Can AI Transform Graduate Computer Science Admissions? Preemptive Bias Detection in Automated Review Systems

Ananya Prakash, Virginia Polytechnic Institute and State University

Ananya Prakash is a Master's in Computer Science graduate from Virginia Tech. She is an interdisciplinary researcher with interests in Machine Learning, Natural Language Processing, CS Education and AI Ethics.

Mohammed Seyam, Virginia Polytechnic Institute and State University

Mohammed Seyam is a Collegiate Associate Professor in the Computer Science Department at Virginia Tech. He is a researcher and educator in the fields of Software Engineering, Human-Computer Interaction, and Computer Science Education. Additionally, he is the CS Department Coordinator for Experiential Learning, where he leads several initiatives to enhance students' learning through out-of-classroom experiences, including the CS Study Abroad program. Mohammed has 20+ years of experience in teaching university level courses, and he presented and conducted multiple talks and workshops in different countries. Among other courses, he taught: Software Engineering, Database Systems, Usability Engineering, and Software Project Management.

Can AI Transform Graduate Computer Science Admissions? Preemptive Bias Detection in Automated Review Systems

Abstract

The number of graduate students has been increasing rapidly to meet industry demands, with over 200% increase in competitive fields like computer science (CS) in the past decade. This has led to several universities adopting AI in their admissions processes for various tasks such as evaluating transcripts, extracting important information from essays, and scoring applications. Past implementations of AI for decision-making in admissions have often led to issues surrounding bias with the potential to have long-standing effects on diversity. With the minimal change in diversity in graduate CS education over the last decade and the recent removal of affirmative action for admissions, it is critical to evaluate the potential biases that may arise from AI-based admissions. In this paper, we propose that while AI could be leveraged to increase the efficiency of the decision-making process for student admissions, it is imperative to continuously identify and tackle bias that emerges from AI-based systems for admissions. Based on previous work, we identify some key sources of bias in the context of admissions and the role of AI in predicting outcomes and assisting decision-making. We propose a framework to preemptively detect bias that may be inferred by a machine learning model using exploratory data analysis, clustering, subgroup discovery, and feature importance. By implementing our proposed framework on a dataset of applications to two graduate CS programs of an R1 public university, we demonstrate how universities may tackle the challenges of using AI for admissions. Our work provides evidence that demographic features like age, gender, birth nation, and race may lead to inferred bias and highlights the importance of bias detection to create fair AI admissions systems.

1. Introduction

Over the last few decades, jobs in the technology industry have become far more competitive, with more students earning master's and doctorate level degrees for jobs motivated by nearly a 20% higher salary than bachelor's degree holders as per the U.S. Bureau of Labor Statistics [1]. According to the National Center for Education Statistics (NCES) [2], the number of graduates with a master's degree has grown from 14,990 in 2000 to 51,338 in 2019, a 242% increase over two decades. Similarly, the number of graduates with a doctorate has grown from 779 to 2790 in the same period, an increase of 258%. While this increase in pursuits of postgraduate degrees in the field reflects the rapid growth of the industry, universities still grapple with the task of evaluating increasingly large volumes of applications.

Several large universities adopt a holistic review approach for admissions that is time-consuming and relies heavily on skilled human reviewers. The average time taken for each full review could vary between 10-30 minutes based on the skills of the reviewer [3]. A survey conducted by Intelligent in 2023, an education magazine [4], reported that 50% of 400 surveyed institutions already used Artificial Intelligence (AI) in their admissions process, and an additional 30% planned to do so in 2024. AI gives universities the advantage of increased efficiency, allowing them to focus their limited resources on other critical tasks like selecting students for financial

aid and scholarships [5]. Therefore, it is essential to innovate AI systems that assist in the admissions process while still minimizing the possibility of biased outcomes.

The rapid development of the technology industry led to an increased number of graduate degree holders yet the diversity among these graduates has not shown comparable growth. For instance, the male-to-female ratio among master's graduates has remained nearly constant in the last decade at 2:1 with 66% males and 33% females [2]. This supports Cuny and Aspray's observation that fewer women enroll in graduate computer science (CS) programs, with numbers dropping from master's to doctorate levels [6]. Despite the time gap between the 2002 study [6] and the NCES report from 2023 [2], the findings align, emphasizing the persistence of the issue.

Research shows that the diversity gap is further exacerbated for graduate degrees because of issues like lower GPAs or poor undergraduate experiences, low access to resources for standardized tests for underrepresented communities, and financial limitations [7]. Studies by the NCES also reveal that students of underrepresented minorities (URM) including Hispanic, Black, and American Indian or Alaskan Native students constitute a far lower percentage of graduate students compared to White and Asian students in science and engineering fields in the United States [8]. With the Supreme Court's decision to ban affirmative action in 2023 [9], a policy that earlier allowed universities to consider race as a criterion in the admissions process, there may be an increased threat to diversity, given its already deficit state in CS graduate education. With the minimal change in diversity over the last decade and the ban on affirmative action practices, it is imperative to find a solution to the challenge of diversity.

Despite AI's potential to optimize efficiency and reduce the workload of human reviewers, it also presents new risks. One such risk is bias, a phenomenon exhibited in AI systems that can amplify and perpetuate undesirable negative effects on individuals, organizations, and society [10]. The survey [4] also found that universities typically use AI to review letters of recommendation, transcripts, and essays and to communicate admission decisions to applicants. Though AI may be causing little to no harm in analyzing objective criteria like transcripts, previous studies have highlighted its ability to learn sensitive attributes like gender from letters of recommendation [11] and gender and household income from personal essays [2], which could potentially induce bias in the admissions process. To fully optimize the admissions process, machine learning systems may be employed to make final decisions on applications, as done by nearly 87% of the survey respondents [4]. However, institutions must pay careful attention to the details of how their model is trained.

While most large public universities already employ a rubric or some algorithms to make admissions decisions [5], machine learning models might learn unintended patterns from historical data that further perpetuate biases. This study aims to uncover the potential biases that a machine learning model trained on historical data may infer in its training phase, which we will henceforth refer to as inferred bias. The rapid adoption of machine learning for admission reviews reported in [4] highlights the importance of carefully analyzing what the machine learning model may infer during training to prevent any unprecedented biases from being perpetuated by the model. We examine and evaluate the potential for such inferred biases by investigating the following research questions:

RQ 1: How has the affirmative action ban impacted demographic diversity in graduate computer science education?
RQ 2: What are the independent features that may increase the likelihood of a positive or negative decision?
RQ 3: What are the intersectional attributes that may lead to bias in the model?

By studying these questions, this paper aims to demonstrate how potential inferred biases can be discovered before applying machine learning models to automate admission by analyzing trends and distributions in the data, clustering, subgroup discovery, and feature importance. While we focus on detecting inferred bias in the context of graduate admissions for computer science, these methods can be extended to the admission processes of various programs and universities. By performing similar systematic bias discovery methods, universities may develop machine learning solutions that improve the efficiency of admissions reviews, decrease the possibility of biased results, and encourage diversity in graduate computer science education. Thus, this paper intends to present opportunities for universities to increase the efficiency of the admissions review process while minimizing potential harm to diversity emerging from the application of machine learning systems for decision-making in the context of graduate admissions.

2. Background and Related Work

This work focuses on detecting potential biases that may be inferred by a machine learning system developed to determine admission decisions for applications. In the following subsections, we will explore some of the necessary background that is pertinent to this study.

2.1 Admission decision pipeline and sources of bias

The admission process in U.S. universities typically requires applicants to submit various materials for evaluation such as transcripts, personal essays, multiple letters of recommendation, standardized test scores, and additional materials like extra-curricular certificates. When applying to graduate programs, applicants often submit additional essays that help the university understand their qualities, experience, and beliefs. These include essays on leadership, academic research, community service, and personal and professional ethics. Therefore, the data consists of numerical features such as standardized examination scores and Grade Point Averages (GPA), along with textual data from the essays and letters of recommendation. Applications also collect personal information including but not limited to the applicant's name, address, gender, and ethnicity. Figure 1 details the potential stages in the admissions pipeline where bias could emerge and where AI is currently used as per the Intelligent survey [4].

In the context of university admissions, features like gender and ethnicity are usually examined for bias, as done by Kahlor et al. [13]. Contrarily, several existing studies on machine learning systems for admissions also seem to exclude demographic features to remove bias [14, 15]. This could be due to these studies simulating the admission process in California, where Proposition 209[1] prohibits the use of gender and ethnicity for admission reviews. In the GRADE system [3], the authors note that gender and ethnicity were assigned zero weight when passed to the model

[1] California Proposition 209: https://lao.ca.gov/ballot/1996/prop209_11_1996.html

as features, concluding that demographic features do not contribute to the model's decisions. Another study analyzing the effect of the test-optional policy on admission decisions [16] examined bias along the features of gender, ethnicity, and first-generation applicants.

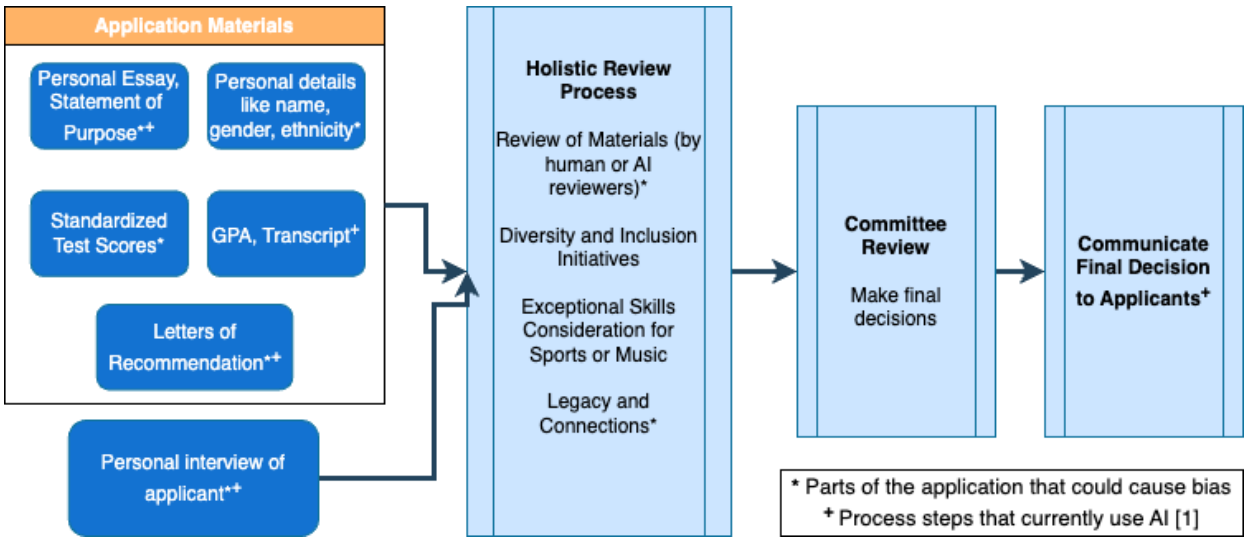


Figure 1. Admission Decision Pipeline

Though gender and ethnicity are the most commonly considered sensitive features when examining bias, other features such as nationality, first-generation students and median household income may also induce bias. Another occurrence of bias that is often neglected is intersectional feature bias of observations that belong to more than one protected group, i.e., as a combination of multiple sensitive attributes [17]. This can be identified by scanning the dataset for subgroups with increased bias [18]. The holistic review process involves various application materials, with one of the primary mediums for students to share their narratives being essays. However, these essays often contain personal stories that may reveal demographic information related to an applicant as found in a study [12], creating a potential for bias. Similarly, sensitive attributes of the applicant may also be inadvertently extracted from letters of recommendation [11], proving to be another potential source of bias.

2.2 Prior Work on AI in Admissions

An important and necessary precursor to identifying bias in the admission process is understanding the different ways in which AI has been previously applied to tackle admissions. While numerous studies have examined the feasibility of using AI in the admission process, there is limited structured work that compares and categorizes the various applications. To effectively describe the different applications, we classified these systems into two main categories as shown in Figure 2:

- AI-predicted decision-making: We refer to AI-predicted decision-making as the process where an AI algorithm is directly used to predict the admission decision for a given observation containing an applicant's information.

- AI-assisted decision-making: We define AI-assisted decision-making as the process where AI is used in the admission review pipeline to aid human evaluators. This involves making more information available for decision-makers, such as by validating essay scores or extracting and comparing transcript information.

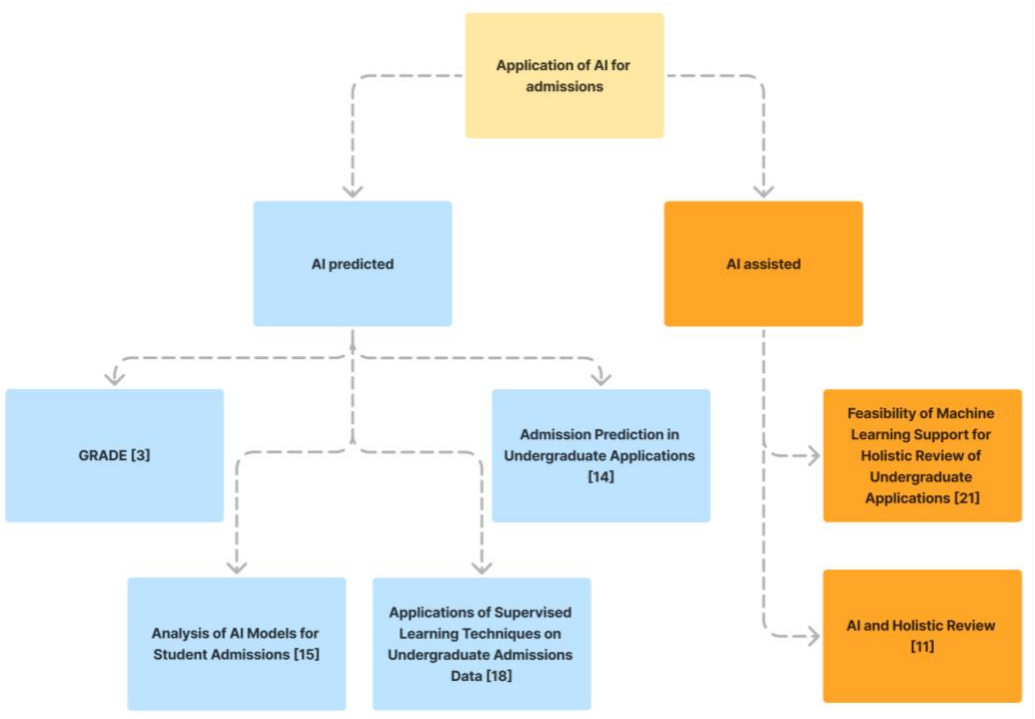


Figure 2. Categorization of AI applications for admission review

Several studies have explored the domain of AI-predicted decision-making for admissions, such as the GRADE system [3], deep learning algorithms to predict undergraduate admissions [15], and a Multilayer Perceptron and Support Vector Machine implementation [19]. Another study [16] developed an AI model to predict admission decisions with the test-optional policy. [15] and [19] developed a classifier trained on historical data to determine admission offers and used feature selection to identify feature contributions to the classifier's output. Both these works attempted to identify which features contribute more to the decisions, with the first study directly making inferences from learned weights of the classifier and the second study using LIME [20].

While most researchers argue that using AI for admission prediction reduces personal biases from human evaluators' decisions, training on historical data could also produce biased outcomes and lead to a vicious feedback loop between the data and algorithm [21]. A few studies have also applied AI to create systems that aid decision-making in admissions. In an experimental study to increase the efficiency of the holistic review process while sensitizing reviewers, Alvero et al. [12] used AI to extract hidden sensitive attributes in personal essays. Another interesting use case of AI to support holistic review is its application in validating review scores assigned by application reviewers [14]. Both these studies demonstrate the possibility of AI systems to aid admissions decision-making but require further research and experimentation to prove their effectiveness. Therefore, AI can be applied to directly make admissions decisions or as a

supporting tool to aid human reviewers in decision-making. Both these applications offer benefits but must be implemented carefully to ensure that they do not deliver biased outputs.

3. Proposed Framework to Detect Inferred Bias

This study aims to uncover the potential biases that a machine learning model trained on historical data may learn in its training phase. As established in the introduction, while admissions decisions are made holistically through different rubrics specific to each university, they are often made by human reviewers, who not only evaluate objectively but also look for qualities that indicate student success in the university. When a machine learning model trains on historical data, it can potentially learn patterns in admissions outcomes as rules for decision-making, regardless of whether the admissions committee intended for these to be the rules. We refer to this as inferred bias, wherein a machine learning model develops biases in its algorithm while training based on inferences from the data.

We introduce a framework to systematically detect potentially inferred biases, as shown in Figure 3. Once the data is preprocessed, exploratory data analysis (EDA) is performed in phase 1 to reveal distributions and trends. This is crucial to understanding relationships between the features and the decision variable, as well as other patterns in the data that may be useful to analyze our results. As part of the EDA, we also perform clustering to identify potential subsets in the data to further extract patterns in subgroups of the dataset that may lead to inferred biases. The second phase involves searching for intersectional feature bias through subgroup discovery. Subgroup discovery is used to identify a combination of features that could have a higher tendency toward a particular decision outcome [17]. Phase 3 includes training a machine learning model to extract the feature importance of the dataset features and subsequently identify demographic features that the model considers important but may not be typically prioritized by an admissions review committee.

This framework would help ascertain how bias might be introduced into machine learning models, by comprehending the data, feature linkages to the admission decision, intersectional feature bias, and the early effects of the affirmative action ban. Using this framework, researchers and universities may analyze the induction of bias to a model and potentially rectify it to mitigate bias propagation, while still increasing the efficiency of admissions committees by using AI for admissions. Figure 3 displays the proposed framework and relates it to corresponding research questions in this study.

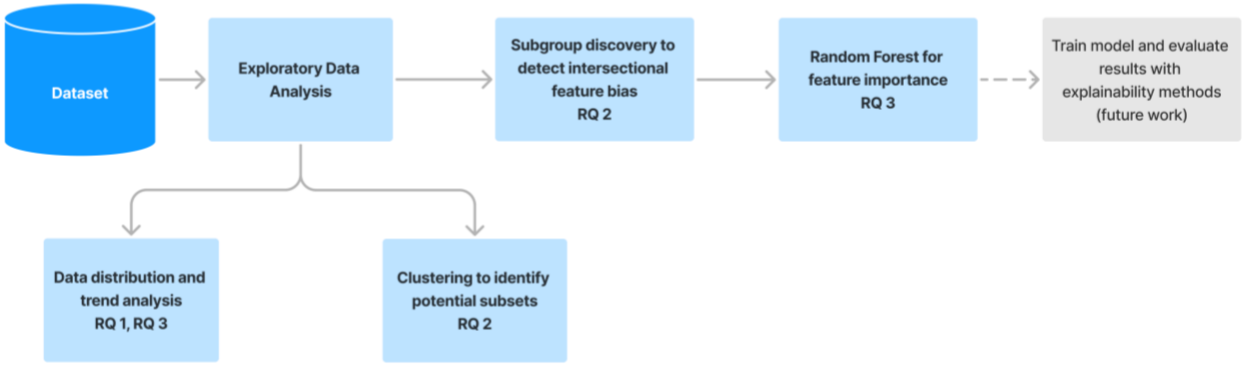


Figure 3. Detection of Inferred Bias

4. Experiments and Findings

4.1 Dataset

The dataset consists of 14850 observations with 11 features of graduate applications to a competitive Computer Science department of an R1 public university in the United States. These are from applications to three distinct graduate-level programs: a research-focused Master of Science (MS) program, an industry-focused Master of Engineering program (MEng), and a Doctor of Philosophy (PhD) program. The data spans a period of ten years from 2014 to 2024 for the MS and PhD programs but was only available for 2 years for the MEng program. To maintain uniformity in results, we dropped all observations of the MEng program. Our final dataset consists of the 9315 observations of the MS and PhD programs available from 2014-2024.

Though application data consists of standardized test scores, GPA, essays, personal statements, and research experience for graduate applications, the scope of this study is limited to non-essay data. The data was preprocessed to impute missing values with mean for numerical features and median for categorical features. Optional fields such as ‘Birth Nation', ‘Race’, ‘Current or Former Military’, and ‘First Generation’ were imputed with placeholder values such as ‘Unknown’ and -1 to indicate that they were opted out. Since the applicants had undergraduate education from various countries, the GPA scores were rescaled to the standard U.S. 4.0 scale for uniformity. The GRE, TOEFL, and IELTS scores were aggregated from the initial dataset, which contained separate features for each subsection of these standardized tests. Unlike the centralized application reviews done for undergraduate applications, the graduate admissions process for the data-providing institution consists of decentralized application reviews from the CS department where reviewers examine the applicant information along with essays and supporting documents submitted, to determine admission decisions. The university uses volunteer readers to score essays, and the average of multiple reviews is considered for evaluation along with the remaining application details.

4.2 Exploratory Data Analysis

4.2.1 Data Distributions

The first stage of EDA involves examining the distributions of various demographic features and their contribution to the subset of accepted and rejected applications. The features used from the dataset can be categorized into demographic and non-demographic features as shown in Table 1.

Table 1. Features classified as demographic and non-demographic

Demographic	**Non-Demographic**
Person Age	GPA
Gender	IELTS
Birth Nation	TOEFL
Citizenship (Primary)	GRE
Race	Decision
Current or Former Military	
First Generation	

Since the paper is focused on determining potential bias upon applying a machine learning model to the data, we thoroughly examine the distributions of various features categorized as demographic in Table 1. The dataset consists of applications to two graduate programs in the department of computer science, with 6317 observations belonging to the MS degree and 2998 observations belonging to the PhD degree. Some of the demographic features like 'Current or Former Military' and 'First Generation' only had a small subset of values filled in, where 8.6% of applications reported as first-generation students and 0.6% of applications reported having military experience. This minority makes it challenging to include these features in different analyses since a majority of the values were unfilled and had to be imputed.

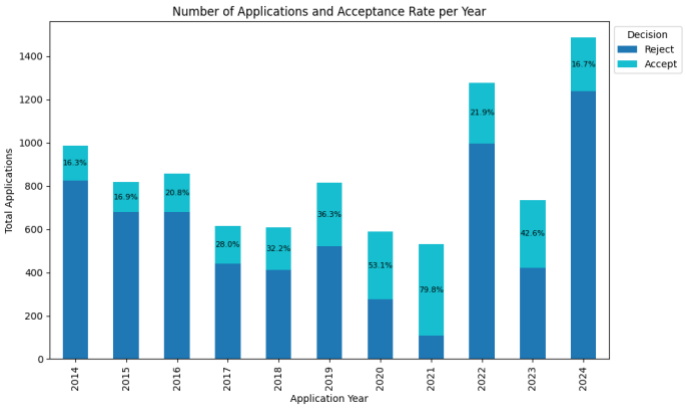


Figure 4. Applications acceptance rate

Figure 4 reveals the acceptance rate of the CS graduate programs of our case university. The average acceptance rate is around 36% with significantly higher acceptance in 2019, 2020, 2021,

and 2023. This might be related to the lower number of applications received during the pandemic or due to the expansion of the programs in recent years.

Since this study focuses on graduate data, notably, there is a wide range of age values in the population, as many applicants may have had some industry experience or completed multiple degrees before applying for their graduate degree. By visualizing the distribution as shown in Figure 5, we cannot draw any apparent relationship between the age and the application decision, since the accepted and rejected observations seem to span the entire range. Through further analysis in the following sections, we can examine if there may be a deeper relationship or potential bias with the age feature.

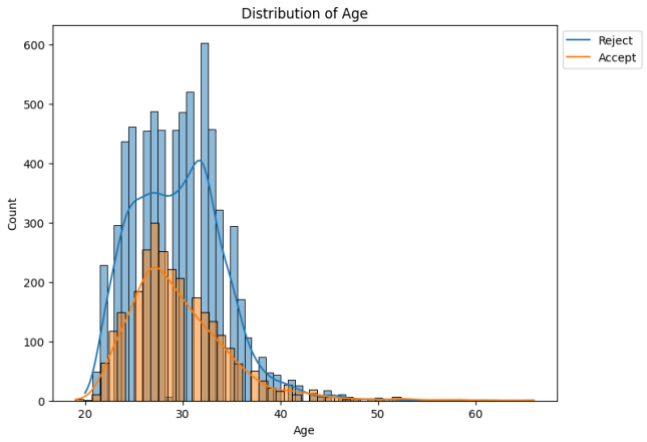


Figure 5. Distribution of Age across 'Accept' and 'Reject' observations

While several previous studies have focused on machine learning bias for undergraduate data [12, 15, 16, 19], few have conducted similar studies on graduate data [11, 13]. One of the peculiarities of experimenting with graduate application data is that it includes a significant number of applications from outside the United States. This can result in findings that are vastly different from undergraduate data due to the majority of undergraduate applications typically being from within the United States with similar prior education. However, in the case of graduate applications, there has been an increasing trend in the number of international applicants. This is evident in Figure 6 which shows the 10 countries with the highest number of applications, of which the United States comes third with 883 applications over 10 years. This aligns with reports by the NCES [22] that 44% of STEM master's degrees conferred in 2019-20 were by international students.

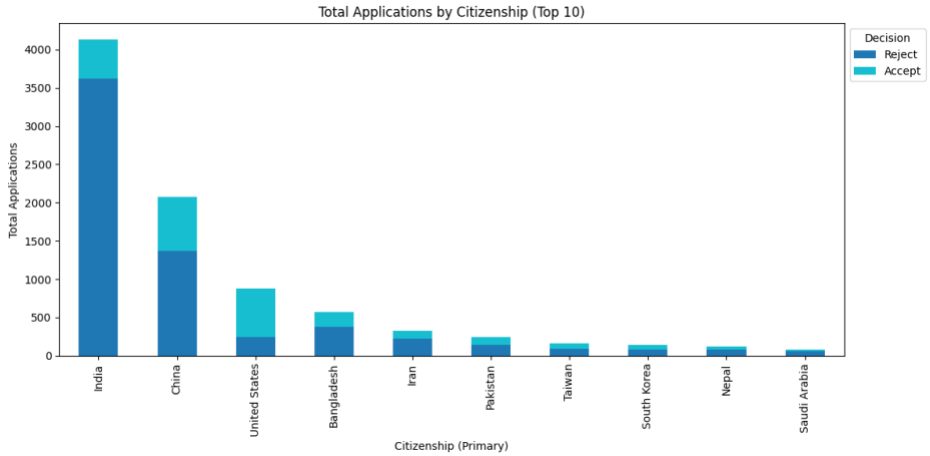


Figure 6. Top 10 Applications by Citizenship

By examining the distribution of reported race, we found an increasing trend of applicants opting out of reporting race[2], aligning with findings from [23]. The majority of U.S. applicants reported their race as Caucasian (63%), followed by Asian (25%). Among the URMs, around 6% of applicants reported race as Black and 5% of applicants reported race as Hispanic within the US. The American Indian and Pacific Islander groups have very low percentages of 1% and 0.3% respectively. The majority (86%) of the international applicants reported race as Asian, which aligns with findings from Figure 6, where the top countries include India and China.

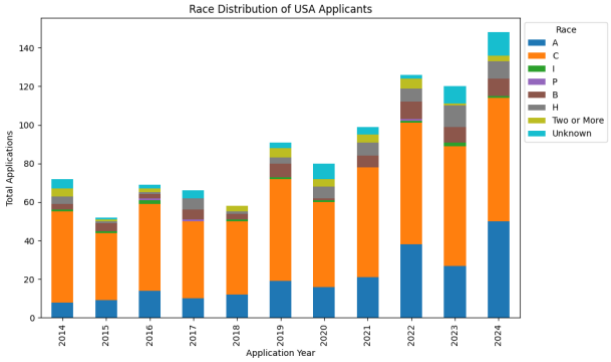
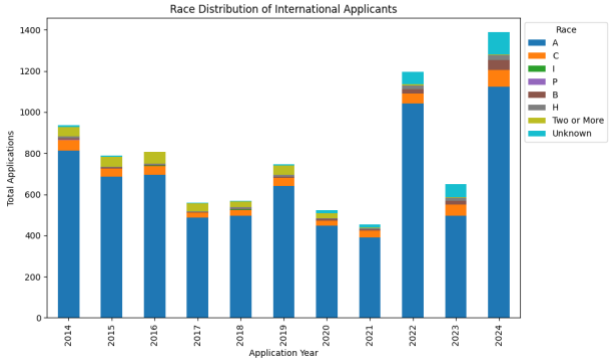


Figure 7a. Race of USA applicants

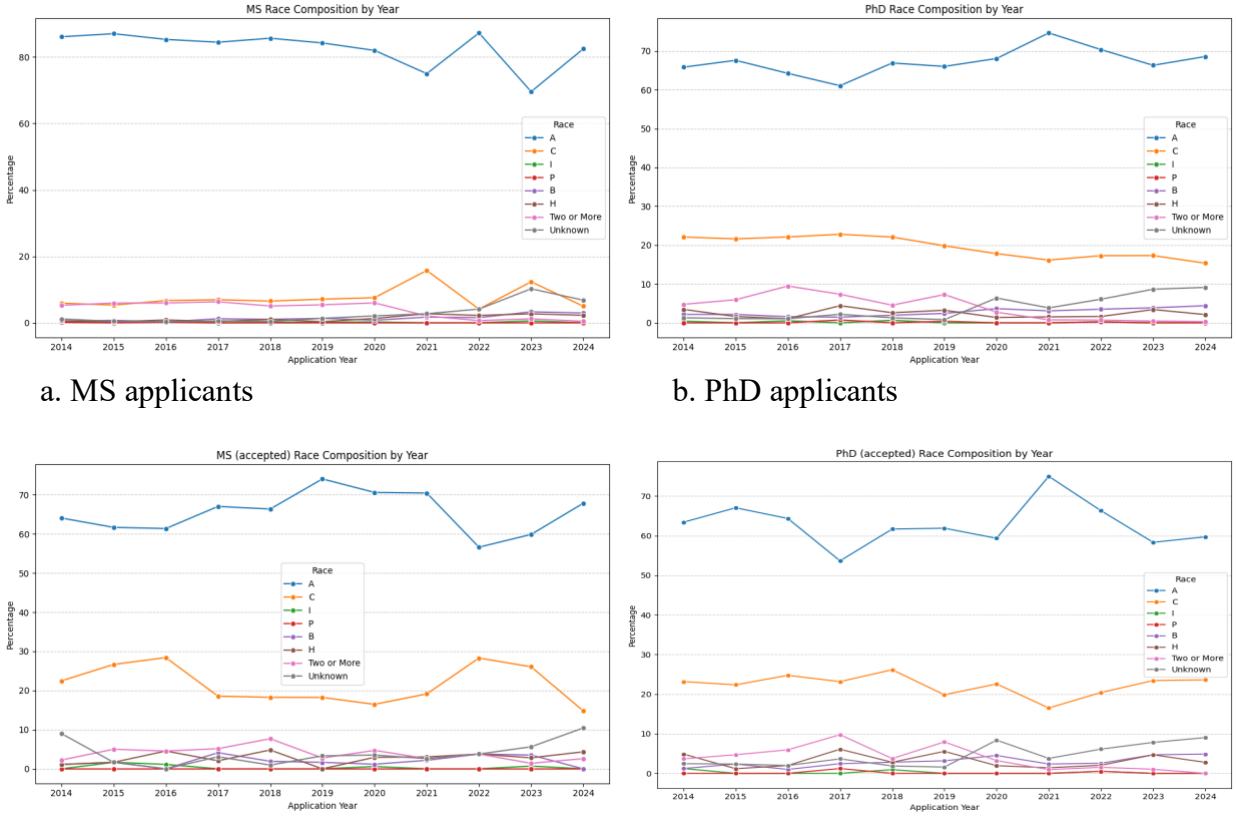


7b. Race of international applicants

The number of applicants opting out of specifying race or marking it 'Unknown' has been increasing in recent years. While it is too early to draw conclusions about the effect of the 2023 Supreme Court ruling for affirmative action ban [9], we observe that the number of applicants opting out of reporting race has risen by 66% from 2023 to 2024. Several universities across the U.S. have reported a similar trend [23], indicating increased concern around disclosing race among applicants after the ban of affirmative action.

[2] Race notations: A – Asian, B – Black, C – Caucasian, H – Hispanic, I – American Indian, P – Pacific Islander or Alaskan Native

We also examined the race composition of the two graduate computer science degrees offered by the university. Among the underrepresented minority groups (URMs), we observe a decline in the percentage of applicants who reported race as Black by 12.8% and Hispanic by 17.9% for the MS degree. However, for the PhD degree, the increasing trend of students reporting race as Black continues, with the percentage increase changing from 11.3% to 13% in the 2022-23 and 2023-24 application cycles, whereas there is a decline in Hispanic. When comparing U.S. and international applicants with reported race Black, there is a positive change for U.S. applicants from 5.4% to 5.6% whereas international applicants declined from 3.58% to 2.94%. The percentage of PhD applicants reporting race as Hispanic sharply declined after the affirmative action ban from 3.6% to 2.2%, a decline of 38%. However, the admitted class has contrasting results with a sharp decline of Black students for the MS program but a steady increase for the PhD program. Hispanic students have contrasting results in the admitted class, with an increase in the MS program and a slight decrease in the PhD program. These findings are evident in Figure 8.



a. MS applicants

b. PhD applicants

c. MS admitted class

d. PhD admitted class

Figure 8 Race Distribution of MS and PhD students (applicants and admitted class)

Finally, we also examined the gender distribution of applicants for the 2014-2024 period. We found that the gender ratio of applicants has remained nearly constant through the years, with the percentage of females (F) averaging 26.6% as shown in Figure 9. There are minimal observations (0.3%) that have marked their gender as neither (N).

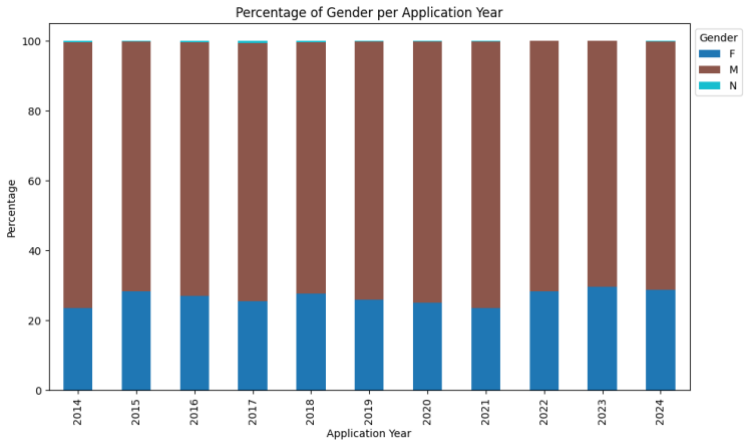


Figure 9. Gender Distribution of Applicants

4.2.2 Clustering

A common unsupervised learning approach for identifying subgroups in a dataset is clustering. In this study, we used clustering to identify underlying patterns among different subgroups that may lead to a positive or negative decision on an application. Since we are experimenting with admissions data, the two classes of data would be those with the decision 'admit' and those with the decision 'reject'. However, clustering into these two classes would not render much insight into the contributions of different features towards the admission decision. Therefore, we assumed an unknown number of clusters in the data and performed Hierarchical Density-based Spatial Clustering (HDBScan) [24] to segregate the observations into subsets for each cluster.

We chose the density-based clustering approach as it is robust to noisy data since it detects dense clusters and categorizes the remaining observations as noise [25]. It can also tackle high-dimensional data, unlike other clustering algorithms like KMeans, especially when there is no predefined number of clusters. The initial clustering algorithm yielded 13 distinct clusters with 71% coverage of the dataset, leaving only 29% of the data points as noise. However, this achieved a low Density Based Clustering Validation (DBCV) score of 0.15, indicating that the clustering was not optimal. After fine-tuning the parameters of the HDBScan algorithm, we achieved an improved DBCV score of 0.33 with 7 dense clusters shown in Figure 10, although this resulted in a reduced coverage of 56%.

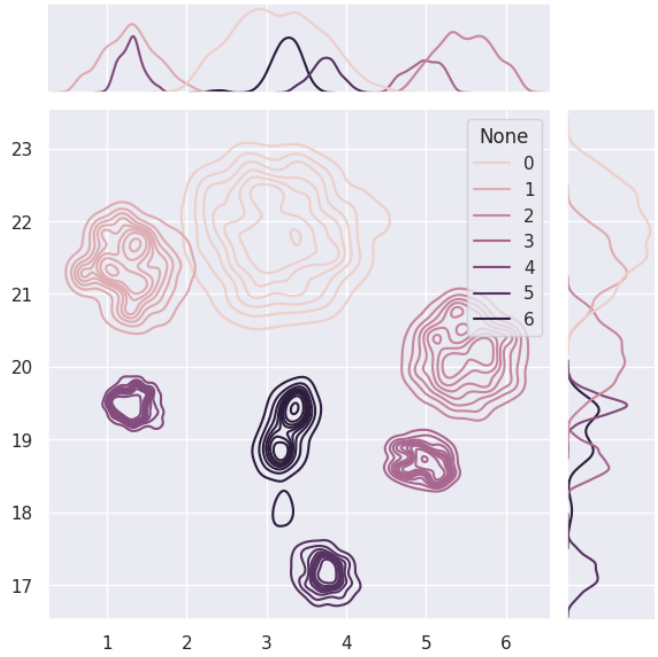


Figure 10. HDBScan clustering results

Among the seven clusters, we noticed three potential types of clusters: a. homogenous clusters with a majority of rejected applications, b. heterogeneous clusters with a mixture of accepted and rejected applications, and c. homogenous clusters with a majority of accepted applications. For the type a cluster, we noticed that 'Birth Nation' India has a negative correlation with the decision, while 'Birth Nation' South Korea and Pakistan have a positive correlation with it. 'GPA' also has a positive correlation with the decision. This could imply that being from India might lead to a negative decision but may also reflect the applicants in this cluster since the cluster contains 96% applications with 'Birth Nation' India, 4 applications with 75% acceptance from Pakistan, and 2 applications with 100% acceptance from South Korea. Type b cluster had an overall acceptance rate of 36%, with the decision having a negative correlation to 'Age' and a positive correlation to 'GPA'. Type c clusters had an acceptance rate of 71%, with positive correlations for 'GPA' and 'Birth Nation' United States. It had negative correlations for 'TOEFL' with 'Birth Nation' being Bangladesh or Jordan and 'Race' being 'Two or More'. Since all the correlations mentioned are in the absolute range of [0.08, 0.2], they are only mildly correlated with the decision. It is also important to note that the correlations depend on the demographic features of the subset of data within each specific dense cluster. Nonetheless, when a machine learning model trains on such data, it may consider these apparent correlations as rubrics for decision-making, which can have severe negative impacts on class diversity.

4.3 Detection of Intersectional Feature Bias

Many have studied the bias-contributions of individual features in the dataset in the context of university admissions [11, 12, 13, 15, 16, 19]. While these are useful in determining which features may lead to a biased outcome, they do not consider all possible combinations of features that may lead to a bias, also known as intersectional feature bias. As argued by Wamburu et al. in

[26], systematic scanning [27] without presupposing bias-inducing features may reveal subsets of features that contribute to bias that are difficult to discover otherwise. One way to tackle intersectional feature bias is through subgroup discovery, which is a statistical approach to extract subgroups that have an increased likelihood of achieving a particular target outcome [18]. We applied the subgroup discovery algorithm using the Pysubgroup library [28] to our dataset with the application decision set as the target variable. We limited our search space to only include features defined as demographic in Table 1. However, we had to exclude the features 'Current or Former Military' and 'First Generation', since these did not have a sufficient number of filled values and distorted the results. We extracted the top ten subsets with the highest quality and listed those with quality above the threshold of 0.01 in Table 2 and Table 3.

The results indicate a weak bias towards the target variable since the quality values are low. However, this may be limited by the distribution of data in the dataset we used, as well as its small size. The quality metric represents the deviation of observed outcomes from expected outcomes for each subset, hence indicating bias. For the positive outcome subsets, i.e., with observations that had the decision 'Accepted', we observe that the common intersections of features are with 'Birth Nation' being United States or China, 'Gender' being male, and 'Race' being 'C' or 'A'. The subgroups with the highest coverage of more than 20% are Birth Nation=='United States', Birth Nation=='China', and (Birth Nation=='China' AND Race=='A'). Subset coverage indicates how much of the overall dataset constitutes that particular subset. The lift metric represents the ratio of the target class in the subset as against its prevalence in the entire dataset. The results show that an applicant in subgroups 1 and 2 is nearly 2.5 times more likely to receive an admit, an applicant in subgroup 6 is 2.6 times more likely to receive an admit and an applicant in subgroup 4 is nearly 2.7 times more likely to receive an admit.

Table 2. Subgroup discovery results for Decision 'Accepted' (quality >=0.01)

Index	Subset	Quality	Subgroup Coverage	Lift
1	Birth Nation=='United States'	0.04	0.22	2.46
2	Birth Nation=='United States' AND Gender=='M'	0.03	0.18	2.47
3	Race=='C'	0.03	0.19	2.08
4	Birth Nation=='United States' AND Race=='C'	0.02	0.13	2.67
5	Gender=='M' AND Race=='C'	0.02	0.15	2.14
6	Birth Nation=='United States' AND Gender=='M' AND Race=='C'	0.02	0.11	2.62
7	Birth Nation=='China'	0.01	0.26	1.17
8	Birth Nation=='China' AND Race=='A'	0.01	0.24	1.17

We can extrapolate similar findings for the negative outcome subsets, i.e., when the application decision is 'Rejected'. The features identified through subset scanning include 'Birth Nation' being India, 'Race' being Asian, and 'Gender' being male or female. Table 3 is sorted by quality and displays the top subgroups discovered with quality greater than or equal to 0.01. The lift value is highest for subgroups 3 and 4 with an increased likelihood of rejection of 1.25 times for features including birth nation, gender, and race.

Table 3. Subgroup discovery results for Decision 'Rejected' (quality >=0.01)

Index	Subset	Quality	Subgroup Coverage	Lift
1	Birth Nation=='India'	0.07	0.55	1.23
2	Birth Nation=='India' AND Race=='A'	0.06	0.46	1.23
3	Birth Nation=='India' AND Gender=='M'	0.06	0.39	1.25
4	Birth Nation=='India' AND Gender=='M' AND Race=='A'	0.05	0.32	1.25
5	Race=='A'	0.03	0.77	1.06
6	Gender=='M' AND Race=='A'	0.02	0.56	1.07
7	Birth Nation=='India' AND Gender=='F'	0.02	0.15	1.20
8	Birth Nation=='India' AND Gender=='F' AND Race=='A'	0.02	0.13	1.20

Overall, we observe low bias for both the target outcomes but identify features that could lead to a biased outcome. We also note that the increased likelihood is higher for the positive target class (accept) than the negative class (reject).

4.4 Random Forest Feature Importance

A straightforward explainability method to understand the perception of training features by machine learning models is the visualization of the feature importance. We trained a Random Forest classifier on our training set which contained 84% of the original dataset filtered by application year from 2014-2023. The resulting feature importances with a significance greater than 0.01 are displayed in Figure 11. While we observe the GRE, TOEFL, and GPA scores among the top features with high importance, demographic features like age, citizenship, birth nation, and race have also been identified by the model as highly important. First-generation and military experience features, though present in the feature importance graph, may also be there due to the majority of their values being imputed missing values with –1. The model may have presumed that these features having value –1 correlate to a decision class. The major cause for concern from the feature importance graph is that features including age, the applicant's citizenship and birth nation being India or the United States, and race being Asian are all inferred as important features by the model. These may simply be the demographic features of the 'accept' and 'reject' subgroups after the admission decision was made and are unlikely to be decision-making factors that were considered by the admission review committee. However, they are inferred as the most important independent features by the random forest model, which illustrates that the model may be learning some biases.

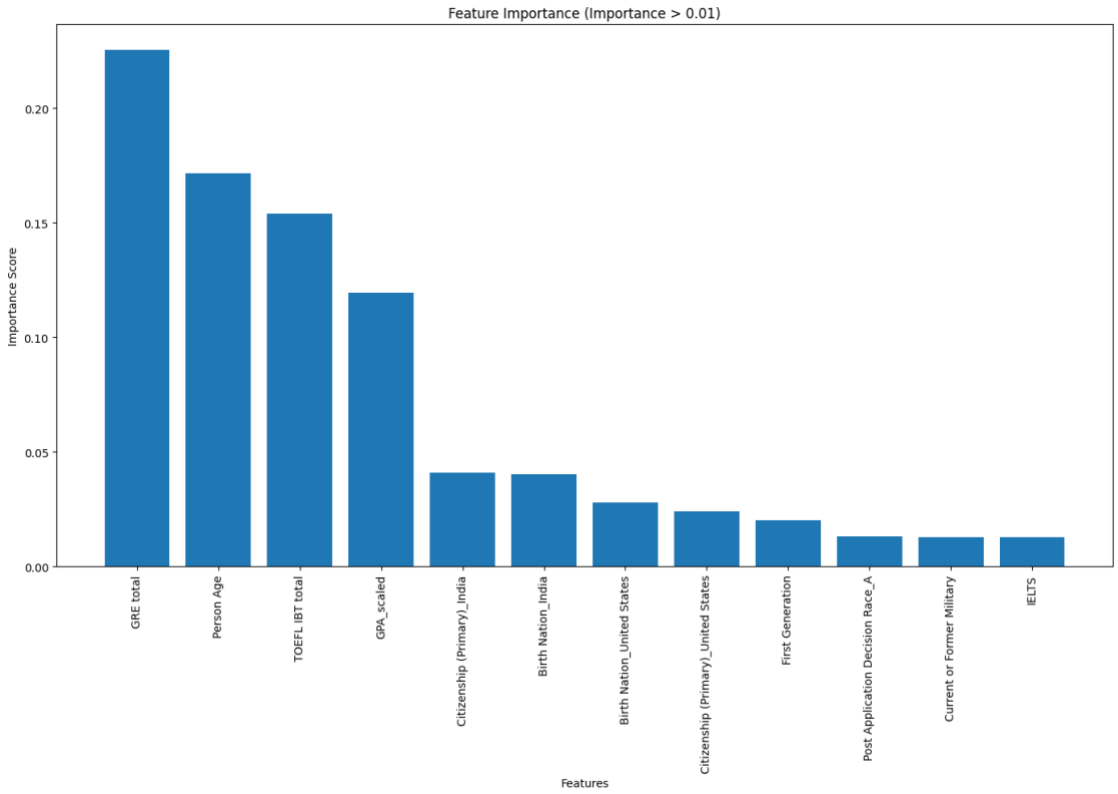


Figure 11. Feature Importance using Random Forest Classifier

5. Discussion

5.1 Impact of the Affirmative Action Ban

Affirmative action was introduced in the early 2000s in the United States as a national initiative to address historical injustices in the lives of women and ethnic minorities by guaranteeing them some advantage in college admissions and employment possibilities. Though its historical significance was to provide more opportunities for underrepresented communities to excel in academia and industry, it did not imply that universities would explicitly use race as a criterion for scoring applications but rather use it as an additional characteristic after consolidating already highly qualified applicants, within constitutional limits [29]. In 2023, the Supreme Court ruled against affirmative action practices and declared that universities may not conduct race-conscious admission reviews [9]. Preliminary effects of this decision have already been analyzed for various universities based on their published 2024 admissions data [24, 30, 31]. A study published in 2024 analyzed the enrolment changes for the undergraduate class of 2028 across different U.S. universities that previously practiced race-conscious admissions [32]. They provide fair warning that these trends may be the typical changes that occur year-on-year with the admissions cycles and therefore it is not conclusive if they are a result of the affirmative action ban. Nonetheless, inferences can be made about the change in the class population of URMs. A vast majority of the universities in the study do reflect a decline in the number of Black and Hispanic students, with a few universities having contrasting outcomes.

Our findings from 4.2.1 indicate similar changes in our graduate admissions dataset, where Black and Hispanic students' applications have reduced in the 2024 admissions cycle for the MS program compared to 2023. Since the applications from the American Indian and Alaskan Native populations are already low, we could not draw significant conclusions for these groups in the 2024 cycle. A positive observation is that the rate of change year on year has increased from 2023 to 2024 by 15% in the case of Black PhD applicants. From Figure 7, we observe a significant increase in the number of applicants opting out of reporting race in 2024, categorized as 'unknown'. Though this had a pre-existing increasing trend for applications from within the US, we observe a sudden increase among international applicants. This aligns with reports [23] that an increasing number of students are choosing to not disclose their race after the affirmative action ban, especially in highly competitive universities. In conclusion, with respect to RQ 1, we observe trends in our dataset that concur with observations from across competitive universities in the US that there is a common hesitation to report race and a decline in URM applicants. However, it remains premature to make strong claims with a single year of data post the affirmative action ban and will be more significant after a few years of observation.

5.2 Potential for Inferred Bias

Through various experiments detailed in section 4, we explored RQ 2 and demonstrated that several features could contribute to the inferred bias of a machine learning model. The feature importance ranking with the Random Forest classifier in section 4.4 rendered demographic features like age, birth nation, and race as highly important features. This indicates that the model's generalized algorithm for deciding if a particular application should be granted admission factors in these features with a high weightage. Though a deep learning model or a neural network may improve performance, the Random Forest classifier offers interpretability, which is paramount when modeling human-centric processes. The density-based clustering in section 4.2.2 reveals results that coincide with our findings from the feature importance ranking, such as the birth nation India having a higher tendency to be rejected and the birth nation United States having a higher tendency to be accepted. Another finding from the clustering was that 'Age' had a negative correlation to the decision for some clusters, which aligns with our finding from the feature importance ranking that age had an importance of around 0.17. This implies that age is another feature that may be used for decision-making by the machine learning model.

We explored RQ 3 through our analysis in section 4.3 using subgroup discovery. We found various permutations of birth nation being India, gender being male or female, and race being Asian displaying bias towards an application being rejected. This augments our findings from sections 4.4 and 4.2.2, where experiments revealed the same demographic features as having a higher likelihood of being rejected. Likewise, the subgroup discovery yielded permutations of the birth nation being United States or China, gender being male and race being Caucasian or Asian to have a higher likelihood of being accepted. In a discussion detailed in [33], the author mentions that members of the admission review committee are highly skilled at triangulation - a process in which they can easily determine if the characteristics of an applicant remain consistent throughout different application materials, and are therefore able to tactfully make decisions on the student's potential success in university. This means that while reviewers might be looking for certain specific qualities, they do not process applications by strictly adhering to an algorithm, which allows for flexibility of human judgment based on context. However, as seen in

the case of GRADE [3], machine learning models do not have a similar ability, and might harm diversity by measuring the success of students based directly on historical admissions trends of accepted students. Therefore, it is critical to be aware of the potential biases that a model may infer from the dataset and rectify these before the model is applied in practice.

5.3 Bias or a Reflection of the Applicant Population

The experiments in section 4 demonstrated different approaches to examine the dataset for biases that may be inferred by a machine learning model. However, we must also note that many of these biases stem from distributions in the dataset of historical admits and rejects of applications. This raises the question of whether these are truly biased outcomes or simply a reflection of the data itself. For example, the negative class, i.e., the decision being 'rejected', was often biased towards applications having birth nation India. Though the data indicates a bias, this does not necessarily imply that the admission reviewers will reject applications on the basis of the birth nation being India. Figure 6 illustrates that the dataset contains around 4000 observations falling in the category of birth nation India, which accounts for nearly 43% of the dataset. Since the dataset itself is not evenly distributed and certain demographic qualities exist in larger proportions in the dataset, our findings conclude these as biases. Nonetheless, Figure 6 also shows us that the acceptance rate is higher for applications from the United States and China, despite the total number of applications being lower. This may have led our analysis to find a bias for applications from these two countries towards the 'accepted' class.

We must also note that the dataset is from a university in the United States, which may be a reason that there is a higher acceptance rate for applicants from the United States, since there may be students who were previously enrolled in the same university or other well-recognized universities in relevant programs. If the results of our bias detection experiments are a reflection of the data, then one might ask why a machine learning model learning the same patterns might be problematic. This is because human reviewers make multiple considerations based on the unique attributes of each application and do not adhere to a specific formula for making decisions despite using a rubric of minimum requirements that indicate the applicant's potential to succeed in the program. On the contrary, a machine learning algorithm learns patterns in the data as rules, which determine how it will classify observations in the future. This will then lead to algorithmic bias and potentially harm the diversity of the graduate student population. Therefore, it is imperative to examine the data for potential biases that a model might infer and minimize the risk of reduced diversity caused by a potentially biased model. Once a model is developed, universities could also apply explainability methods to validate if these inferred biases have been learned by the machine learning model during training and seek methods to rectify them through bias correction measures.

5.4 Opportunities to Increase Efficiency and Improve Diversity

With the affirmative action ban in place and based on the findings in this study regarding the use of a machine learning model to efficiently conduct admission reviews, we identified two main issues concerning diversity for universities to consider: a. demographic composition of the

applicant pool and b. demographic composition of the admitted population. We propose that universities tackle both these issues in order to ensure that diversity is facilitated in their programs. The first issue may be tackled by conducting an increased number of inclusivity initiatives that encourage applicants from URMs to apply, along with bridge courses and workshops that may provide support for students with a poor academic history or lack of access to resources for standardized tests to boost their profile. Additional financial support initiatives may also encourage students of these groups to apply. When applying machine learning models for admissions processing, the second issue can be tackled by thoroughly investigating inferred bias from data used as input to the machine learning models by using a framework similar to ours shown in Figure 3. They may also conduct post-modeling analysis and implement bias correction methods to ensure that machine learning models do not just repeat history and instead give fair consideration to all applicants, especially those from URMs. It is also critical to develop these models with incremental improvements with human evaluation or develop a human-in-the-loop system such that the model decisions are validated by a reviewer.

6. Limitations and Future Work

Bias in the admissions process is a challenging problem to tackle, as it can emerge in various stages as shown in section 2.1. Studying bias in the context of graduate admissions data, for graduate computer science programs in our case, brings added complexity. Previous studies have often focused on bias in machine learning for admission in undergraduate data [12, 15, 16, 19], which tends to be less complex than graduate data due to the uniformity in application details since they are often derived through the Common App [34]. Since most applicants apply from within the United States, they have a similar educational background, without different GPA scales or standardized testing requirements based on country of origin. This leads to more consistent and less sparse data when considering undergraduate applications. Graduate applications, however, may require additional essays and standardized tests that are required for a specific subset of applicants or are optional.

Feature distributions and admission review criteria also differ vastly between undergraduate and graduate data. For example, age is usually consistent across undergraduate applications since most prospective students apply immediately after the completion of high school, but the same feature has a much higher variance in graduate data since applicants may pursue graduate education directly after their undergraduate education or at any stage of their career. Graduate admissions data also has a significantly lower volume of data per admissions cycle, owing to its significantly lower intake compared to undergraduate programs. In addition to this, the process of admission review varies not only between different universities but also between the undergraduate and graduate programs in the same university. Undergraduate applications are typically reviewed centrally by the university whereas graduate admission review may be conducted by a specific department's professors and staff since essays can be specific to the field. Therefore, it is difficult to generalize decision-making criteria as application reviews may be unique to each department.

Admissions data is not easily available online and is highly confidential. Most of the prior works have independently sourced their datasets from different universities, and limited such studies exist due to the challenge of data that is difficult to access. We conducted our analyses on

graduate Computer Science applications to two programs of a single university. While our methodology yielded interesting results, these cannot be generalized to all datasets across different universities, due to the different criteria, policies and nuances of admission reviews in different universities. Nonetheless, our work contributes a framework for the detection of inferred bias and it will be interesting to explore in the future of different datasets reveal similar findings.

This study discusses the potential bias that may be inferred from the data by a machine learning model during training but was limited in that it only analyzed non-essay data. Future work could include the various essay texts listed in section 2.1 in the analyses as input data for the machine learning model, much like the human review process would. This might reveal further potential biases that may be hidden in text data, as demonstrated in [12]. The study could also be extended by training various machine learning models and analyzing their outcomes to validate if the determined biases are indeed inferred and transferred into the model as algorithmic bias. This can be done by applying explainability methods like LIME [20] for local explanations and SHAP [35] for global explanations. We can also use fairness assessment suites like IBM's AIF360 [36] and Aequitas [37], which provide a range of metrics that help assess fairness and bias in the model outcomes.

7. Conclusion

The increasing number of graduate CS applications necessitates AI-based admissions processing to increase efficiency and allow universities to allocate the limited resources available in admissions committees towards other critical tasks like financial aid decisions. While such solutions may benefit universities immensely, they also present a risk to diversity, due to models training on historical data and inferring bias. This issue is more critical than ever due to the recent policy changes surrounding affirmative action practices in university admissions. To elucidate the potential risk of using machine learning for admission decisions, we put forth three main research questions regarding the effect of the affirmative action ban, the independent features that could induce bias in the machine learning model, and the potential intersectional bias of subgroups in the data. We proposed a framework to detect inferred biases from data systematically, using exploratory data analysis, clustering, subgroup discovery, and feature importance. The trends post the affirmative action ban showed a decline in applications from Black and Hispanic students for the MS program, which aligned with trends across universities in the U.S. Contrarily, we found an increase in Black applicants for the PhD program, but this did not reflect in the accepted class of students. We also found a significant increase in the number of applicants opting out of reporting race by 66% from 2023 to 2024, showing a widespread diversity concern among applicants post the affirmative action ban. Nonetheless, with only one year of data, it may be too soon to conclude that this is due to the affirmative action ban. While examining the sources of inferred bias, we identified that age, birth nation, and race contributed independently to inferred bias. Additionally, we found that birth nation, race, and gender contributed intersectionally to inferred bias. Future work should focus on addressing the limitations of this study by incorporating textual data from applications into the bias detection framework, implemented the framework using different datasets, and applying machine learning models to study their outputs with explainability models. In conclusion, this study highlights the need for bias detection in AI-based admissions reviews and provides a framework

to discover inferred biases before applying machine learning models to automate graduate admissions.

8. References

[1] B. Reyell, "Unpacking the Earning Potential of a Graduate Degree," Northeastern University Graduate Programs, Sep. 19, 2024. https://graduate.northeastern.edu/knowledge-hub/earning-potential/ (accessed Jan. 15, 2025).

[2] "Digest of Education Statistics," *Ed.gov*, 2021. https://nces.ed.gov/programs/digest/d23/tables/dt23_325.35.asp

[3] A. Waters and R. Miikkulainen, "GRADE: Machine-Learning Support for Graduate Admissions," *AI Magazine*, vol. 35, no. 1, pp. 64–75, 2014, doi: 10.1609/aimag.v35i1.2504.

[4] "8 in 10 Colleges Will Use AI in Admissions by 2024," Intelligent, Sep. 27, 2023. https://www.intelligent.com/8-in-10-colleges-will-use-ai-in-admissions-by-2024/

[5] C. Claybourn, "Is AI Affecting College Admissions?," *US News & World Report*, 2023. https://www.usnews.com/education/best-colleges/articles/is-ai-affecting-college-admissions

[6] J. Cuny and W. Aspray, "Recruitment and retention of women graduate students in computer science and engineering: results of a workshop organized by the computing research association," SIGCSE Bull., vol. 34, no. 2, pp. 168–174, Jun. 2002, doi: 10.1145/543812.543852.

[7] J. Daniel, "Diversifying Graduate Student Enrollment: What We Know and What We're Learning | S&T Policy FellowsCentral," Aaaspolicyfellowships.org, Mar. 10, 2023. https://www.aaaspolicyfellowships.org/blog/diversifying-graduate-student-enrollment-what-we-know-and-what-were-learning#_edn3 (accessed Jan. 15, 2025).

[8] E. Grieco, "Diversity and STEM: Women, Minorities, and Persons with Disabilities 2023 | NSF - National Science Foundation," Nsf.gov, 2023. https://ncses.nsf.gov/pubs/nsf23315/report/graduate-enrollment-in-science-and-engineering#overall-minority-enrollment (accessed Jan. 15, 2025).

[9] S. Wood, "How Does Affirmative Action Affect College Admissions?," U.S. News & World Report, Nov. 03, 2022. Available: https://www.usnews.com/education/best-colleges/applying/articles/how-does-affirmative-action-affect-college-admissions

[10] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, "Towards a standard for identifying and managing bias in artificial intelligence," National Institute of Standards and Technology (U.S.), Gaithersburg, MD, NIST SP 1270, Mar. 2022. doi: 10.6028/NIST.SP.1270.

[11] Y. Zhao, Z. Qi, J. Grossi, and G. M. Weiss, “Gender and Culture Bias in Letters of Recommendation for Computer Science and Data Science Masters Programs,” Sci Rep, vol. 13, no. 1, p. 14367, Sep. 2023, doi: 10.1038/s41598-023-41564-w.

[12] A. J. Alvero et al., “AI and Holistic Review: Informing Human Reading in College Admissions,” in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York NY USA: ACM, Feb. 2020, pp. 200–206. doi: 10.1145/3375627.3375871.

[13] G. Kalhor, T. Zeraati, and B. Bahrak, “Diversity dilemmas: uncovering gender and nationality biases in graduate admissions across top North American computer science programs,” EPJ Data Sci., vol. 12, no. 1, p. 44, Oct. 2023, doi: 10.1140/epjds/s13688-023-00422-5.

[14] B. M. Neda and S. Gago-Masague, “Feasibility of Machine Learning Support for Holistic Review of Undergraduate Applications,” in 2022 International Conference on Applied Artificial Intelligence (ICAPAI), Halden, Norway: IEEE, May 2022, pp. 1–6. doi: 10.1109/ICAPAI55158.2022.9801571.

[15] A. Priyadarshini, B. Martinez-Neda, and S. Gago-Masague, “Admission Prediction in Undergraduate Applications: an Interpretable Deep Learning Approach,” in 2023 Fifth International Conference on Transdisciplinary AI (TransAI), Sep. 2023, pp. 135–140. doi: 10.1109/TransAI60598.2023.00040.

[16] K. Van Busum and S. Fang, “Analysis of AI Models for Student Admissions: A Case Study,” in Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, Tallinn Estonia: ACM, Mar. 2023, pp. 17–22. doi: 10.1145/3555776.3577743.

[17] A. Wang, V. V. Ramaswamy, and O. Russakovsky, “Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation,” in 2022 ACM Conference on Fairness, Accountability, and Transparency, Jun. 2022, pp. 336–349. doi: 10.1145/3531146.3533101.

[18] S. Helal, “Subgroup discovery algorithms: A survey and empirical evaluation,” Journal of Computer Science and Technology, vol. 31, pp. 561–576, 2016, Available: https://api.semanticscholar.org/CorpusID:255153310

[19] T. Lux, R. Pittman, M. Shende, and A. Shende, “Applications of Supervised Learning Techniques on Undergraduate Admissions Data,” in Proceedings of the ACM International Conference on Computing Frontiers, Como Italy: ACM, May 2016, pp. 412–417. doi: 10.1145/2903150.2911717.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the predictions of any classifier,” San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. doi: https://doi.org/10.1145/2939672.2939778.

[21] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A Survey on Bias and Fairness in Machine Learning,” ACM Comput. Surv., vol. 54, no. 6, p. 115:1-115:35, Jul. 2021, doi: 10.1145/3457607.

[22] “COE - Graduate Degree Fields,” nces.ed.gov. https://nces.ed.gov/programs/coe/indicator/ctb

[23] Z. Schermele, “At selective colleges, fewer students are disclosing race in their applications,” USA TODAY, Oct. 21, 2024. https://www.usatoday.com/story/news/education/2024/10/21/affirmative-action-ban-admissions-effect-2024/75699203007/

[24] Ricardo, D. Moulavi, and J. Sander, “Density-based clustering based on hierarchical density estimates,” J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Eds., Springer Berlin Heidelberg, 2013, pp. 160–172.

[25] “Comparing Python Clustering Algorithms — hdbscan 0.8.1 documentation,” Readthedocs.io, 2016. https://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html

[26] J. Wamburu, G. A. Tadesse, C. Cintas, A. Oshingbesan, T. Akumu, and S. Speakman, “Systematic Discovery of Bias in Data,” in 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan: IEEE, Dec. 2022, pp. 4719–4725. doi: 10.1109/BigData55660.2022.10020781.

[27] D. B. Neill, M. III, and H. Zheng, “Fast subset scan for multivariate event detection,” Statistics in Medicine, vol. 32, Art. no. 13, 2013, doi: https://doi.org/10.1002/sim.5675.

[28] Pysubgroup: Lemmerich, F., & Becker, M. (2018, September). pysubgroup: Easy-to-use subgroup discovery in python. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECMLPKDD). pp. 658-662.

[29] D. Mitchell and E. A. Daniele, “Diversity in American Graduate Education Admissions: Twenty-first-century Challenges and Opportunities,” DigitalCommons@Molloy, 2015. https://digitalcommons.molloy.edu/eas_pub/31/ (accessed Jan. 15, 2025).

[30] P. Waldron, “Race-blind college admissions harm diversity without improving quality | Cornell Chronicle,” Cornell Chronicle, 2024. https://news.cornell.edu/stories/2024/11/race-blind-college-admissions-harm-diversity-without-improving-quality

[31] J. Murphy, “What Happened to Campus Diversity Post-SFFA? Five Findings - Education Reform Now,” Education Reform Now, Oct. 17, 2024. https://edreformnow.org/2024/10/17/what-happened-to-campus-diversity-post-sffa-five-findings/ (accessed Jan. 15, 2025).

[32] J. Murphy, “Tracking the Impact of the SFFA Decision on College Admissions - Education Reform Now,” Education Reform Now, Sep. 09, 2024.

https://edreformnow.org/2024/09/09/tracking-the-impact-of-the-sffa-decision-on-college-admissions/

[33] E. Evaristo, “Balancing the potentials and pitfalls of AI in college admissions | USC Rossier School of Education,” rossier.usc.edu, Dec. 04, 2023. https://rossier.usc.edu/news-insights/news/balancing-potentials-and-pitfalls-ai-college-admissions

[34] The Common Application, Aug. 14, 2018. https://www.commonapp.org/

[35] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” Nov. 24, 2017, arXiv: arXiv:1705.07874. Accessed: May 17, 2024. [Online]. Available: http://arxiv.org/abs/1705.07874

[36] R. K. E. Bellamy et al., “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias,” Oct. 03, 2018, arXiv: arXiv:1810.01943. doi: 10.48550/arXiv.1810.01943.

[37] P. Saleiro et al., “Aequitas: A Bias and Fairness Audit Toolkit,” Apr. 29, 2019, arXiv: arXiv:1811.05577. Accessed: May 22, 2024. [Online]. Available: http://arxiv.org/abs/1811.05577