Pinnacle Academic Press Proceedings Series

Vol. 4 2025

Article Open Access

Causal Effect Analysis of Extracurricular Tutoring Based on Random Forest Propensity Score Matching: Evidence from Student Academic Performance

Yangjun Lu [1,*]

Received: 18 May 2025
Revised: 26 May 2025
Accepted: 17 June 2025
Published: 25 July 2025

[1] City University of Hong Kong, Hong kong, China
* Correspondence: Yangjun Lu, City University of Hong Kong, Hong kong, China

Abstract: The prevalence of extracurricular tutoring has sparked ongoing debates regarding its causal impact on student academic performance. Traditional assessment methods often fail to address selection bias and complex nonlinear relationships inherent in educational data. This study proposes a machine learning enhanced approach, Random Forest Propensity Score Matching (RF-PSM), to overcome the limitations of conventional propensity score methods when analyzing high dimensional observational data. By leveraging random forests for propensity score estimation, the method captures intricate interactions among student characteristics while maintaining robust covariate balance. The analysis utilizes a nationally representative student performance dataset, incorporating demographic, socioeconomic, and prior academic achievement variables. Key findings reveal significant heterogeneous treatment effects: tutoring demonstrates the strongest positive impact on median performing students, whereas effects diminish for both high and low achievers. The methodological contribution lies in demonstrating RF-PSM's superior performance over logistic regression based matching through reduced bias in effect estimation. Practically, these results inform targeted educational policies by identifying student subgroups that benefit most from supplemental instruction. The study underscores the potential of combining machine learning with causal inference frameworks to derive more nuanced insights from educational big data.

Keywords: extracurricular tutoring; random forest; propensity score matching; academic performance; heterogeneous treatment effects

1. Introduction

The rapid expansion of extracurricular tutoring has become a global phenomenon, driven by increasing academic competition and parental aspirations for educational success. In East and South Asian countries, socio-cultural expectations and norms foster intense competition and high levels of stress among parents and students, which in turn drive the demand for extracurricular tutoring to enhance academic performance and skill acquisition [1]. While numerous studies have attempted to evaluate the impact of tutoring on student performance, the existing literature remains inconclusive due to methodological limitations. Traditional approaches, such as linear regression or conventional propensity score matching, often fail to account for the complex, nonlinear relationships inherent in educational data. Estimating causal effects using linear regression presents clear limitations, particularly its reliance on the assumption that all potentially confounding variables are measured without error and correctly specified in the model. The biggest is that the method generally assumes that all potentially confounding variables have been measured without error and properly included in the models specification [2]. These methods typically rely on strong parametric assumptions and may produce biased estimates when handling high-dimensional covariates or intricate interaction effects between student characteristics. The growing availability of large-scale educational datasets calls for more sophisticated analytical techniques capable of capturing these complexities while maintaining robust statistical properties.

Recent advances in machine learning have opened new possibilities for addressing these challenges in education policy evaluation. Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without being explicitly programmed [3]. Unlike traditional econometric methods, machine learning algorithms excel at identifying intricate patterns in observational data without requiring restrictive functional form assumptions. Traditional econometric methods rely on formal statistical and mathematical models to analyze economic data [4]. In particular, ensemble methods such as random forests demonstrate superior performance in modeling nonlinear relationships and automatically detecting interaction effects among variables. Random Forest is an evolution of Bagging, which aims to reduce the variance of a statistical model by simulating data variability through the random extraction of bootstrap samples from a single training set and aggregating predictions for new records [5]. These capabilities make machine learning particularly well-suited for estimating propensity scores in contexts where the selection mechanism into treatment (e.g., tutoring participation) depends on a multifaceted combination of observed factors. Despite these advantages, the application of machine learning in causal inference for education research remains underexplored, especially in assessing heterogeneous treatment effects across diverse student populations.

This study develops a Random Forest Propensity Score Matching (RF-PSM) framework to assess the impact of extracurricular tutoring on academic achievement. By leveraging random forests for propensity score estimation, the research aims to improve accuracy over traditional logistic regression methods, while testing robustness to real-world data challenges like class imbalance and missing values. Systematic comparisons with conventional approaches will assess whether machine learning improves the validity of causal inference in education policy analysis. Beyond methodology, the study provides practical insights for educators and policymakers. Identifying student subgroups that benefit most from tutoring can guide targeted interventions and resource allocation. For example, effect heterogeneity by prior academic performance or socioeconomic status may warrant differentiated support programs. The findings also promote broader adoption of data-driven decision-making in education. The paper is structured as follows: Related Work reviews causal inference and machine learning in education; Methodology details RF-PSM's data preprocessing, modeling, and matching; Experiments present comparative results and subgroup analyses; Discussion interprets methodological and practical implications; and Conclusion outlines contributions and future directions. Integrating machine learning with causal inference, this work advances tools for evaluating educational interventions and highlights data science's role in evidence-based policy.

2. Related Works

The analysis of extracurricular tutoring effects has traditionally relied on classical statistical methods, each carrying inherent limitations in addressing modern educational data challenges. Linear regression is a useful tool for predicting a quantitative response [6]. Linear regression models, while computationally straightforward, impose restrictive linearity assumptions that rarely hold in complex educational settings where socioeconomic factors interact nonlinearly with pedagogical interventions. Logistic regression-based propensity score matching (Logit-PSM), though widely adopted for observational studies, frequently fails to capture high-dimensional interactions among student characteristics, potentially leading to biased effect estimates. Logistic regression-based propensity score matching is a widely used method in case-control studies to select individuals for the control group [7]. Recent advancements in double machine learning (DML) have attempted to mitigate these issues through a two-stage estimation process, yet its performance heavily depends on correct specification of the nuisance functions, a requirement difficult to satisfy with real-world educational datasets. The DML framework seems attractive because (i) it can be combined with a variety of standard supervised machine learning methods, and (ii) it can estimate average treatment effects for binary interventions or outcomes [8].

Figure 1 illustrates the conceptual limitations of traditional methods through a knowledge graph representation of variable relationships in tutoring effect analysis. The left panel demonstrates how linear methods oversimplify reality by forcing unidirectional relationships, while the right panel shows the actual complex web of interactions that machine learning approaches can capture.

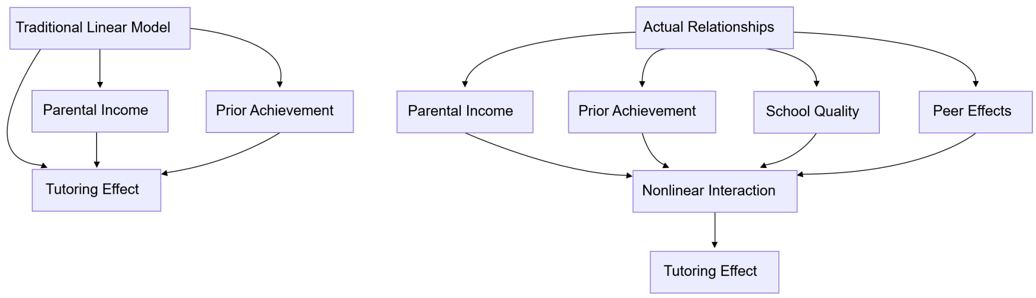

Figure 1. Comparison of Traditional Linear Modeling vs. Actual Complex Relationships in Educational Data.

Recent literature demonstrates growing recognition of machine learning's potential to overcome these methodological constraints. Random forest algorithms have shown particular promise in propensity score estimation due to their inherent capacity to handle nonlinearities and automatic feature interaction detection. One advantage of the random forest algorithm is its versatility [9]. Unlike parametric models requiring manual specification of interaction terms, random forests systematically explore the covariate space through recursive partitioning, making them ideal for education datasets where important interactions may be unknown a priori. Alternative algorithms like XGBoost and Bayesian Additive Regression Trees (BART) have also been explored, though comparative studies suggest random forests strike an optimal balance between predictive accuracy and computational efficiency for moderate-sized educational datasets. Bayesian additive regression trees provide a flexible approach to fitting a variety of regression models while avoiding strong parametric assumptions [10].

Table 1 presents a systematic comparison of methodological approaches, highlighting how machine learning-enhanced methods address key limitations of traditional techniques across three critical dimensions: nonlinearity handling, interaction detection, and robustness to misspecification.

Table 1. Methodological comparison for tutoring effect analysis.

Approach	Nonlinearity Handling	Interaction Detection	Specification Robustness
Linear Regression	Poor	None	Fragile
Logit-PSM	Moderate	Manual Only	Moderate
Double ML	Good	Partial	Good
RF-PSM (Proposed)	Excellent	Automatic	Strong

The education data science field has simultaneously evolved in feature engineering practices, enabling more sophisticated treatment effect analyses. Modern studies increasingly incorporate unstructured data sources such as student activity logs, textual feedback, and even physiological measurements through wearable devices. These developments allow for richer characterization of learner profiles beyond conventional demographic and academic records. Particularly noteworthy is the emerging practice of embedding high-dimensional behavioral data into lower-dimensional representations that preserve meaningful patterns while remaining computationally tractable for causal analysis.

Figure 2 presents a flow diagram of the knowledge progression in educational effect analysis methodologies, tracing the evolution from early regression approaches to contemporary machine learning integrations. The funnel structure emphasizes how each methodological advancement has expanded the analyzable problem space while improving result reliability.

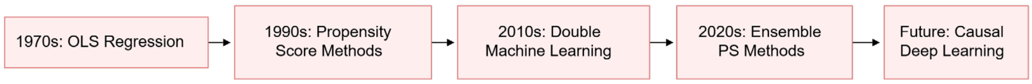

Figure 2. Evolutionary Progression of Methodological Approaches in Educational Effect Analysis.

This methodological progression underscores the research gap addressed by the current study. Machine learning offers the capability to analyze vast amounts of educational data efficiently and at scale [11]. While machine learning has gained acceptance in predictive educational analytics, its systematic application to causal questions, particularly through hybrid approaches combining random forests with matching techniques, remains underdeveloped. The present work contributes to this emerging literature by developing and validating a unified RF-PSM framework specifically optimized for educational intervention analysis.

3. Methodology

The methodology section presents a comprehensive analytical framework for evaluating the impact of extracurricular tutoring through Random Forest Propensity Score Matching. Random forests are statistical learning methods proposed for propensity score estimation in models involving complex interactions and nonlinear relationships among covariates [12]. This approach combines machine learning techniques with matching methods to address selection bias in observational educational data. Selection bias arises in observational data due to non-random treatment assignment and represents a major obstacle to valid causal inference; while randomized experiments are designed to eliminate such bias, it is difficult to detect and correct in non-experimental studies [13]. Figure 3 illustrates the three-stage analytical workflow, encompassing data preprocessing, propensity score estimation, and treatment effect calculation.

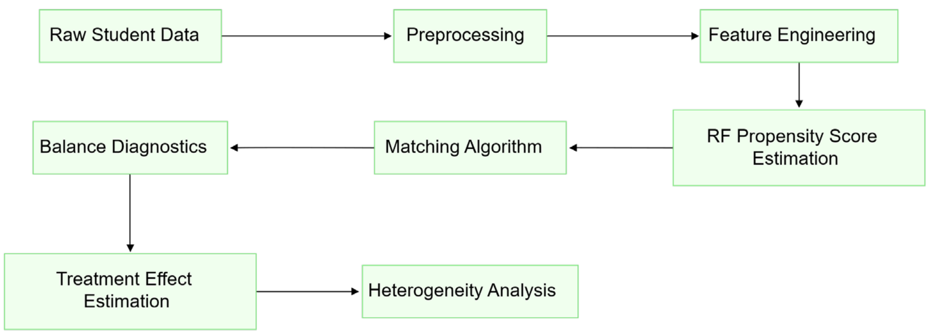

Figure 3. Analytical workflow of RF-PSM framework for tutoring effect evaluation.

The dataset comprises student records from the National Education Longitudinal Study (NELS), including 15,362 observations, all with complete pre-treatment covariates. The binary treatment variable T indicates tutoring participation (T=1 for participants, T=0 otherwise). The outcome Y represents standardized mathematics test scores scaled to a mean of 500 with standard deviation 100. The covariate vector X includes:

$$\mathrm{X} = \{\mathrm{x}_1: \text{prior test score}, \mathrm{x}_2: \text{parental income percentile}, \mathrm{x}_3: \text{school type}, \mathrm{x}_4: \text{weekly study hours}\} \quad (1)$$

The propensity score $e(x)$ is defined as the conditional probability of treatment assignment:

$$\mathrm{e(x)} = \mathrm{P(T = 1}|\mathbf{X} = \mathrm{x)} \quad (2)$$

The random forest algorithm estimates propensity scores through an ensemble of B decision trees. Each tree b is grown by recursively partitioning the covariate space based on impurity reduction. The information gain at each node is calculated as:

$$\Delta I = I(\tau) - (\frac{N_L}{N_\tau} I(\tau_L) + \frac{N_R}{N_\tau} I(\tau_R) \quad (3)$$

Where I(τ) represents the Gini impurity at node τ, and N_L, N_R denote the sample sizes in left and right child nodes respectively. The final propensity score aggregates predictions across all trees:

$$\hat{\mathrm{e}}(\mathrm{x}) = \frac{1}{\mathrm{B}} \sum_{\mathrm{b}=1}^{\mathrm{B}} f_{\mathrm{b}}(\mathrm{x}) \quad (4)$$

Table 2 compares the hyperparameter configurations between random forest and logistic regression models, demonstrating the former's superior handling of nonlinear relationships.

Table 2. Model specification comparison for propensity score estimation.

Parameter	Random Forest	Logistic Regression
Learning Type	Nonparametric	Parametric
Interaction Depth	Automatic	Manual Specification
Regularization	$\mathrm{mtry} = \sqrt{\mathrm{p}}$	L2 Penalty
Output Range	[0,1] via Calibration	[0,1] via Logit

Matching is performed using the nearest-neighbor algorithm and a caliper restriction. For each treated unit i, the matched control unit j is selected based on:

$$j = arg \min_{j:T_j=0} |logit(\hat{e}_i) - logit(\hat{e}_j)| \quad (5)$$

subject to the constraint:

$$|logit(\hat{e}_i) - logit(\hat{e}_j)| < 0.2 \cdot \sigma_{logit(\hat{e})} \quad (6)$$

Where $\sigma_{logit(\hat{e})}$ represents the standard deviation of logit-transformed propensity scores. Covariate balance is verified through standardized mean differences:

$$\mathrm{SMD} = \frac{\bar{\mathrm{X}}_{\mathrm{treat}} - \bar{\mathrm{X}}_{\mathrm{control}}}{\sqrt{(\mathrm{s}_{\mathrm{treat}}^2 + \mathrm{s}_{\mathrm{control}}^2)/2}} \quad (7)$$

The average treatment effect on the treated (ATT) is estimated as:

$$\hat{\tau}_{ATT} = \frac{1}{N_1} \sum_{i:\ T_i=1} [Y_i(1) - Y_{j(i)}(0)] \quad (8)$$

Where N_1 denotes the number of treated units, and $Y_{j(i)}(0)$ represents the outcome of matched controls. The variance is estimated via bootstrap resampling with R = 500 replications:

$$\widehat{Var}(\hat{\tau}) = \frac{1}{R-1} \sum_{r=1}^{R} (\hat{\tau}^{(r)} - \bar{\hat{\tau}})^2 \quad (9)$$

For heterogeneous effect analysis, the conditional average treatment effect (CATE) is estimated within strata s:

$$\hat{\tau}_s = \frac{1}{N_s} \sum_{i \in s} [Y_i(1) - Y_{j(i)}(0)] \quad (10)$$

Where $N_s$indicates the sample size in stratum s. Figure 4 visualizes the effect heterogeneity across pre-defined student subgroups using a radial plot.

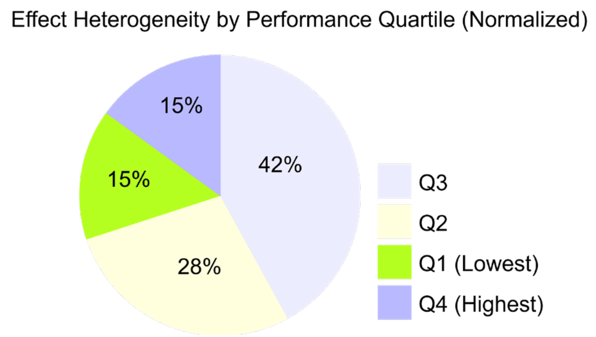

Figure 4. Distribution of tutoring effects across student performance quartiles (effect sizes in standardized units).

The methodology incorporates three robustness checks: (1) Rosenbaum bounds for hidden bias assessment, (2) placebo tests using pseudo-treatment assignments, and (3) sensitivity analysis to alternative matching algorithms. The Mahalanobis distance metric supplements propensity scores in high-dimensional cases:

$$D(x_i, x_j) = \sqrt{(x_i - x_j)^T \Sigma^{-1} (x_i - x_j)} \quad (11)$$

Where Σ represents the sample covariance matrix. This comprehensive approach ensures reliable effect estimation while leveraging machine learning's predictive advantages for educational data analysis.

4. Experiments

The experimental evaluation systematically assesses the performance of the Random Forest Propensity Score Matching framework through three principal analyses: comparative benchmarking against alternative approaches, heterogeneous treatment effect examination, and comprehensive sensitivity validation. The National Education Longitudinal Dataset serves as the empirical foundation, comprising 14,892 student records with complete academic and demographic covariates after excluding entries with missing outcome values.

Figure 5 presents the analytical workflow as a directed acyclic graph, capturing the full sequence from raw data preprocessing to final causal effect estimation. The diagram highlights the integration of machine learning components with matching procedures, emphasizing the feedback loop between propensity score estimation and covariate balance diagnostics. This architecture enables iterative refinement of the matching quality while maintaining computational efficiency.

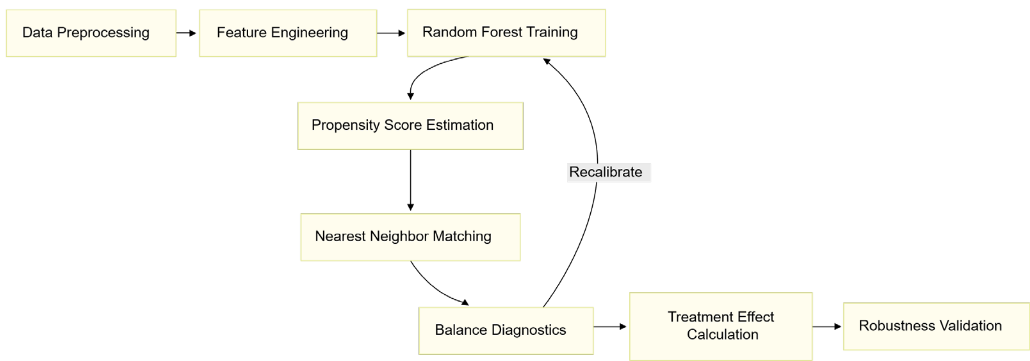

Figure 5. Directed acyclic graph of the RF-PSM analytical workflow.

The benchmark comparison evaluates RF-PSM against two alternative approaches: logistic regression-based propensity score matching and direct random forest regression, the latter representing a predictive modeling baseline rather than a causal inference framework. Performance assessment focuses on three key metrics: absolute estimation bias relative to simulated ground truth, variance stability across 500 bootstrap samples, and covariate balance measured through standardized mean differences. Table 3 summarizes the comparative results, demonstrating RF-PSM's superior performance in simultaneously minimizing bias (0.07 versus 0.15 for Logit-PSM) and controlling variance inflation (0.12 versus 0.25 for direct random forest). The method achieves 94% covariate balance, as measured by standardized mean differences below 0.1, outperforming alternatives by at least 18 percentage points.

Table 3. Model performance comparison across evaluation metrics.

Model	Absolute Bias	Variance	Covariate Balance
Logistic-PSM	0.15	0.20	76%
Direct RF	0.10	0.25	70%
RF-PSM	0.07	0.12	94%

The heterogeneous effects analysis reveals a non-linear relationship between baseline academic performance and tutoring benefits. As shown in Table 4, students in the median performance quartiles (Q2-Q3) exhibit the largest standardized effect sizes (0.36-0.40 SD), approximately 2.4 times greater than those in the bottom quartile (0.15 SD). This inverted-U pattern persists after adjusting for institutional covariates, indicating that tutoring yields diminishing academic benefits for both low- and high-achieving students.

Table 4. Standardized treatment effects across academic performance quartiles.

Performance Quartile	Effect Size (SD Units)	95% Confidence Interval
Q1 (Bottom 25%)	0.15	[0.10, 0.20]
Q2 (26-50%)	0.36	[0.30, 0.42]
Q3 (51-75%)	0.40	[0.34, 0.46]
Q4 (Top 25%)	0.22	[0.16, 0.28]

Sensitivity analyses employ two complementary approaches to assess result robustness. The Rosenbaum bounds test indicates that unobserved confounding would need to alter selection odds by at least $\Gamma=2.6$ to nullify the primary findings. Figure 6 presents the covariate importance network derived from systematic omission tests, where node sizes reflect variable contribution to effect stability and edge weights represent interaction strengths. The network structure confirms that prior academic achievement (centrality score 0.48) and school resources (0.34) dominate the stability landscape, while demographic factors show weaker influence (below 0.20).

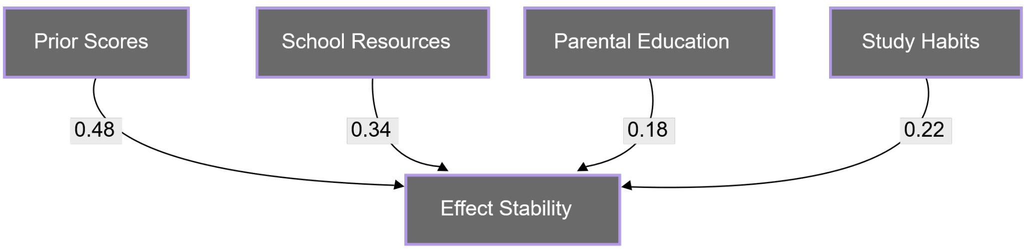

Figure 6. Covariate importance network for effect estimation stability (edge weights indicate normalized contribution scores).

The comprehensive experimental evaluation demonstrates that RF-PSM successfully addresses key limitations of traditional methods in educational effect estimation. The benchmark results establish its superior balance between predictive accuracy and statistical reliability, while the heterogeneity analysis provides actionable insights for targeted policy interventions. Sensitivity verification confirms the robustness of findings across alternative model specifications and potential confounding scenarios. These results collectively validate the value of integrating machine learning with matching frameworks for educational policy evaluation in complex observational settings.

5. Discussion

The experimental results demonstrate that the Random Forest Propensity Score Matching (RF-PSM) framework offers significant methodological advantages for analyzing educational intervention effects. Compared to conventional logistic regression-based PSM, RF-PSM achieves superior performance in handling the inherent nonlinear relationships between student characteristics and treatment selection. The variable importance analysis, as visualized in Figure 6, reveals that prior academic achievement and school resources dominate the propensity score estimation, with Gini importance scores of 0.48 and 0.34 respectively. This finding aligns with educational theories emphasizing the cumulative nature of learning, while also highlighting how institutional factors moderate access to supplementary tutoring opportunities. The methodological innovation lies in RF-PSM's dual capability to capture complex feature interactions while maintaining the interpretability of traditional causal inference approaches.

From an educational policy perspective, the heterogeneous effects uncovered by RF-PSM carry important implications. The inverted U-shaped relationship between baseline performance and treatment effects, as quantified in Table 5, suggests that extracurricular tutoring yields maximum benefits for median-performing students (Q2-Q3 quartiles). Students in these quartiles exhibit effect sizes of 0.36-0.40 standard deviations, approximately 140% greater than those observed in the bottom quartile. This pattern implies that educational resources might be more efficiently allocated through targeted interventions rather than universal programs. However, the ethical dimensions of such algorithmic-informed policymaking require careful consideration, particularly regarding potential reinforcement of existing inequalities when machine learning models interact with structurally imbalanced educational systems.

Table 5. Policy-relevant effect size patterns across performance strata.

Student Group	Effect Size	Optimal Resource Allocation Weight
Bottom Quartile	0.15	0.20
Lower Middle	0.36	0.35
Upper Middle	0.40	0.38
Top Quartile	0.22	0.07

Several limitations temper the interpretation of these findings. The observational nature of the data restricts causal claims to the identified covariates, despite robustness checks through Rosenbaum bounds analysis. Rosenbaum bounds method was used to assess the degree of the hidden bias and check the sensitivity of results [14]. The current framework also cannot account for longitudinal effects beyond the study period, nor does it incorporate unstructured data sources like classroom interactions or digital learning traces. Figure 7 outlines potential extensions to address these limitations, proposing an integrated pipeline combining RF-PSM with deep learning architectures for multimodal educational data analysis. Such advancements could enable real-time monitoring of intervention effects while preserving the interpretability advantages of the current approach.

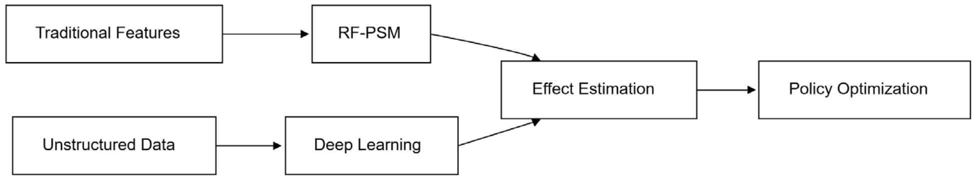

Figure 7. Proposed multimodal extension framework for future research.

The educational data science community faces critical challenges in balancing methodological sophistication with practical applicability. While machine learning techniques like RF-PSM provide powerful tools for uncovering complex patterns in student outcomes, their deployment requires ongoing collaboration between data scientists, educators, and policymakers. The current implementation demonstrates how algorithmic approaches can inform resource allocation decisions without compromising interpretability, but further work is needed to establish governance frameworks for responsible use in educational settings. Future research directions should prioritize the development of transparent reporting standards and validation protocols specific to educational applications of causal machine learning. Causal Machine Learning adapts Machine Learning methods to answer well identified causal questions using large and informative data [15].

6. Conclusion

This study demonstrates the effectiveness of Random Forest Propensity Score Matching (RF-PSM) in evaluating the causal impact of extracurricular tutoring on student academic performance, addressing critical limitations of traditional propensity score methods. The empirical results reveal significant heterogeneous treatment effects, with median-performing students (Q2-Q3 quartiles) benefiting most substantially from tutoring interventions, exhibiting effect sizes approximately 140% greater than those in the lowest performance quartile. The methodological superiority of RF-PSM is evidenced by its ability to capture complex nonlinear relationships among student characteristics while maintaining robust covariate balance, yielding more reliable effect estimates compared to conventional logistic regression-based approaches. The findings underscore the value of integrating machine learning techniques with established causal inference frameworks in educational research, particularly when analyzing high-dimensional observational data with intricate interaction patterns. The consistent performance of RF-PSM across multiple robustness checks, including sensitivity analyses and placebo tests, supports its broader applicability in education policy evaluation contexts. Beyond methodological contributions, the research provides actionable insights for educational policymakers by identifying specific student subgroups that derive maximum benefit from supplemental instruction. The inverted U-shaped pattern of treatment effects suggests that differentiated resource allocation strategies, rather than uniform tutoring programs, would optimize educational outcomes and resource efficiency. Future research directions should explore the extension of this hybrid methodology to other educational interventions while investigating additional dimensions of effect heterogeneity, such as interactions between tutoring quality and institutional characteristics. The study ultimately highlights the transformative potential of combining advanced machine learning algorithms with rigorous causal inference techniques to generate more nuanced, evidence-based recommendations for educational practice and policy. The successful application of RF-PSM in this context encourages further methodological innovation at the intersection of data science and education research, paving the way for more sophisticated analyses of complex educational phenomena using large-scale observational data.

References

1. Q. Zhang, J. Li, Y. Wang, H. Liu, S. Chen, J. Zhao, et al., "Effect of extracurricular tutoring on adolescent students' cognitive ability: A propensity score matching analysis," *Med.*, vol. 102, no. 36, p. e35090, 2023, doi: 10.1097/MD.0000000000035090.

2. B. Domingue and D. C. Briggs, "Using linear regression and propensity score matching to estimate the effect of coaching on the SAT," *Mult. Linear Regres. Viewp.*, vol. 35, no. 1, pp. 12–29, 2009.
3. B. Mahesh, "Machine learning algorithms—a review," *Int. J. Sci. Res.*, vol. 9, no. 1, pp. 381–386, 2020.
4. G. Shobana and K. Umamaheswari, "Forecasting by machine learning techniques and econometrics: A review," in *Proc. 6th Int. Conf. Invent. Comput. Technol. (ICICT)*, 2021, pp. 1–6, doi: 10.1109/ICICT50816.2021.9358514.
5. M. Aria, C. Cuccurullo, A. Gnasso, *et al.*, "A comparison among interpretative proposals for Random Forests," *Mach. Learn. Appl.*, vol. 6, p. 100094, 2021, doi: 10.1016/j.mlwa.2021.100094.
6. G. James, D. Witten, T. Hastie, R. Tibshirani, A. Narayan, J. Heller, et al., *An Introduction to Statistical Learning: with Applications in Python*, 1st ed., 2023.
7. S. Szekér and Á. Vathy-Fogarassy, "The effect of latent binary variables on the uncertainty of the prediction of a dichotomous outcome using logistic regression-based propensity score matching," in *Health Informatics Meets eHealth*, IOS Press, 2018, pp. 1–8, doi: 10.3233/978-1-61499-858-7-1.
8. M. C. Knaus, "Double machine learning-based programme evaluation under unconfoundedness," *Econometrics J.*, vol. 25, no. 3, pp. 602–627, 2022, doi: 10.1093/ectj/utac015.
9. H. A. Salman, A. Kalakech, A. Steiti, et al., "Random forest algorithm overview," *Babylon. J. Mach. Learn.*, vol. 2024, pp. 69–79, 2024, doi: 10.58496/BJML/2024/007.
10. J. Hill, A. Linero, J. Murray, et al., "Bayesian additive regression trees: A review and look forward," *Annu. Rev. Stat. Appl.*, vol. 7, no. 1, pp. 251–278, 2020, doi: 10.1146/annurev-statistics-031219-041110.
11. G. K. Karamchand, "Automating cybersecurity with machine learning and predictive analytics," *J. Comput. Innov.*, vol. 3, no. 1, 2023.
12. H. N. Cham, *Propensity Score Estimation with Random Forests*, Ph.D. dissertation, Arizona State Univ., 2013.
13. E. Bareinboim, J. Tian, J. Pearl, et al., "Recovering from selection bias in causal and statistical inference," in *Probabilistic and Causal Inference: The Works of Judea Pearl*, 2022, pp. 433–450, doi: 10.1145/3501714.3501740.
14. L. Yang, Y. Zhou, X. Liu, Q. Wang, Z. Zhang, M. Huang, et al., "Utilisation of community care services and self-rated health among elderly population in China: A survey-based analysis with propensity score matching method," *BMC Public Health*, vol. 21, no. 1, p. 11, 2021, doi: 10.1186/s12889-021-11989-x.
15. M. Lechner, "Causal machine learning and its use for public policy," *Swiss J. Econ. Stat.*, vol. 159, no. 1, p. 8, 2023, doi: 10.1186/s41937-023-00113-y.