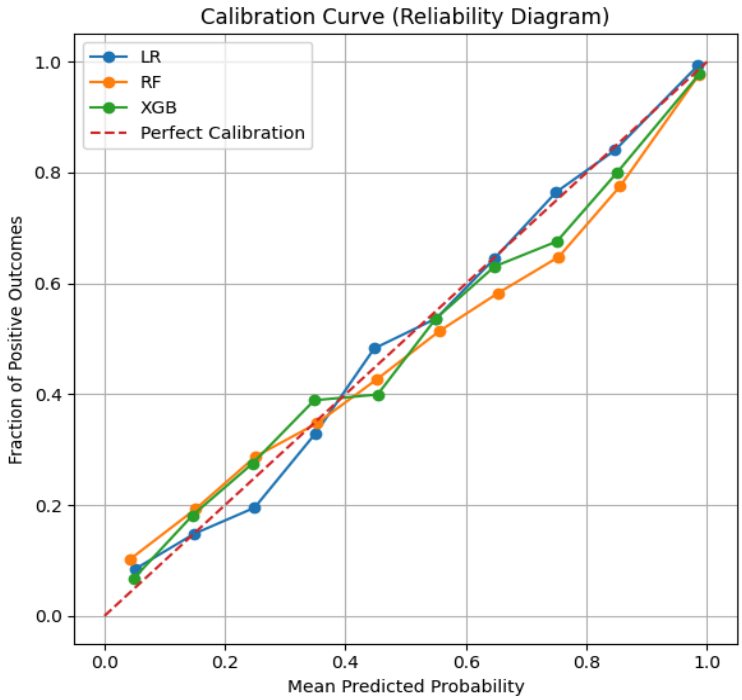

Figure 4: Reliability Diagram for Probability Calibration

Figure 4 presents the reliability diagram comparing predicted probabilities with observed outcomes for logistic regression, Random Forest, and XGBoost. The curve for logistic regression lies closer to the diagonal, indicating better alignment between predicted probabilities and observed outcomes. In contrast, Random Forest and XGBoost show noticeable deviations, particularly in the mid-probability region. The Brier scores are:

- Logistic Regression: 0.134
- Random Forest: 0.138
- XGBoost: 0.129

Although XGBoost achieves a slightly lower Brier score, its calibration curve indicates overconfidence in certain probability ranges. Logistic regression, on the other hand, provides more consistent probability estimates across the full range.

This distinction is important because the abstention mechanism relies on probability thresholds to determine whether a prediction should be accepted or deferred. If probabilities are not well calibrated, the model may assign high confidence to incorrect predictions, reducing the effectiveness of abstention.

D. Threshold Selection and Effect of τ

The threshold τ controls which predictions are accepted, and which are deferred for human review.

During validation, multiple values of τare evaluated using F1-score to identify an initial candidate. This process results in a preliminary threshold of $\tau = 0.05$, which achieves very high accuracy but extremely low coverage.

However, selecting τ based solely on validation F1-score leads to overly conservative behaviour, where a large proportion of predictions are rejected. Therefore, the final threshold is selected based on the trade-off between coverage and selective risk, rather than accuracy alone.

a. Effect of τ on Performance

Table 2: Classification Performance and Coverage–Risk Trade-off across Threshold Values (τ)

τ (Threshold)	Accuracy	Precision	Recall	F1 Score	Coverage	Abstention Rate	Selective Risk
0.05	0.987	1	0.983	0.992	0.263	0.737	0.013
0.1	0.964	0.993	0.952	0.972	0.357	0.643	0.036
0.15	0.946	0.981	0.929	0.954	0.442	0.558	0.054
0.2	0.927	0.96	0.911	0.935	0.527	0.473	0.073
0.25	0.911	0.94	0.897	0.918	0.61	0.39	0.089
0.3	0.893	0.92	0.881	0.9	0.693	0.307	0.107
0.35	0.873	0.903	0.857	0.879	0.769	0.231	0.127
0.4	0.851	0.879	0.835	0.856	0.846	0.154	0.149
0.45	0.825	0.851	0.81	0.83	0.924	0.076	0.175

Table 2 presents the classification performance and abstention–coverage trade-off across different values of τ. The results show a clear pattern: as τ increases, coverage increases while accuracy decreases. This behaviour occurs because increasing τ expands the acceptance regions toward the decision boundary, where predictions are more uncertain. As a result, more predictions are made, but a larger proportion of them are error-prone.

b. Theoretical Interpretation of τ

The threshold τ defines two acceptance regions:

$$[0, \tau] \; and \; [1 - \tau, 1]$$

Predictions falling within these regions are considered sufficiently confident, while those between τ and $1 - \tau$ are rejected.

As τ increases, these acceptance regions expand toward the decision boundary, incorporating lower-confidence predictions. Conversely, smaller values of τ restrict predictions to high-confidence regions, improving reliability at the cost of coverage.

c. Selection of Optimal Threshold

The final operating point is selected as:

- $\tau = 0.20$
- Coverage = 52.7%
- Selective Risk = 0.073

Although lower values of τ achieve higher accuracy, they result in very low coverage and are not practical for decision-making. The selected value balances reliability and usability by maintaining acceptable error while ensuring that predictions are made for a sufficient proportion of students.

This value represents a balanced trade-off where the model avoids highly uncertain predictions while still maintaining practical coverage. Lower values of τ result in very high accuracy but impractically low coverage, while higher values increase coverage at the cost of reliability.

E. Risk-Coverage Trade-off

The relationship between coverage and selective risk is analysed to evaluate the effectiveness of the abstention mechanism in improving decision reliability.

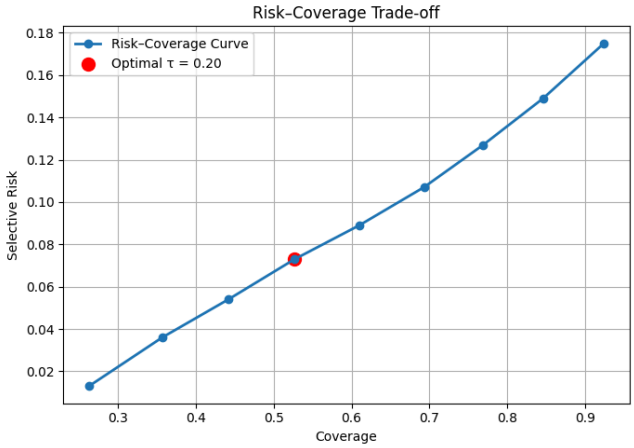

Figure 5: Risk-Coverage Trade-off for Abstention-Based Prediction

Figure 5 presents the risk-coverage curve, showing how selective risk changes as the proportion of accepted predictions increases.Selective risk increases steadily as

coverage increases. This occurs because higher coverage includes predictions closer to the decision boundary, which are more uncertain and more likely to be incorrect. This shows that abstention improves reliability by filtering out these error-prone predictions rather than modifying the model itself. At lower coverage levels, the model makes predictions only on high-confidence instances, resulting in very low error. As coverage increases, predictions closer to the decision boundary are included, leading to a gradual rise in selective risk.

The highlighted point at $\tau = 0.20$ corresponds to a coverage of approximately 52.7% and a selective risk of 0.073. This point represents the selected operating threshold, where the model maintains a balance between reliability and practical usability.

This trend reflects the behaviour of selective classification, where expanding the acceptance region introduces more uncertain and error-prone predictions. Conversely, restricting predictions to high-confidence regions reduces error but limits the number of decisions made.

The smooth and monotonic increase in the curve indicates that prediction uncertainty is well aligned with error. This confirms that the abstention mechanism effectively identifies unreliable predictions and improves overall decision quality without modifying the underlying model.

F. Decision Behaviour and Abstention Patterns

The distribution of predicted probabilities is analysed to understand how the abstention mechanism operates in practice.

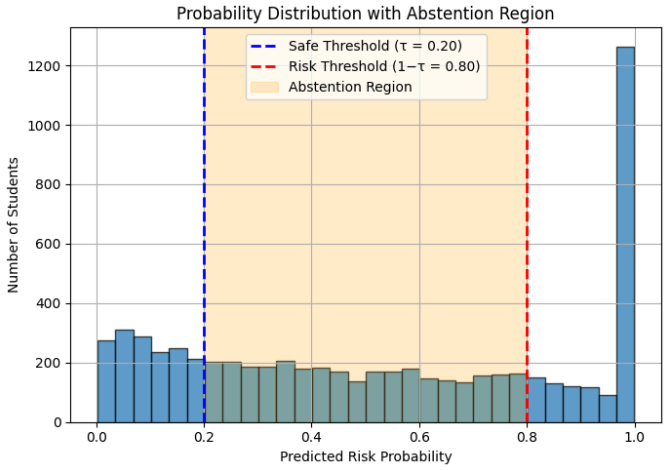

Figure 6: Distribution of Predicted Probabilities with Abstention Region

Figure 6 presents the distribution of predicted risk probabilities, along with the acceptance thresholds and abstention region.

Predictions are concentrated near 0 and 1, representing high-confidence cases where the model is more certain about its decisions. In contrast, the central region contains predictions with intermediate probabilities, indicating higher uncertainty. The thresholds at $\tau = 0.20$ and $1 - \tau = 0.80$ define the acceptance boundaries. Predictions below 0.20 are classified as SAFE, and those above 0.80 are classified as RISK, while predictions between these thresholds are deferred for further review.

The concentration of predictions in this middle region reflects the presence of uncertainty near the decision boundary, where small changes in input can alter the predicted class. This makes such predictions more error-prone, which justifies excluding them from automatic decisions.

Importantly, the abstention region aligns with this uncertain zone, indicating that the model selectively avoids making decisions where confidence is low. This shows that the abstention mechanism improves reliability by filtering out uncertain predictions rather than forcing a classification. Sample outputs further illustrate this behaviour.

Table 3: Sample Predictions Illustrating Abstention Decisions

Student ID	True Label	Predicted Probability	Decision
23252	0	0.157	SAFE
30849	1	0.991	RISK
21169	0	0.807	RISK
30521	0	0.269	HUMAN_REVIEW
29539	0	0.461	HUMAN_REVIEW
18908	0	0.057	SAFE
29351	1	0.573	HUMAN_REVIEW
11315	1	1	RISK
31609	1	0.756	HUMAN_REVIEW
25510	0	0.277	HUMAN_REVIEW

Predictions labelled as HUMAN REVIEW consistently correspond to mid-range probabilities, confirming that abstention is driven by uncertainty rather than random selection.

This behaviour is central to the proposed approach, as it demonstrates that the model does not attempt to force decisions in ambiguous cases. Instead, it defers such cases, thereby reducing the likelihood of incorrect predictions and improving overall decision reliability.

G. Error Reduction through Abstention

The effect of abstention on classification errors is examined using confusion matrices before and after applying the abstention mechanism.

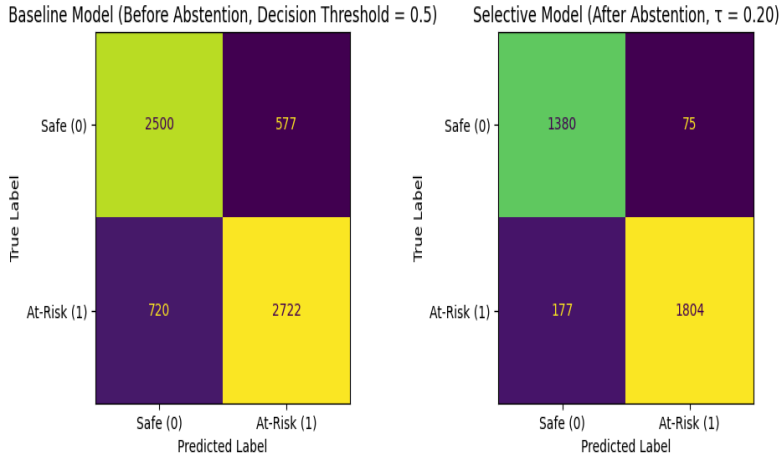

Figure 7: Confusion Matrix Comparison Before and After Abstention

Figure 7 compares the confusion matrices of the baseline model (without abstention) and the selective model (with $\tau = 0.20$). In the baseline model, a noticeable number of misclassifications occur, including 577 false positives (safe students predicted as at risk) and 720 false negatives (at-risk students predicted as safe).

After applying abstention, these errors are substantially reduced. In the selective model, false positives decrease to 75, and false negatives decrease to 177. This reduction is achieved by removing uncertain predictions rather than correcting them, which confirms that abstention improves reliability by avoiding high-risk decisions.

H. Generalisation and Model Stability

a. Feature Importance

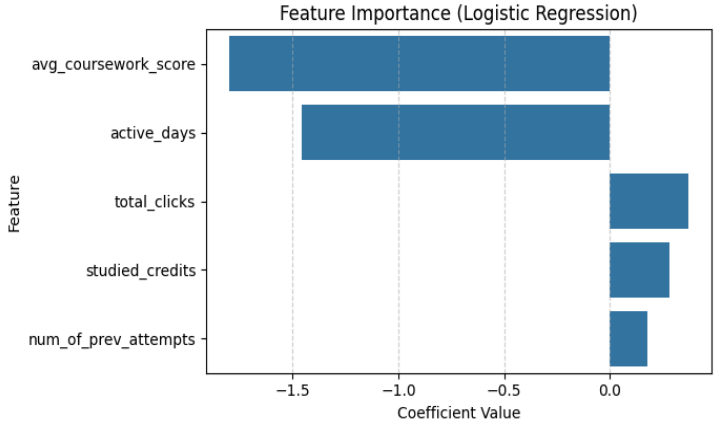

Figure 8: Logistic Regression Feature Importance

Figure 8 presents the feature importance derived from logistic regression coefficients. Assessment scores and engagement-related features contribute most strongly to prediction, while demographic attributes have a smaller influence. This indicates that the model relies primarily on behavioural and academic signals, which are more directly related to student performance outcomes. The absence of extreme coefficient values also suggests stable parameter estimation without dominance from any single feature.

b. Learning Curve Analysis

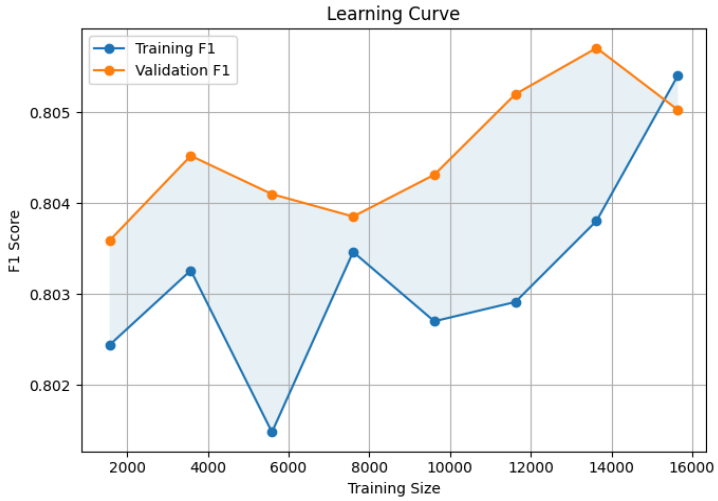

Figure 9: Learning Curves for Training and Validation Performance

Figure 9 presents the learning curves for training and validation F1-scores across increasing training sizes.The training and validation curves remain closely aligned across all sample sizes, with F1-scores consistently in the range of approximately 0.802 to 0.807. The gap between the two curves is minimal throughout, indicating that the model does not overfit the training data.

In typical overfitting scenarios, the training performance is significantly higher than the validation performance. In contrast, both curves here follow a similar trend, suggesting that the model is learning patterns that generalise well to unseen data. As the training size increases, the validation performance shows a slight improvement and then stabilises, while the training performance follows a similar pattern. This indicates that the model has captured the dominant structure in the dataset, and additional data provides only marginal improvement.

c. Cross-Validation Stability

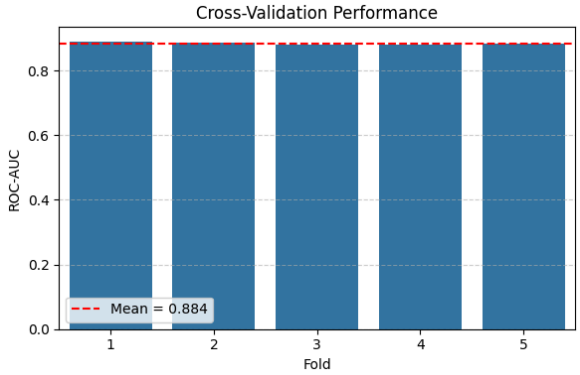

Figure 10: Cross-Validation Performance Across Data Splits

Figure 10 presents the ROC-AUC scores obtained from cross-validation across different data splits. The model achieves a mean ROC-AUC of approximately 0.884 with a very low standard deviation (0.0034), indicating consistent performance across folds. The low variation across splits confirms that the model's performance is not dependent on a particular train-test partition and that it generalises reliably across different subsets of the data.

Taken together, the feature importance, learning curve, and cross-validation results show that the model operates in a stable learning regime. The absence of overfitting, balanced feature contributions, and consistent performance across folds indicate that the observed improvements from the abstention mechanism reflect robust and generalisable behaviour.

d. Robustness under Noise and Feature Variation

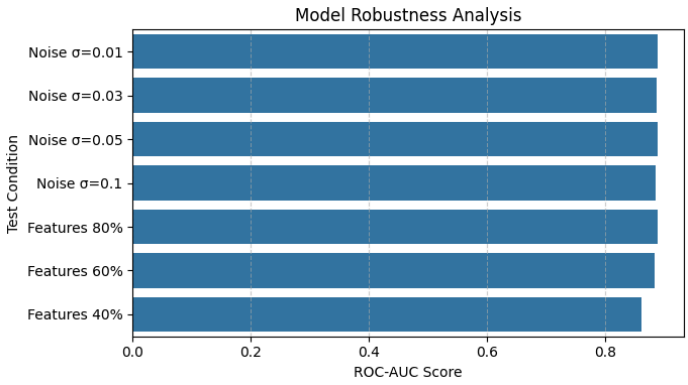

Figure 11: Model Robustness under Noise and Feature Variation

Figure 11 presents the model performance under different noise levels and feature availability conditions. The model maintains stable performance across all tested conditions, with ROC-AUC values remaining above approximately 0.85. When noise is introduced (σ = 0.01 to 0.1), only a slight decrease in performance is observed. Similarly, reducing the number of input features from 80% to 40% leads to a gradual decline, but the model continues to perform consistently. This indicates that the model is not highly sensitive to small perturbations in the data or moderate reductions in feature availability.

The stability across these conditions suggests that the model has learned robust patterns rather than relying on specific features or noise-free inputs. This is important for practical deployment, where data may be incomplete or noisy.

I. Uncertainty Quality Analysis

To evaluate whether predicted uncertainty reflects the reliability of model decisions, the average uncertainty is compared between correct and incorrect predictions.

Table 4: Uncertainty Comparison between Correct and Incorrect Predictions

Uncertainty Quality Analysis		
Model	**Avg Uncertainty (Correct Predictions)**	**Avg Uncertainty (Incorrect Predictions)**
LR	0.1686	0.3273
RF	0.1351	0.2842
XGB	0.1367	0

Table 4 presents the average uncertainty values for correct and incorrect predictions across different models. Across all models, incorrect predictions consistently show higher uncertainty than correct predictions. For example, in logistic regression, the average uncertainty increases from 0.1686 for correct predictions to 0.3273 for incorrect ones.

This pattern is also observed in Random Forest and XGBoost, where incorrect predictions are associated with noticeably higher uncertainty values.This indicates that uncertainty is not random, but systematically higher for error-prone predictions.This relationship is central to the proposed approach, as it allows the model to identify unreliable predictions using probability-based uncertainty. By deferring predictions with high uncertainty, the abstention mechanism reduces the likelihood of incorrect decisions.

6. DISCUSSIONS

The results show that errors are concentrated in uncertain regions near the decision boundary. By deferring these cases, the model reduces error among accepted predictions and improves reliability. Uncertainty is consistently higher for incorrect predictions, confirming that predicted probabilities provide a reliable signal for identifying such cases. This supports the use of threshold-based abstention. Although XGBoost achieves slightly higher ROC-AUC, logistic regression is more suitable due to its stable probability estimates, which are required for consistent decision thresholds. This shows that the improvement comes from handling uncertainty rather than increasing model complexity, which is important in educational settings where incorrect decisions can have practical consequences.

7. CONCLUSIONS

This study developed an abstention-aware student risk prediction framework that integrates probabilistic classification with confidence-based decision thresholds. The results demonstrate stable discrimination performance (ROC-AUC = 0.889) and balanced classification outcomes across varying threshold levels. The selective prediction mechanism effectively controls the trade-off between reliability and coverage. At moderate thresholds ($\tau = 0.20$), the model achieves strong predictive performance while maintaining responsible abstention behaviour. The framework successfully implements a human-in-the-loop approach, ensuring that uncertain cases are referred for review rather than forced into automated decisions. Overall, the study confirms that decision reliability can be improved by avoiding uncertain predictions rather than increasing model complexity, improving the practical applicability of AI in educational risk prediction.

8. SUGGESTIONS AND RECOMMENDATIONS

Future research should explore advanced uncertainty modelling techniques such as Bayesian learning, conformal prediction, and deep ensemble methods to further enhance abstention calibration. Additional studies may investigate real-time deployment within learning management systems to support early intervention strategies. Evaluating fairness, bias mitigation, and demographic sensitivity is recommended to ensure equitable AI decision-making. Expanding validation across multiple institutions and diverse educational contexts would strengthen generalisability. Further research may also incorporate temporal modelling to capture longitudinal student behaviour patterns. Integrating explainable AI techniques could improve transparency and educator trust in automated risk assessments.

ACKNOWLEDGMENTS

The authors would like to sincerely thank their academic supervisor for continuous guidance and support throughout this research. We are especially grateful to Dev Gurung, PhD candidate, for his valuable discussions, technical insights, and encouragement during the development of this work. We also thank the School of Information Technology at Deakin University for providing the necessary computational resources and research facilities. The Open University Learning Analytics Dataset (OULAD) is acknowledged for enabling the empirical evaluation of the proposed framework. Finally, we appreciate the constructive feedback from peers and reviewers, which helped improve the overall quality of this study.

REFERENCES

[1] Z. Papamitsiou and A. A. Economides, "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence," *Educational Technology & Society*, vol. 17, no. 4, pp. 49–64, 2014.

[2] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Scientific Data*, vol. 4, Art. no. 170171, 2017, doi: 10.1038/sdata.2017.171.

[3] A. Alamri, M. Watson, and M. Watson, "Using learning analytics to predict at-risk students in online higher education," *Sustainability*, vol. 11, no. 21, p. 5958, 2019.

[4] S. Slade and P. Prinsloo, "Learning analytics: Ethical issues and dilemmas," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1510–1529, 2013.

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794.

[7] C. Piech *et al*., "Deep knowledge tracing," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015, pp. 505–513.

[8] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Machine Learning (ICML)*, 2016, pp. 1050–1059.

[9] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, 2017, pp. 1321–1330.

[10] C. K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, Jan. 1970, doi: 10.1109/TIT.1970.1054406.

[11] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4878–4887.

[12] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York, NY, USA: Springer, 2005.

[13] J. J. Levy and A. J. O'Malley, "Don't dismiss logistic regression: The case for sensible extraction of interactions in the era of machine learning," *BMC Medical Research Methodology*, vol. 20, no. 1, pp. 1–12, 2020.

[14] Y. Hua, T. S. Stead, and M. R. Ganti, "Clinical risk prediction with logistic regression: Best practices and applications," *Academic Medicine & Surgery*, 2025.

[15] T. Srinivasan, J. Hessel, and Y. Choi, "Selective 'selective prediction': Reducing unnecessary abstention in language models," in *Findings of the Association for Computational Linguistics (ACL)*, 2024.

[16] A. Shah, Y. Bu, and G. W. Wornell, "Selective regression under fairness criteria," in *Proc. 39th Int. Conf. Machine Learning (ICML)*, 2022, pp. 19548–19562.

[17] G. Siemens, "Learning analytics: The emergence of a discipline," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1380–1400, 2013, doi: 10.1177/0002764213498851.

[18] J. Zhou, J. Li, and P. Liu, "Categorical data encoding in machine learning: A survey," *Journal of Big Data*, vol. 10, no. 1, pp. 1–27, 2023, doi: 10.1186/s40537-023-00756-3.

[19] Y. I. Kim and S. Kim, "Feature scaling and normalization techniques in predictive modeling," *Applied Sciences*, vol. 12, no. 14, p. 7201, 2022, doi: 10.3390/app12147201.

[20] L. J. Garcia, M. M. Sanchez, and A. R. Martinez, "Handling missing data in machine learning: Techniques and applications," *Information Sciences*, vol. 512, pp. 348–365, 2020, doi: 10.1016/j.ins.2019.10.042.