Engineering Educators Bringing the World Together

2025 ASEE Annual Conference & Exposition

Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025 ASEE

Paper ID #48637

Data-Driven Insights into Academic Success: Analyzing ten years of student academic records in an Electrical and Computer Engineering department

Mr. Weiyu Sun, Georgia Institute of Technology

Weiyu Sun is pursuing doctoral degree of Electrical and Computer Engineering (ECE) at Georgia Institute of Technology since Fall 2024. He earned BE and ME degrees of ECE at Nanjing University in China. His research interests/fields include Trustworthy AI, AI4Education, AI4Science, bio-signal processing, and foundation model.

Dr. Jacqueline Rohde, Georgia Institute of Technology

Jacqueline (Jacki) Rohde is the Assessment Coordinator in the School of Electrical and Computer Engineering at the Georgia Institute of Technology. Her interests are in sociocultural norms in engineering and the professional development of engineering students.

Dr. Liangliang Chen, Georgia Institute of Technology

Liangliang Chen received a Ph.D. degree in the School of Electrical and Computer Engineering at Georgia Tech. Prior to that, he received a B.B.A. degree in business administration, a B.S. degree in automation, and an M.Eng. degree in control engineering from Harbin Institute of Technology. His research interests are machine learning theory and applications.

Yiming Guo, Georgia Institute of Technology

Yiming Guo is pursuing a Master of Science degree in Electrical Engineering at the Georgia Institute of Technology. He received his Bachelor of Science degree at University of California, Los Angeles. His primary interests involve machine learning and circuit design.

Dr. Ying Zhang, Georgia Institute of Technology

Dr. Ying Zhang is a Professor and Senior Associate Chair in the School of Electrical and Computer Engineering at the Georgia Institute of Technology. She is the director of the Sensors and Intelligent Systems Laboratory. Her research interests are centered on systems-level interdisciplinary problems across multiple engineering disciplines, with AI-enabled personalized engineering education being one of her current research focuses.

Data-Driven Insights into Academic Success: Analyzing ten years of student academic records in an Electrical and Computer Engineering department

Abstract

This empirical research paper presents an analysis of a longitudinal dataset covering 10 years of student academic performance using statistical and machine learning methods, contextualized within the School of Electrical and Computer Engineering (ECE) at a large, public, research-intensive institution in the Southeast United States. This investigation expands upon a research work presented at the 2024 ASEE Conference, which identified predictors of student academic success in an upper-level microelectronic circuits course from a smaller dataset with fewer predictors. In this study, we have expanded the analysis in several dimensions (i.e., time scale, predictor variables, and outcome variables) and have also developed machine learning models to predict student performance in the core microelectronic circuits course in the ECE curriculum. Altogether, this analysis indicates opportunities to help program leaders provide students with early, effective, and personalized support, enhancing academic success across the diverse backgrounds that students bring to their undergraduate studies. This work builds on prior work in engineering education modeling predictors of academic success through academic records, but is contextualized to a specific undergraduate program, which allows for a fine-grained interpretation of results that may be transferable to other institutions.

Introduction and Background

Many engineering educational researchers have worked with large-scale datasets of students' academic records to better understand influential factors on students' performance [1, 2, 3, 4]. Such datasets enable robust statistical analyses that uncover generalizable trends across diverse student populations, providing valuable insights into the systemic influences on student outcomes, as well as to identify students who may need additional support to achieve the academic success of which they are capable. These studies have shed light on critical factors such as high school preparation (e.g., [5]) and first-year experiences (e.g., [6]), which collectively influence students' persistence and success in engineering undergraduate programs.

Building on this body of work, this paper leverages the unique opportunity afforded by the large enrollment in the School of Electrical and Computer Engineering to create a dataset that balances statistical power with the ability to capture department-specific nuances. Unlike national datasets, which are invaluable for understanding macro-level trends, institution-specific datasets allow for a deeper exploration of localized factors, including curriculum design, pedagogical practices, and departmental policies, that may uniquely impact student performance and engagement. This localized focus can also reveal long-standing legacies within the department that influence student outcomes, offering opportunities for targeted interventions and reforms.

In addition to drawing on traditional statistical methods, this work incorporates machine learning techniques to analyze complex, high-dimensional data. Machine learning approaches enable the identification of non-linear relationships and interactions among variables that might otherwise remain hidden in conventional analyses. These methods can support the development of personalized interventions, enabling institutions to proactively address challenges faced by individual students or specific student populations. This work seeks to harness the power of machine learning not only to enhance the interpretability of complex educational data but also to provide actionable insights that can guide curriculum development, academic advising, and institutional policies aimed at improving student outcomes.

This work expands upon a previous analysis [7]. The expanded dataset spans from Fall 2014 to Spring 2024 and includes data for approximately 1,600 students. The dataset includes students' high school performance (e.g., whether taking advanced placement courses, SAT scores), academic records at the current institution (e.g., grades and number of attempts in various courses), and additional characteristics such as transfer credits, transfer history, gender, and ethnicity. Given the large scale of the institution and the diverse backgrounds of its students, the dataset exhibits significant heterogeneity across the attributes. By applying data-mining techniques to this dataset, we can gain valuable insights into the factors that influence students' academic performance throughout their time at the institution.

The analysis in this paper concentrated on a 4-credit hour, junior-level course on semiconductor architecture that is notorious within the department for its difficulty. We refer to this course as "ECE 301" (a pseudonym). Of the 1,592 students who have taken ECE 301 in the study time frame, 15.3% earned a final grade of a "D", "F", or "W" and 19.0% earned a "C" in their first trial. Among these students, around 65% of students were direct matriculation, meaning that around a third of students were transfer or readmitted students, although this number has fluctuated over time. For the purposes of this study, a representative subset of approximately 790 student samples was selected from the full dataset. This subset was chosen to balance computational feasibility with the ability to capture meaningful trends and variations within the data. A heuristic analysis was conducted on this subset, with findings presented in the main text.

Literature Review

Recently, more and more research and projects [8] have been conducted to analyze factors that predict students' journeys and academic success in academic institutions. With the trend, many related datasets are crafted, such as StudentLife [9] and MIDFIELD [1]. Some of these datasets can be specific to a certain case (such as within a specific institution/course), while others collected data widely, compiling academic record data across different institutions. From these records, researchers can capture critical factors associated with students' academic success by data-mining these collected academic data [10]. Nevertheless, with student populations becoming increasingly diverse, educators aim to discover wider and more detailed factors to facilitate more precise and personalized assistance for those at-risk students. For example, Hu and Rangwala [11] used deep learning tools to analyze the impact of a certain course grade on the student's academic performance instead of only relying on overall GPAs. Aggarwal et al. [12] highlighted the advantages of incorporating non-academic factors, such as parental annual income and the state of residence, into the evaluation of a student's future performance. Some studies [13, 14] have also reported the importance of factors such as motivation and belongingness, and althougth linking these factors with academic record shows immense progress, it requires additional data collection procedures. Altogether, these existing works indicate the need for more fine-grained and diverse student academic data to better support students. With a similar goal of exploring the detailed factors influencing students' academic success, this work represents the early quantitative phase of a larger mixed-method project aimed at identifying opportunities to support ECE students' academic success.

Purpose and Research Questions

This work aims to answer the following questions via data-mining the collected student data on ECE 301:

(1) What are the relative influences of student characteristics, course characteristics, and student prior academic performance on students' final grades in ECE 301?

(2) To what extent can we identify students who are at risk of struggling in ECE 301 using information contained within their prior academic records?

Methods

Data Creation

Previously, the authors [7] described the creation of an academic record dataset for ECE students at an institution in the United States. Although this dataset provided valuable insights into student success in a microelectronic circuits course, there were some noted limitations. For example, the authors were unable to access information about credits students brought to the institution, whether through prior coursework or pathways such as Advanced Placement (AP) testing. As a result, the dataset did not capture the complete academic record of students, particularly for transfer students. We have worked alongside institutional data management experts to expand the initial dataset in [7] and support more robust analyses. This expansion includes expanding the time frame (Fall 2014 — Spring 2024, compared to Fall 2016 — Spring 2023), adding new student- and course-level variables, and collaboratively building systems for updating the dataset each semester. The project was approved by the institution's Institutional Review Board, and the data provided was de-identified using unique identifiers to safeguard students' information. We outline the dataset's variables and structure to support researchers conducting department-level analyses of student academic performance as follows. Details on the measurement of various student and course characteristics can be found in [7].

- *Student Characteristics*: Each row represents a student-semester instance and includes information such as the student's unique identifier, gender, race, citizenship status, residency code, matriculation term, admit type, and standardized test scores (SAT or SAT equivalent). High school GPA and the name of the high school were also included if provided during admissions.
- *Transfer Credits*: Each row represents a student-course instance and includes the student's unique identifier, course details (subject code and name) for which transfer credits were awarded, and the source of those credits (e.g., prior institution, AP testing, SAT II testing, etc.).
- *Course History*: Each row represents a student-course instance and includes the student's unique identifier, term code, course details (subject code and name), final grade, instructor

name, and the student's cumulative GPA and term GPA.

In addition to the institutional dataset, we utilized two supplementary data sources. First, to contextualize transfer credits, we categorized institutions based on their Carnegie classification. Second, we gathered data on the average grade in the course for each course section in each semester. This information allowed us to interpret grade data as relative performance, comparing students in similar contexts while accounting for potential variations between instructors. Using these sub-datasets, we constructed a student-level dataset, where each row represents an individual student with their academic information aggregated into a series of columns. The variables examined include:

- Admission type (direct matriculation from high school versus other matriculation pathways)
- SAT Math score (for directly matriculated students; converted from ACT if necessary)
- AP counts (number of AP exams a student used for transfer credits in STEM courses)
- Student cumulative GPA of the term they take the latest pre-course (on a 4.0 scale)
- Final grades in courses within the ECE 301 pre-requisite chain. This chain has changed over time and includes grades on an "A–F" scale, with a "T" representing transfer credits. The courses are: Physics II, Differential Equations (DiffEQ), Multivariate Calculus (Calc3), Programming, Digital System Design, and Circuit Analysis.
- Normalized final grade in courses within the ECE 301 pre-requisite chain (numeric). To account for variations in grading standards among instructors, each student's pre-requisite grade is "normalized" by dividing it by the average grade of the specific section when the credits were earned at the focal institution.
- High school information, including students' SAT/ACT scores and AP test data.

Compared with [7], the analysis in this paper does not include students' demographic variables (gender and race/ethnicity) or instructor identity. Note that gender and race were not identified as significant predictors in the prior work [7]. The focus of this paper is more on mutable attributes, such as proficiency in specific pre-requisites. The enhanced dataset offers a robust foundation for exploring factors contributing to student success in ECE.

Data Analysis

We categorized students based on their matriculation history to separately analyze directly matriculated students and transfer students. Although these students share the same classrooms, their academic records exhibit significant differences that warrant careful consideration. First, directly matriculated students typically completed ECE 301's core pre-requisites (such as Physic II and Circuit Analysis) at the focal institution. This provides a detailed record of their proficiency, reflected through a range of letter grades. In contrast, transfer students often bring in credits for pre-requisites (shown in Figure 3), which are recorded as a "T" (transfer) on their academic records. This limits insights into their knowledge acquisition and retention. Second, the academic record's ability to capture students' academic histories differs between groups. Transfer credits are recorded in the semester they are recognized by the focal institution (often the students' first semester there), rather than the semester the courses were originally completed. This distinction affects how academic histories are evaluated over time. As a result, it is more practical to treat these two groups as separate when analyzing students at different stages of their academic journey.

To determine whether we can identify students who may be at risk of struggling in ECE 301 early, allowing us to provide the necessary support to enhance their academic success, we designed different stages to investigate prediction models. The stages are designed based on several prerequisite courses (feature courses) students are required to take before enrolling in ECE 301.

For directly matriculated students, we selected three ECE foundation courses: Physics II, Multivariate Calculus (Calc 3), and Differential Equations (DiffEQ), along with four ECE major courses: Digital System Design, Programming, Circuit Analysis, and Digital Design. These courses are typically taken in a specific sequence, as illustrated in Figure 4(a). This chronological order provides a useful framework for analyzing students at different stages of their academic journey and offering timely, personalized support to identified at-risk students. The stage design for directly matriculated students is summarized in Table 1.

- Stage 1 includes students' high school information, grades from two introductory courses – Physics II and Digital System Design – and their cumulative GPAs for the corresponding semester. This information is available by the end of the second term (as shown in Figure 4(a), on average, students complete Physics II and Digital System Design during their first and second semester, respectively).
- Stage 2 adds DiffEQ and Calc3, which are generally completed by the end of the third term, to the feature list of Stage 1. Students' cumulative GPAs are also updated.
- Stage 3 incorporates all prerequisite courses, typically completed by the fourth term, in addition to updated cumulative GPAs and high school information.

For transfer students, the stage design and associated feature list are different, as some prerequisite courses are transferred from other institutions (as shown in Figure 3). For example, Calc3, Physics II, and DiffEQ are commonly transferred. All these transfer credits are marked as "T", resulting in information redundancy and high feature correlation among these course features. Therefore, we retained Physics II from these three courses to explore the relationship between students' performance in ECE 301 and their previous institution. Target encoding was used to replace the "T" grades with the average ECE 301 grade of the corresponding group of students who transferred from the same class of institution, as classified by the Carnegie Classification of Institutions of Higher Education. Additionally, as shown in Figure 4(b), the sequence of prerequisite courses for transfer students differs slightly from that of directly matriculated students due to the transferred credits. For example, the average term for completing Circuit Analysis is earlier, as many transfer students receive credit for this course prior to enrollment. The stage design for transfer students is outlined in Table 2:

- Stage 1 includes students' high school information, grades from Physics II and Digital System Design, and their cumulative GPAs for the corresponding semester. This information is commonly available by the end of the first term, as shown in Figure 4(b).
- Stage 2 adds Circuit Analysis, which are generally completed by the end of the first or second term, to the feature list of Stage 1. Students' cumulative GPAs are also updated.
- Stage 3 incorporates all prerequisite courses, typically completed by the third term, in addition to updated cumulative GPAs and high school information.

Table 1: The stage design for directly matriculated students

Stage	**Feature List**
1	Physics II, Digital System Design, cumulative GPA, high school information
2	Stage 1 + DiffEQ + Calc3
3	Stage 2 + Circuit Analysis + Programming + Digital Design

Table 2: The stage design for transfer students

Stage	**Feature List**
1	Physics II, Digital System Design, cumulative GPA, high school information
2	Stage 1 + Circuit Analysis
3	Stage 2 + Programming + Digital Design

For the prediction target, we classify students' performance in ECE 301 into two categories: grades A and B are labeled as good performance ("0"), while grades C, D, F, and W are labeled as poor performance ("1"). By including grade C in the "poor performance" category, the program can proactively target a larger group of students for intervention, ensuring that the students who are on the borderline receive the resources needed to improve their performance before facing academic difficulties. With the designed stages and prediction targets, machine learning tools are applied to classify and analyze both directly matriculated students and transfer students.

For the machine learning model, we selected random forest [15] for prediction due to several reasons:

1) *Ability to handle feature correlation:* One notable characteristic of our data is the potential correlation between different features, such as grades in various pre-requisite courses, as illustrated in Figure 5. Random forest mitigates this issue by using different data subsets for each decision tree, effectively reducing the impact of feature correlation.
2) *Robustness to outliers:* In practice, a student's performance in a specific course may be affected by some non-academic factors, such as illnesses, resulting in outliers in the data. Random forest reduces the impact of such outliers through majority voting during prediction.
3) *Efficiency in student performance prediction:* Random forest has been demonstrated to be one of the most effective tools in student performance prediction [8].

We performed 10-fold cross-validation and averaged the performance over three random seeds. The evaluation metrics used were accuracy and recall to assess prediction performance.

In addition to evaluating the prediction performance of the trained random forest model, we aim to identify the importance of individual features within the collected data. Understanding feature importance allows us to explore the factors contributing to why some students struggle in ECE 301. For this analysis, we use SHapley Additive exPlanations (SHAP) [16] to analyze the feature importance for the trained random forest model. Compared to other methods, such as Gini importance or permutation importance, SHAP offers greater robustness to correlations between

Table 3: Longitudinal Enrollment and Student Characteristics

Term	Overall Enrollment	% Women	% Under-represented	% Transfer
Fall 2014	151	11.92	15.89	44.37
Spring 2015	111	15.32	25.23	43.24
Fall 2015	112	11.61	26.79	37.50
Spring 2016	72	18.06	15.28	37.50
Fall 2016	102	15.69	26.47	38.24
Spring 2017	71	16.90	19.72	45.07
Fall 2017	88	25.00	29.55	36.36
Spring 2018	83	14.46	21.69	33.73
Fall 2018	83	22.89	24.10	38.55
Spring 2019	75	14.67	24.00	40.00
Fall 2019	71	14.08	21.13	36.62
Spring 2020	52	26.92	23.08	26.92
Fall 2020	75	24.00	22.67	26.67
Spring 2021	61	16.39	14.75	27.87
Fall 2021	60	26.67	18.33	28.33
Spring 2022	59	15.25	16.95	40.68
Fall 2022	74	14.86	24.32	32.43
Spring 2023	67	14.93	22.39	44.78
Fall 2023	61	13.11	29.51	39.34
Spring 2024	64	26.56	28.12	25.00
Total	1592	17.34	22.55	37.00

prerequisite features, making it a more reliable tool for our analysis.

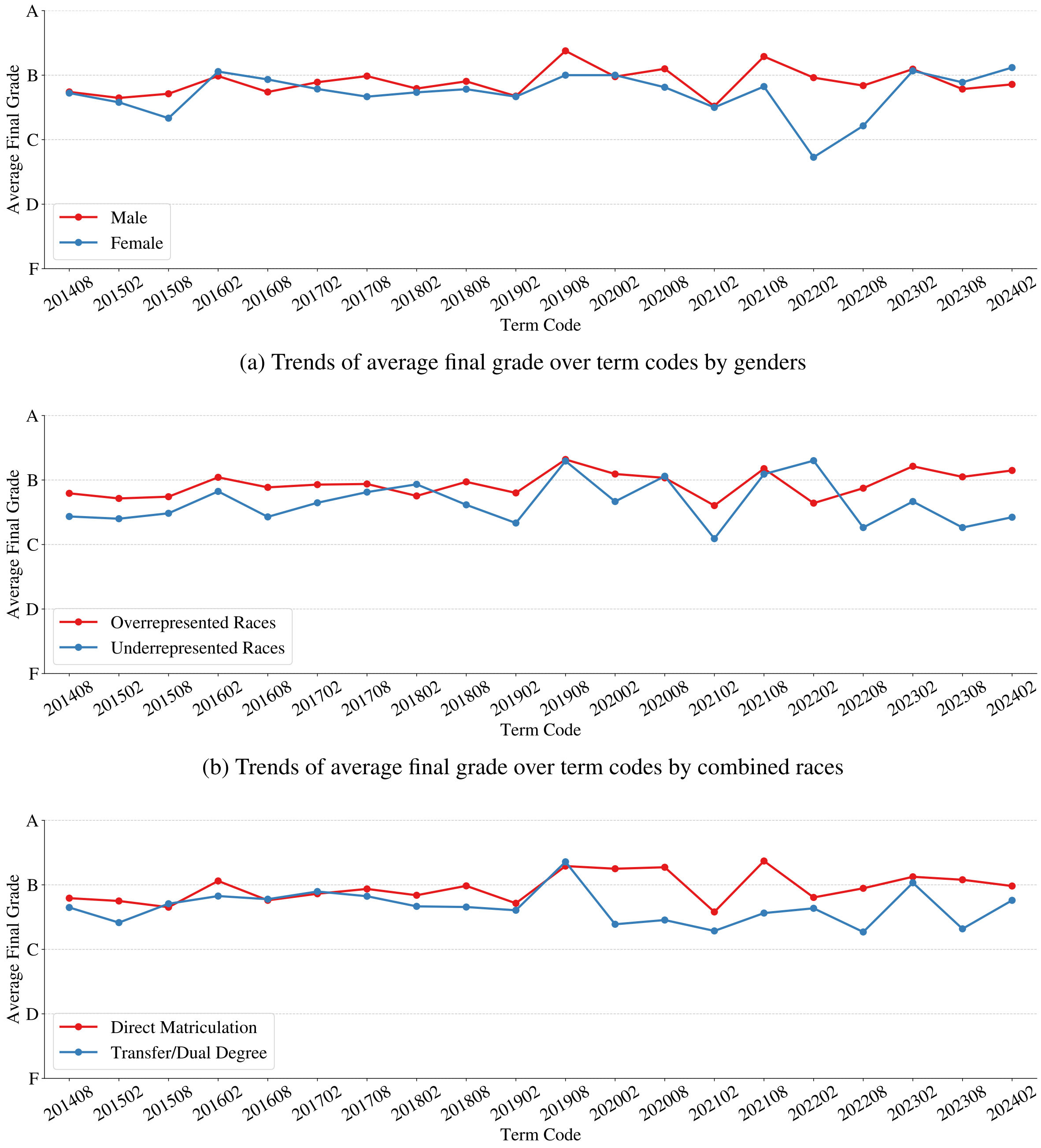
Results

Basic Statistical Information of the Dataset

Between Fall 2014 and Spring 2024, a total of 1,592 students received final grades in ECE 301. Below, we first present count data detailing the composition of students enrolled in ECE 301. We then analyze how final grades varied based on student characteristics, course characteristics, and prior academic performance.

Count Information

Table 3 disaggregates enrollment numbers over time by gender, race/ethnicity, and transfer status. Overall, women made up 17.3% of the course population, students from minoritized racial or ethnic backgrounds made up 22.5%, and transfer students constituted 37.0% on average, though these proportions have fluctuated over time. Because of the challenges of presenting the three-dimensional intersections of these categories, Figure 1 provides separate line charts showing final grade performance by gender, race/ethnicity, and transfer status. Figure 1(a) indicates that women



(a) Trends of average final grade over term codes by genders

(b) Trends of average final grade over term codes by combined races

(c) Trends of average final grade over term codes by transfer statuses

Figure 1: Trends of average final grade over term codes by various demographic and status groups

generally performed at the same level as their male peers, except during the Spring and Fall 2022 semesters, after which their average performance returned to parity. Figure 1(b) reveals that, starting in Fall 2022, students from minoritized racial backgrounds experienced lower performance. These trends are not easily attributed to course characteristics (e.g., modality or instructor), suggesting the need for further investigation. Finally, as shown in Figure 1(c), transfer students performed comparably to directly matriculated students before Spring 2020. However, a performance gap emerged during that semester, persisted until it closed in Spring 2023, and then began to fluctuate over the past two terms. This finding inspired us to treat directly matriculated students and transfer students as separate groups for comparison and analysis.

ECE 301 Performance

Figure 2 presents the ECE 301 grade distribution from Fall 2014 to Spring 2024. It shows that about 30–50% students received grades lower than a “B”, with this percentage varying across terms. Among these underperformed students, the majority received a “C,” except in Fall 2019, which exhibited a different pattern.

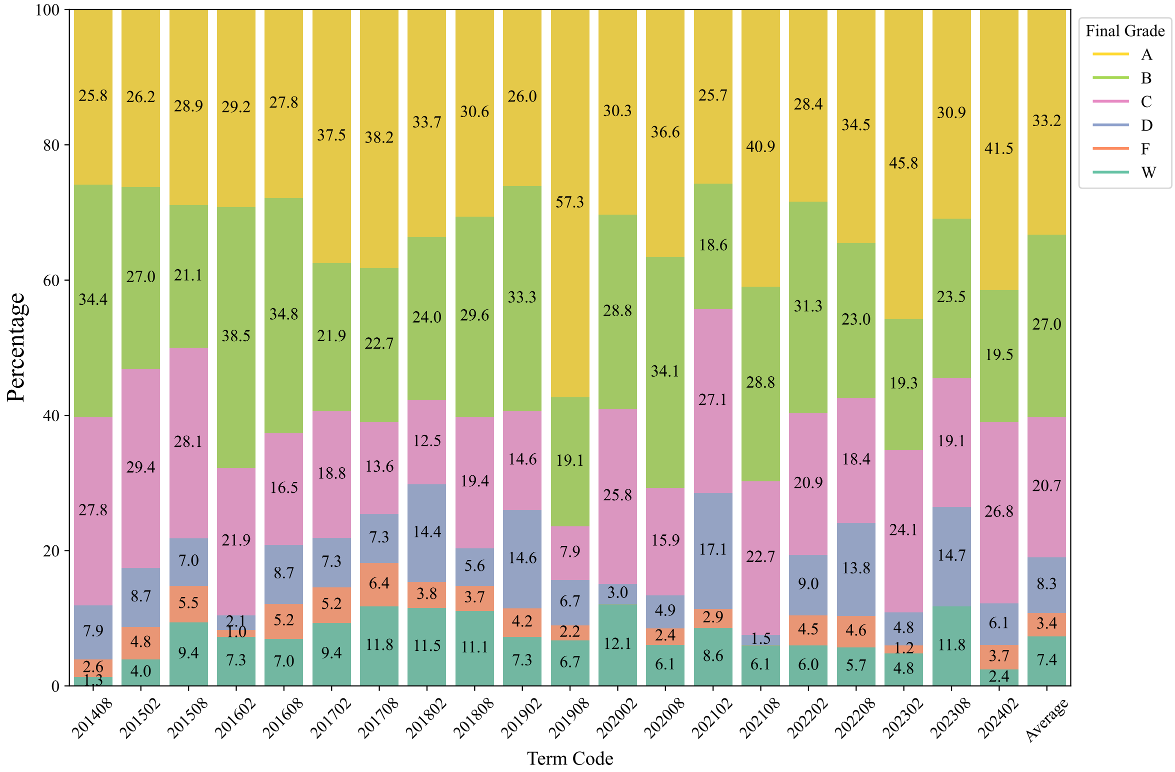


Figure 2: Percentage of students with each grade by term code

Pre-requisite Courses

Figure 3 shows the percentage of transfer students who transferred the corresponding pre-requisite course from another institution. It indicates that most transfer students transferred credits for Physics II, Calc3, and DiffEQ from their previous institutions, while 27% students transferred credits for Circuit Analysis. A high proportion of transferred pre-requisite courses reduces data quality, as the institution records only a “T” for these courses instead of in-institution letter grades (e.g., “A” or “D”). In contrast, courses such as Digital System Design, Digital Design, and Programming are less impacted by transfer credits.

Figure 4 illustrates the average number of terms students spent after enrolling at the institution before taking specific courses. For directly matriculated students, Physics II and Digital System Design are typically taken after the first term, Calc3 and DiffEQ after the second term, and Programming, Circuit Analysis, and Digital Design after the third term. ECE 301 is generally taken after the fifth term. For transfer students, the sequence is slightly altered, as some courses, such as Physics II and Circuit Analysis, are often transferred before enrollment. This timeline provides valuable insights for designing time-based stages to identify and support students earlier in their academic journey.

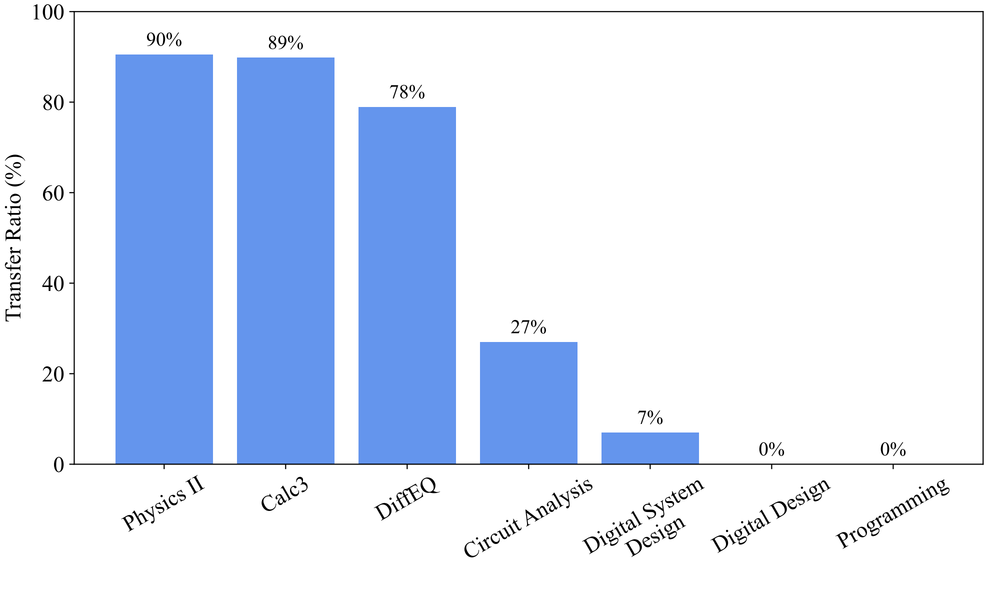


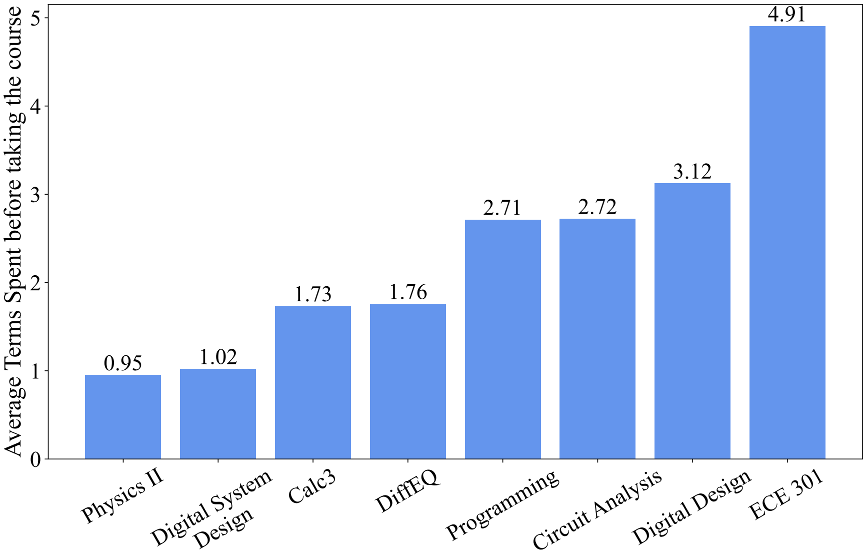
Figure 3: The transfer ratio of different pre-courses within transfer students

Performance Correlation

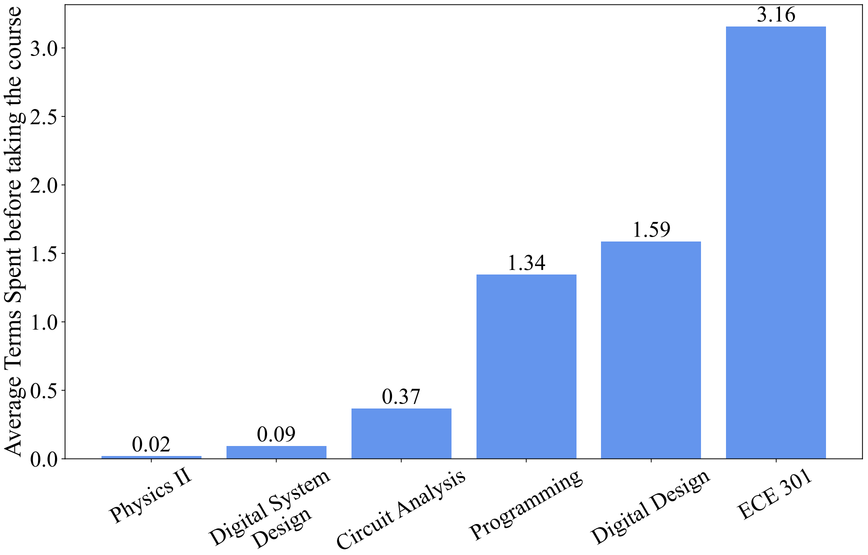
Figure 5 presents the correlation between students' high school information, their pre-requisite course performance, and their performance in ECE 301. The correlation matrix reveals that many pre-requisite courses have positive relationships with ECE 301, suggesting that strong performance in these courses is associated with success in ECE 301. Notably, cumulative GPA exhibits the strongest correlation with ECE 301 performance. In contrast, the relationship between students' high school information (e.g., SAT Math scores and AP count) and ECE 301 is weaker. This is unsurprising, as high school academic information is temporally distant from ECE 301 and reflects a more basic level of knowledge. Consistent with this, subsequent feature importance analyses (Figures 6 and 8) confirm that these two high school features are among the least influential predictors of ECE 301 performance.

Prediction and Analysis

For this analysis, we sampled students with complete records of the aforementioned pre-courses, yielding 540 directly matriculated students and 249 transfer students. For these two groups, we iterated through each stage defined in Tables 1 and 2, testing all possible feature combinations for training the random forest model. For each combination, we set the number of trees to 300

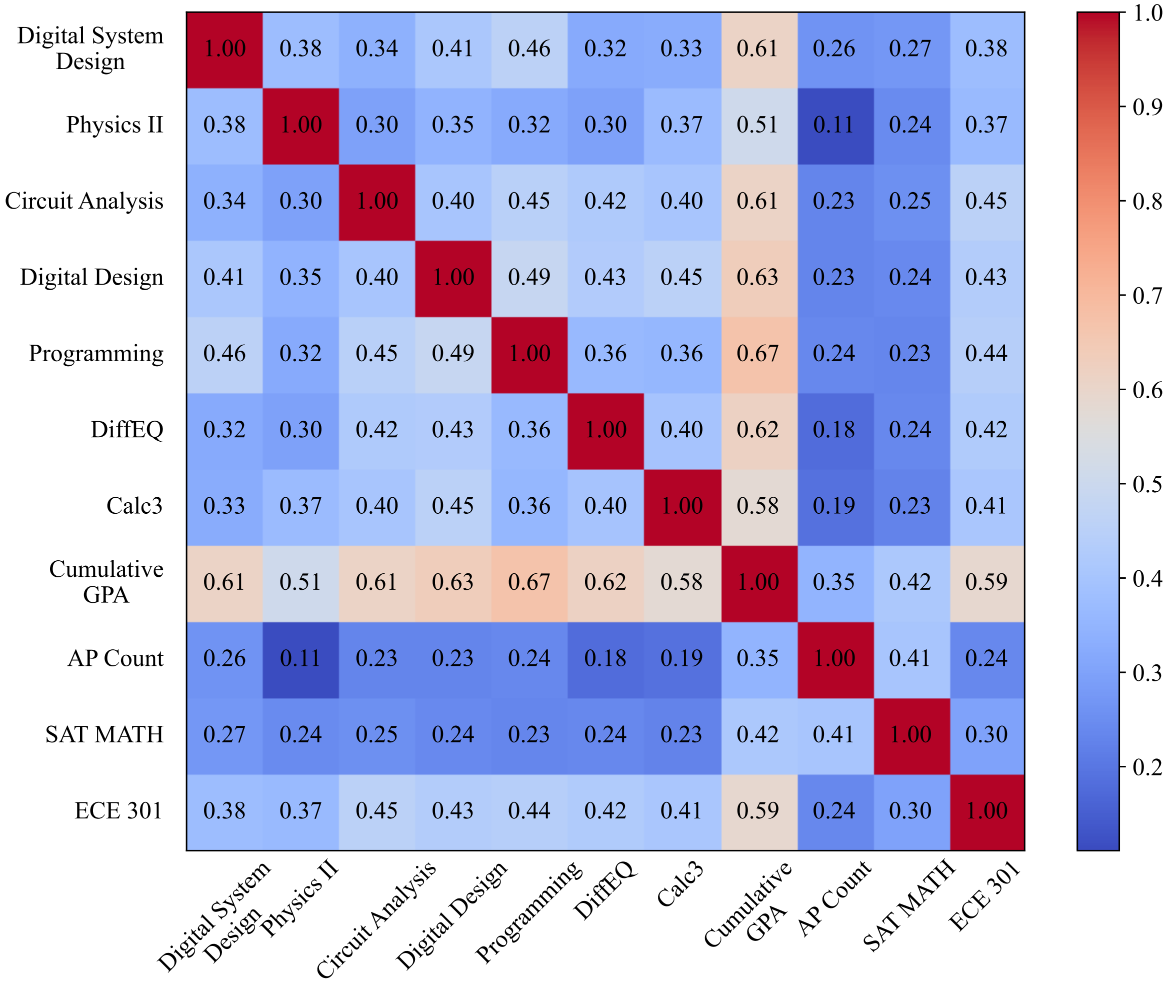


(a) Directly matriculated student



(b) Transfer student

Figure 4: The average term (including summer terms) spent for students (after enrolling in the institution) before taking certain courses recorded in our data



	Digital System Design	Physics II	Circuit Analysis	Digital Design	Programming	DiffEQ	Calc3	Cumulative GPA	AP Count	SAT MATH	ECE 301
Digital System Design	1.00	0.38	0.34	0.41	0.46	0.32	0.33	0.61	0.26	0.27	0.38
Physics II	0.38	1.00	0.30	0.35	0.32	0.30	0.37	0.51	0.11	0.24	0.37
Circuit Analysis	0.34	0.30	1.00	0.40	0.45	0.42	0.40	0.61	0.23	0.25	0.45
Digital Design	0.41	0.35	0.40	1.00	0.49	0.43	0.45	0.63	0.23	0.24	0.43
Programming	0.46	0.32	0.45	0.49	1.00	0.36	0.36	0.67	0.24	0.23	0.44
DiffEQ	0.32	0.30	0.42	0.43	0.36	1.00	0.40	0.62	0.18	0.24	0.42
Calc3	0.33	0.37	0.40	0.45	0.36	0.40	1.00	0.58	0.19	0.23	0.41
Cumulative GPA	0.61	0.51	0.61	0.63	0.67	0.62	0.58	1.00	0.35	0.42	0.59
AP Count	0.26	0.11	0.23	0.23	0.24	0.18	0.19	0.35	1.00	0.41	0.24
SAT MATH	0.27	0.24	0.25	0.24	0.23	0.24	0.23	0.42	0.41	1.00	0.30
ECE 301	0.38	0.37	0.45	0.43	0.44	0.42	0.41	0.59	0.24	0.30	1.00

Figure 5: The correlation matrix among features in the collected student data

and performed a grid search over three random forest hyperparameters: `max_depth` (3, 5, 7, 10), `min_samples_split` (5, 10, 15), and `min_samples_leaf` (5, 10, 15). We reported the best 10-fold average prediction performance (accuracy and recall) across three random seeds (1000, 2000, 3000) and recorded the corresponding hyperparameters, as shown in Tables 4 and 5.

Table 4: Different stages' best hyperparameters for directly matriculated students

Stage	**Max Depth**	**Min Samples Split**	**Min Samples Leaf**
1	3	5	5
2	10	5	15
3	7	5	15

Results for Directly Matriculated Students

Table 6 presents the best feature combinations and corresponding performance metrics for each

Table 5: Different stages' best hyperparameters for transfer students

Stage	Max Depth	Min Samples Split	Min Samples Leaf
1	5	5	10
2	3	5	5
3	3	5	15

stage. The results show a clear performance improvement as the stages progress, with more prerequisite courses being considered. This indicates that incorporating more recent information, such as grades from Circuit Analysis and Programming, allows for more accurate predictions of students' future performance, achieving an accuracy of 80.91%. Notably, while slightly lower, the early-stage prediction performance remains impressive, with an accuracy of 77.37%. This finding underscores the potential to identify and support students at risk of struggling in a middle-level course several terms in advance.

Table 6: Each stage's best prediction accuracy and recall and the corresponding feature combination for directly matriculated students. All combinations include high school information (i.e., SAT Math score, AP count) by default.

Stage	Feature combination	Accuracy	Recall
1	Digital System Design, Physics II, cumulative GPA	77.37%	76.73%
2	Stage 1 + DiffEQ	78.79%	77.41%
3	Stage 2 + Circuit Analysis + Programming	**80.91%**	**78.11%**

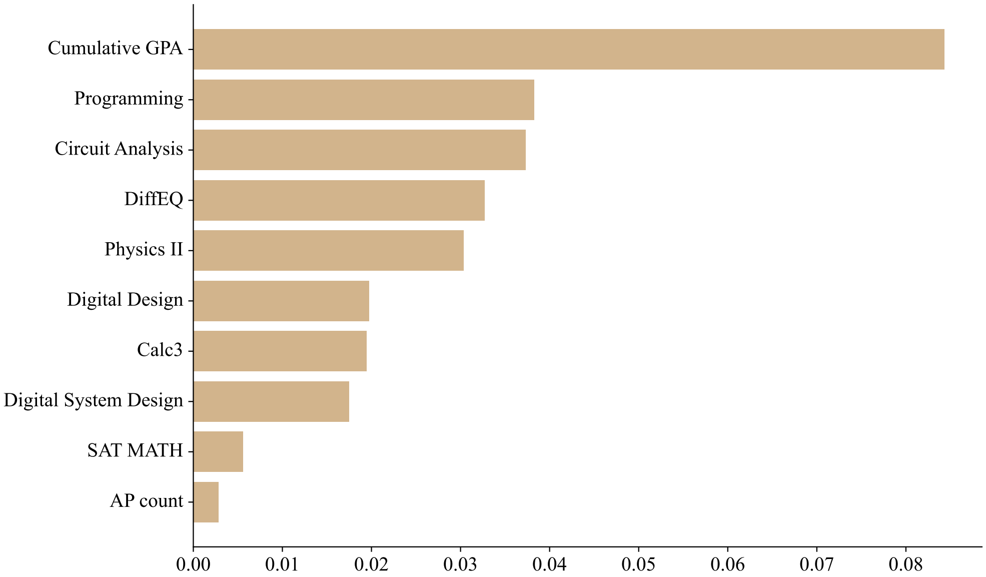


Figure 6: Feature importance for directly matriculated students.

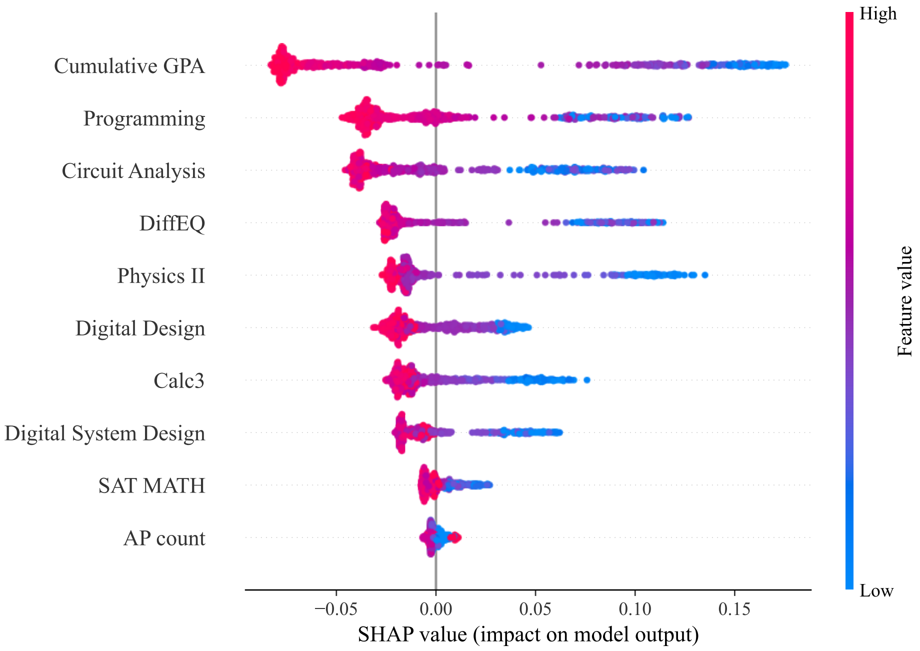


Figure 7: SHAP value of different data points in directly matriculated students

Figure 6 shows the SHAP-based feature importance for all input features in the final stage. Besides cumulative GPA, those higher-level course features (e.g., Circuit Analysis and Programming) play a significant role in predicting students' ECE 301 performance, with the exception of Digital Design. This finding aligns with the feature combination results in Table 6, which does not include Digital Design but covers two other higher-level course features. Conversely, high school academic information, such as SAT Math scores and AP count, shows lower importance. This observations matches our finding of the weak positive correlation between these features and ECE 301 performance in Figure 5. Figure 7 provides a more detailed view of feature importance across different input features. Consistent with Figure 6, it reveals that the impact (SHAP values) of cumulative GPA and higher-level courses is larger and more clearly distributed as these features vary across student samples.

Results for Transfer Students

For transfer students, the model performance improves at higher stages, similar to the results for directly matriculated students (Table 6). However, the accuracy for transfer students is lower compared to directly matriculated students. This discrepancy can be attributed to poorer data quality and fewer available input features. Many prerequisite courses for transfer students, such as Circuit Analysis and Physics II, are recorded as transfer credits ("T"), which provide less detailed information than letter grades (e.g., "A" or "B") available for directly matriculated students. Despite these limitations, the prediction performance in the early stage remains acceptable, achieving 72.26% accuracy and 76.21% recall with very limited features. The high recall (such as 82.48% in Stage 2) is particularly significant, as it indicates that most at-risk students can be identified shortly after completing their first or second terms.

Table 7: Each stage's best prediction accuracy and recall, and the corresponding feature combination for 249 transfer students. All combinations include high school information (i.e., AP count) by default, SAT Math score is not included as we don't have records for about 50% transfer students.

Stage	Feature combination	Accuracy	Recall
1	Physics II, Digital System Design, cumulative GPA	72.26%	76.21%
2	Stage 1 + Circuit Analysis	72.86%	82.48%
3	Stage 2 + Programming + Digital Design	**74.17%**	**85.39%**

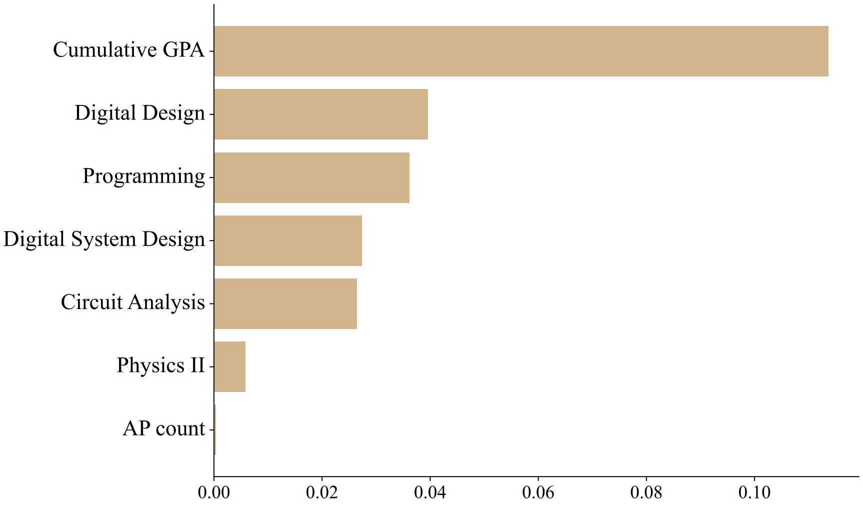


Figure 8: Feature importance for transfer students

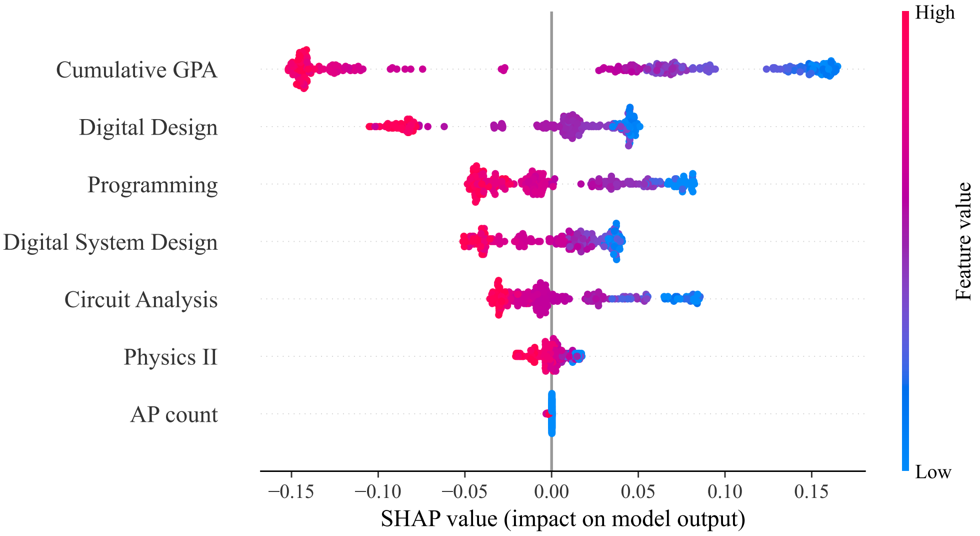


Figure 9: SHAP value of different data points in transfer students

Figures 8 and 9 present the feature importance and SHAP analysis for the input features in Stage 3. High school information, such as AP count, remains the least important feature. Interestingly, "Circuit Analysis," which is an important feature for directly matriculated students, becomes less significant for transfer students. This can be attributed to the fact that many transfer students bring in "Circuit Analysis" credits from their original institutions (as shown in Figure 3). In contrast, Digital Design and Programming are two courses for which students cannot bring transfer credit (as indicated by a 0% transfer ratio in Table 3). These courses play a crucial role in evaluating transfer students' performance.

Discussion

Our stage experiments, summarized in Tables 6 and 7, demonstrate that students' early academic information can serve as a powerful predictor for their success in higher-level courses, even several terms later. By leveraging this predictive power, institutions can more effectively identify students who may be at risk of underperforming, allowing for the implementation of timely interventions and additional support, such as tutoring, mentoring, or personalized academic advising. These efforts can ultimately help students improve their academic outcomes and progress smoothly through their educational journey.

At the same time, we must be mindful of the ways in which this information is used and made public. While predictive analytics hold great promise in enhancing student support, it is important to avoid inadvertently reinforcing biases or stigmatizing individuals, especially over matters they might not have had much control over [17]. The messaging and framing of these developments is key to improve the relevance, accuracy, and context of big data interventions [18].

In practical terms, the importance of features is often more valuable than the accuracy of the model's predictions. By analyzing feature importance, we can identify key factors influencing student success and implement targeted actions to improve outcomes. For instance, Figure 6 highlights the significance of Physics II, an early-stage prerequisite, for ECE 301 performance. This finding suggests that improving the teaching quality of Physics II could positively impact student success in ECE 301. Additionally, the differences in feature importance between directly matriculated and transfer students offer intriguing insights that warrant further investigation. As shown in Figure 3, this is partly caused by the transfer credits of other prerequisite courses, but does it still suggest the potential in investigating the differences between directly matriculated students and transfer students more deeply, as well as the corresponding analytical systems.

In the future, we plan to investigate the predictive accuracy of student data across additional ECE major courses and earlier foundational courses. Additionally, based on a wealth of evidence supporting the predictive capabilities of student attitudes [19], we also hope to take advantage of our in-house analysis to complement the academic record data with non-cognitive survey data. By providing timely, personalized support to at-risk students, we aim to promote greater academic success and improve outcomes throughout their educational journey at the institution.

Conclusions and Implications

In this work, we aim to understand student performance in a junior-level microelectronics course (ECE 301) at a large, public, research-intensive institution in the Southeastern United States. We collected 10 years of academic data, including students' high school records, prior institution data (for transfer students), and grades from prerequisite courses. Students were categorized into directly matriculated and transfer groups, and their data were analyzed across different time stages to predict future performance in ECE 301. Using this data, we constructed a dataset and trained a random forest model to predict student performance in ECE 301. The results demonstrate that at-risk students can be accurately identified at an early stage, offering promising opportunities for timely intervention by educators. To further understand the predictors of student success, we applied SHAP to analyze feature importance and identified several critical prerequisite courses. These findings highlight the potential to enhance educational quality through model-driven feedback. Future work will focus on standardizing the current student performance evaluation system across different engineering courses and time stages, as well as exploring methods to design personalized support for at-risk students based on system feedback.

Acknowledgements

The authors appreciate the support provided by the School of Electrical and Computer Engineering and the College of Engineering at the institution where the study was conducted.

References

[1] S. M. Lord, M. W. Ohland, M. K. Orr, R. A. Layton, R. A. Long, C. E. Brawner, H. Ebrahiminejad, B. A. Martin, G. D. Ricco, and L. Zahedi, "Midfield: A resource for longitudinal student record research," *IEEE Transactions on Education*, vol. 65, no. 3, pp. 245–256, 2022.

[2] S.-M. R. Ting and R. Man, "Predicting academic success of first-year engineering students from standardized test scores and psychosocial variables," *International Journal of Engineering Education*, vol. 17, no. 1, pp. 75–80, 2001.

[3] B. F. French, J. C. Immekus, and W. C. Oakes, "An examination of indicators of engineering students' success and persistence," *Journal of Engineering Education*, vol. 94, no. 4, pp. 419–425, 2005.

[4] L. E. Bernold, J. E. Spurlin, and C. M. Anson, "Understanding our students: A longitudinal-study of success and failure in engineering with implications for increased retention," *Journal of engineering education*, vol. 96, no. 3, pp. 263–274, 2007.

[5] J. De Winter and D. Dodou, "Predicting academic performance in engineering using high school exam scores," *International Journal of Engineering Education*, vol. 27, no. 6, p. 1343, 2011.

[6] K. M. Whitcomb, Z. Y. Kalender, T. J. Nokes-Malach, C. D. Schunn, and C. Singh, "Engineering students' performance in foundational courses as a predictor of future academic

success," *International Journal of Engineering Education*, vol. 36, no. 4, pp. 1340–1355, 2020.

[7] J. Rohde, S. P. Karyekar, L. Chen, Y. Guo, and Y. Zhang, "Predictors of student academic success in an upper-level microelectronic circuits course," in *2024 ASEE Annual Conference & Exposition*, 2024.

[8] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student'performance prediction using machine learning techniques," *Education Sciences*, vol. 11, no. 9, p. 552, 2021.

[9] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, pp. 3–14, 2014.

[10] D. A. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, "Student retention using educational data mining and predictive analytics: a systematic literature review," *IEEE Access*, vol. 10, pp. 72480–72503, 2022.

[11] Q. Hu and H. Rangwala, "Reliable deep grade prediction with uncertainty estimation," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pp. 76–85, 2019.

[12] D. Aggarwal, S. Mittal, and V. Bali, "Significance of non-academic parameters for predicting student performance using ensemble learning techniques," *International Journal of System Dynamics Applications (IJSDA)*, vol. 10, no. 3, pp. 38–49, 2021.

[13] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[14] B. D. Jones, M. C. Paretti, S. F. Hein, and T. W. Knott, "An analysis of motivation constructs with first-year engineering students: Relationships among expectancies, values, achievement, and career plans," *Journal of engineering education*, vol. 99, no. 4, pp. 319–336, 2010.

[15] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[16] S. Lundberg, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

[17] B. Williamson, S. Bayne, and S. Shay, "The datafication of teaching in higher education: critical issues and perspectives," 2020.

[18] C. Klein, J. Lester, T. Nguyen, A. Justen, H. Rangwala, and A. Johri, "Student sensemaking of learning analytics dashboard interventions in higher education," *Journal of Educational Technology Systems*, vol. 48, no. 1, pp. 130–154, 2019.

[19] M. Besterfield-Sacre, C. J. Atman, and L. J. Shuman, "Characteristics of freshman engineering students: Models for determining student attrition in engineering," *Journal of Engineering Education*, vol. 86, no. 2, pp. 139–149, 1997.