JOURNAL OF
STEM EDUCATION:
Innovations and Research
15th Anniversary

Developing a Hybrid Model to Predict Student First Year Retention in STEM Disciplines Using Machine Learning Techniques

Ruba Alkhasawneh
Intel Corporation

Rosalyn Hobson Hargraves
Virginia Commonwealth University

1. Introduction

Understanding the reasons behind the low enrollment and retention rates of Underrepresented Minority (URM) students (African Americans, Hispanic Americans, and Native Americans) in the disciplines of science, technology, engineering, and mathematics (STEM) has concerned researchers for decades. Statistics show that students of color have higher attrition rates compared with other groups, although this trend has been decreasing over the past twenty years (Besterfield-Sacre, Atman, & Shuman, 1997; Mitchell & Daniel, 2007; Fleming, Ledbetter, Williams, & McCain, 2008). These groups tend to enroll in STEM majors in small numbers and leave in higher numbers (Urban, Reyes, & Anderson-Rowland, 2002; Alkasawneh & Hobson, 2009).

Increasing the number of minorities (women and ethnic groups) is a practical way of increasing the workforce pool in STEM fields where white male representation is still dominant. Unfortunately, this solution is difficult for many institutions. Only two out of five African American and/or Hispanic American students remain in their majors and receive bachelor's degrees in a STEM discipline nationwide (Markley, 2005). In order to impact workforce demographics, the population of students choosing STEM majors must change. The literature reflects a substantial interest in increasing URM student retention in higher education (Sidle & McReynolds, 1999; Nave, Frizell, Obiomon, Cui, & Perkins, 2006; Hargrove & Burge, 2002). Retention is of significant interest because of its positive impact on college reputation and workforce demographics (Williford & Schaller, 2005).

Several studies emphasize the importance of identifying college students with higher risk of dropping out in early stages in order to allocate the available resources based upon student needs (Herzog, 2006; Lin, Imbrie, & Reid, 2009). Research by Zhang, Anderson, Ohland, Carter, & Thorndyke (2002) stated that identifying factors that affect student retention could play an effective role in the counseling and advising process for engineering students. This equips institutions to utilize their available resources based upon those groups' needs (Herzog, 2006). Traditional methods of statistical analysis have been used to predict student retention, such as logistic regression (Gaskins, 2009). Recently, research has focused on data mining techniques to study student retention in higher education (Brown, 2007). These techniques are highly accurate, robust with missing data, and do not need to be built on a hypothesis. Data mining is defined as recognizing patterns in a large set of data and then trying to understand those patterns.

1.1 Predictive models of student retention

1.1.1 Tinto's model

Tinto in his model (1975) noted that integration into the college system, academically and socially, impacts students' decision regarding dropping out of college. He added that integration into the college system causes a continuous change in student goals and commitment to graduation, which in turn might generate the decision of persistence or dropping out of college. Tinto's model was based on Durkheim's theory of suicide (Durkheim, 1951) which clearly connected suicide rates to individuals' social integration in the community.

Variables included in this model are individual attributes such as gender and race, pre-college experiences, and family backgrounds. Tinto argues that these variables influence the development of college expectations and commitment to graduation. These expectations and commitments are modified based upon integration into the college system academically and socially to generate a new level of commitment and goals.

The author noted that there is still little information that links race with college dropouts, although it is considered a strong predictor of student persistence. Tinto further added that there isn't enough knowledge about the process of interaction that leads racial groups to drop out and how these processes are affecting their academic and social integration (Tinto, 1975).

1.1.2 Astin's Input-Environment-Output model

Astin (1991) in his book "Assessment for Excellence: The Philosophy and Practice of Assessment and Evaluation" developed a conceptual model which is known as the I-E-O model. The model stated that researchers should focus not only on outcomes when assessing educational programs and practices but also on input characteristics and educational environment. Astin defined student inputs as pre-college characteristics (e.g., race/ethnicity, gender, and family background), college admission tests and high school GPA, and student self-reported data (e.g., goals and college expectations). He addressed the importance of input data because it influences student output data and most likely influences the educational environment (Astin, 1991). Educational environment was defined as everything students experience academically and socially during college that somehow affects their educational outcomes such as joining first year programs and student organizations. In another study, Astin (1999) argued that the lack of involvement in college environment was a significant cause of student withdrawal from college. Educational outcomes refer to the college impact on student.

1.1.3 Terenzini and Pascarella

Terenzini and Pascarella's (1980) study was developed based on Tinto's (1975) model of student dropout using statistical analysis methods. The study used three random samples of freshmen at Syracuse University between 1974 and 1976. A total of four studies were used to test Tinto's model in addition to two studies that focused on the faculty integration part of the model.

Terenzini and Pascarella's major findings are the following:

- Academic and social integration of freshmen were found to be statistically reliable with freshmen persistence.
- Pre-college factors are important in student persistence/dropout based on how they interact with college experiences.
- Frequency and quality of student-faculty contact outside the classroom is positively related to student persistence/dropout behavior.

1.1.4 Other studies

Several reasons have been correlated with college retention with specific focus on the fields of science and engineering such as adequate high school preparation, difficulty in adjusting to college life, lack of engineering community atmosphere, limited exposure to engineering courses in the first and second year, and financial obligations (Nicklow, Kowalchuk, Gupta, Tezcan, & Mathias, 2009). Reason (2003) reported that specific student features such as race/ethnicity, GPA, gender, and institutional features such as selectivity and student integration into academic life are the main factors that affect retention. Anderson-Rowland (1996) in an Arizona State University student survey identified employment demands, financial problems, and family issues as the three main causes of engineering student dropout. Tinto (1995) in his speech "Taking Student Retention Seriously" believed that there are five conditions that support retention, "namely expectation, advice, support, involvement, and learning."

Research has shown that four groups of factors affect the low retention rates of minority students in science and engineering. These include "academic and social integration, knowledge and skill development, support and motivation, and monitoring and advising" (Maton & Ozdemir, 2007).

Furthermore, the literature review identifies first year college success as a significant impact on student retention (Reason, 2003; Crawley, Malmqvist, Ostlund & Brodeur, 2007; Persaud & Freeman, 2005; Sidle & McReynolds, 1999; Nicklow, Kowalchuk, Gupta, Tezcan, & Mathias, 2009; Roberts, 2009). For about two decades, research has shown that student performance and GPA in first and second semesters are crucial predictors of student retention (Heywood, 2005; May & Chubin, 2003; Tan, 2002). Heywood (2005) went as far as identifying the significant role that the first few weeks play in shaping student motivation and attitude toward college life.

1.2 Data Mining Models in predicting student academic success and retention

Research has shown that tracking students who transfer from STEM disciplines to a non-STEM disciplines is an increasingly difficult process (Mendez, Buskirk, Lohr, & Haag, 2008). Thus, several studies have emphasized the importance of identifying college students with higher risk of dropping out in early stages and allocating the available resources based upon student needs (Zhang, Anderson, Ohland, Carter, & Thorndyke, 2002; Dekker, Pechenizkiy, & Vleeshouwers, 2009). Studies have varied in identifying factors that affect student retention the most, especially in their freshman year. Mendez claimed that high school GPA and scores on placement tests, in addition to grades in math, chemistry, and physics, are all strong predictors of engineering student retention (Mendez, Buskirk, Lohr, & Haag, 2008).

Gaskins (2009) has emphasized that student predefined variables such as high school GPA combined with environmental variables such as student living (on campus or off campus) and involvement in first year programs such as a residential living learning community are best predictors of student success. The study was conducted over a ten year period (fall 1997 through fall 2006) and 35,050 students were involved from all majors. Logistic regression was the main statistical method used in this study to categorize students into "retained" and "not retained." The study reported that student success differed between students, institutions, and even different schools within the same institution. As a result, variables of high school GPA, on-campus living and involvement in a first year program were cited as significant in affecting student retention and success in their freshman year.

Data mining methods are becoming more popular and accurate in modeling student performance and retention in higher education. In a data mining project that used 1,508 incoming engineering freshmen at a large Midwestern university during the 2004-2005 academic year, several methods for modeling first year student retention in engineering, such as neural networks, discriminant analysis, logistic regression, and structural equation modeling (Lin, Imbrie, & Reid, 2009), were used. Each model used several pre-college factors that are believed to affect student retention such as high school GPA, standardized tests, and high school math, physics, and chemistry grades to build a framework that predicts engineering student retention. Neural networks proved their superiority among the other four methods used in terms of prediction accuracy.

Herzog (2006) conducted two studies; one focused on studying student retention, which used forty variables, and the other focused on time to degree, which used 79 variables, in all majors. Three-rule induction decision trees (C&RT, CHAID-based, and C5.0) and three backpropagation neural networks (simple topology, multitopology, and three hidden-layer pruned) with a multinomial logistic regression model were compared to examine the most accurate model that predicts student retention and time to degree. To validate the developed models, data was randomly split fifty-fifty to test the accuracy of different models. The study revealed that neural networks and decision tree techniques provided a stronger analysis and better accuracy when predicting student retention and time to degree using a large data set.

2. Research Questions

The purpose of this research was to develop a hybrid framework to model first year student retention for URM comprising African Americans, Hispanic Americans, and Native Americans. Prior to developing this hybrid framework, results of the genetic algorithm and focus groups were analyzed and incorporated.

The examined research questions were:

1. Which student characteristics, environmental influences, and academic support services impact first year student retention in STEM disciplines the most?
2. To what extent did first year college experiences and academic progress affect pre-defined goals of URM students and their intention to graduate with a STEM degree?

Identifying inputs that best contribute to student retention provides significant information for institutions to learn about student needs, how to support student academic success, and how to increase retention in STEM fields. Institutions can also rely on using qualitative analysis to examine students' experiences during the freshman year to acquire useful information on different student retention behaviors from a diverse population. Based on this information, better programs and student services can be developed.

This research has used the neural network technique, which is commonly employed for modeling and machine learning. In this study, the results of genetic algorithms and qualitative methods were incorporated into modeling freshman year retention.

3. Methodology

In this study retention is defined in terms of students who stay in a STEM discipline from the first fall of enrollment to the second fall. Students who switch from one STEM discipline to another are considered retained, while students who switch to a non-STEM major are considered non-retained. Due to the nature of the study in terms of the availability of student information, in which it focuses on fall-to-fall retention at a large public research institution, all students included in this study were enrolled in the fall semester of their sophomore year. The model uses pre-college information, college characteristics, and demographic attributes of students to identify significant factors that impact the decision of persistence/dropout from a STEM discipline. The STEM majors included in this study are: Biology, Chemistry, Physics, Science, Forensic Sciences, Mathematical Sciences, Bioinformatics, Environmental Studies, Computer and Electrical Engineering, Biomedical Engineering, Mechanical Engineering, Chemical and Life Science Engineering, and Computer Science.

Participants of this study fall into two groups:

1) The first group comprised of STEM full-time first year students from the 2007–2009 academic years. Data was obtained from the Office of Institutional Research. The sample size consisted of 1,966 students who started with a STEM discipline in the first fall semester of enrollment. The dataset contains records of both male and female students from different ethnic origins. Ethnic origins are classified as follows: American Indian, Asian, African American, Hispanic, unknown/not specified, and white. In this study, the dataset was divided into two cohorts: first, the majority student cohort that includes a total of 1,468 students and second, the URM student cohort with a total of 498 students. The majority student cohort includes Asian, unknown/not specified, and white ethnic origins, while the URM student cohort includes American Indian, African American, and Hispanic American ethnic origins. The unknown/not specified represents less than 8% of the overall majority student's population. To protect students' anonymity, no identifiable student information was included.

2) Sixty-three participants in a summer transition program at a large public research institution over a three-year period (2008-2010) were invited to participate in the focus groups sessions. The program participants were incoming freshmen in STEM disciplines who were African American, Hispanic American, and Native American. It is a self-selecting program designed to enhance participants' pre-college preparation and ensure a smooth transition into college. Each year, approximately twenty-two participants choose to enroll in the program. Participants' majors were biology, all engineering fields, mathematical sciences, forensic sciences, chemistry, and environmental studies. Of the participants, approximately 59% were female. Sixteen students attended the three meetings conducted in the spring of 2011; two student attendees were not summer transition program participants. These two students responded to an invitation for non-participants to get insight into other freshman year experiences for students who did not have a chance to participate in the program.

This particular group was included because they were exposed to a variety of activities and programs prior to and during their freshman year. It is believed that this group of students would be able to provide valuable responses and compare their experience with their peers who did not participate in any first year programs and/or activities. The group represents diverse backgrounds and ethnic origins and is comparable with the institution's population.

3.1 Data Collection

This study will develop a hybrid model that predicts URM student retention. The model incorporated both relevant factors that are determined using genetic algorithms and qualitative method via focus groups conducted to understand students' first-year experiences.

Two different datasets were used for both quantitative (neural networks and genetic algorithm) and qualitative (focus groups) methods:

1) Data used in this study was obtained from the Office of Institutional Research, covering a three year period (2007–2009) for all freshmen who started with a STEM field. Student inputs that were included have been classified into three categories: demographic, pre-college, and college variables. The demographic variables included in this study are race/ethnicity, residency, and gender. The pre-college variables are honors, math SAT scores (SATM), verbal SAT scores (SATV), combined SAT scores (SATC), high school percentile rank (Rank), and first math course (CourseM). The college variables are term credits attempted in the first fall (TCAF), term credits earned in the first fall (TCEF), credits attempted in the first fall (CAF), credits earned in the first fall (CEF), term credits attempted in the first spring (TCAS), term credits earned in the first spring (TCES), credits attempted in the first spring (CAS), credits earned in the first spring (CES), first mathematics course grade (GradeM), fall term GPA (TGPAF), and spring term GPA (TGPAS). Two response variables were used in this study to build two predictive models: the first is GPA and the second is retention. The first model, GPA model, used all available student inputs except two variables, which are Term GPA in fall and Term GPA in spring. The retention model used all twenty student inputs in addition to GPA. This study included many factors which were identified by most of the related studies as influential factors on student performance and college retention such as race/ethnicity, gender, college GPA, mathematics grades, standardized test scores, and placement test scores.

2) The qualitative data was obtained by conducting focus groups for summer transition program participants over a three year period (2008-2010). The collected data focused on identifying significant student characteristics from the students' point of view. Furthermore, focus group sessions collected information on URM students' first year college academic and social experiences. An approval from the Institutional Review Board for Research Including Human Subjects (IRB) was obtained. Sixteen students participated in the three sessions: 9 in the first session, 5 in the second session, and 2 in the third session.

In this study of 1,966 students it was observed that females represent higher a proportion 0.52 (1,019 samples) compared to males. With a total of 1,468 majority students, it was observed that males represent a higher proportion of 0.55 (809 samples) as compared to females. However, regarding the 498 URM students included, it was observed that females represented a higher proportion, 0.72 (360 samples).

Although females represented a higher percentage in the student population, their retention rate was 10% less than that of males. Students who started with a higher level of mathematics had higher retention rate, and students who did not perform well in their first mathematics course were less likely to be retained in their STEM major. The average SAT score was 1138 and 1062 for retained and non-retained students, respectively. The average high school rank was 78% for retained students and 76% for non-retained. The overall freshman year GPA was 3.0 for retained students and 2.8 for non-retained. The average was 38 college credits for retained students and a total of 33 college credits for non-retained students.

More specifically, URM female students represented a higher proportion in STEM population, i.e. 0.72 (360 samples). However, their retention rate, 70%, was lower than the retention rate of males, which was 81%. Majority students with higher level of mathematics and better grades were more likely to be retained in their STEM major. The average SAT score was 1035 and 980 for retained and non-retained students, respectively. The average high school rank was 78% and the overall freshman year GPA was 2.8 for both retained and non-retained students. The average total college credits earned were 33 for retained students and 31 for non-retained.

3.2 Neural network models design

The FeedForward backpropagation network was used to model first year student retention in STEM disciplines in a large public research institution. The number of hidden layer neurons used was between 2–4 where each neuron has a hyperbolic tangent (tanh) activation function. The network's output layer for predicting the overall student GPA has a linear activation function (purelin) while the output layer for predicting student retention has the hyperbolic tangent activation function. The training function used is Levenberg-Marquardt. The algorithm is an iterative technique that adjusts the weights to minimize the difference between the actual and predicted output.

Each model was built using student inputs in two different ways: 1) using all available student inputs 2) using an optimized dataset which was obtained from the genetic algorithm. Within each model, performance was compared when different student inputs were used. The procedure above was repeated for two different datasets, URM and majority students.

This study focuses on the achievements of URM student in STEM majors. Thus, results obtained from both methods were incorporated to develop a comprehensive model that is able to predict URM students' first year retention accurately.

To validate the neural networks models, the 10 fold cross-validation was used. The training set was randomly divided into 10 parts, nine of which were for training and the rest for testing. The process was repeated 10 times, and then the accuracy of the model was computed.

3.2.1 *Neural network model performance*

The response variable used in this study for retention is a categorical variable of two values, retained or not. The retention model's accuracy (ACC) was calculated by adding the number of correctly predicted retained students (TP) to the number of correctly non-retained students (TN) and dividing the resulting number by the total number of students included (N) as in the following equation:

$$ACC= (TP+TN)/N \qquad (1)$$

Whenever the 10 fold cross-validation is used, prediction error of each training set is calculated and the final error is the total error for all the training sets.

3.2.2 Feature Subset Selection

To build an effective model, it is important to select a non-redundant subset of student inputs which are relevant to the output variable. When using neural networks, learning time is increased if a large set of variables are used. In neural networks, the genetic algorithms technique gives good results for feature selection.

The feature subset selection was used to provide a deep insight into freshman retention and academic success in STEM disciplines. The output of the genetic algorithm is a vector of binary values at the best fitness value, which in our case is the root mean square error (RMSE). The mutation rate used was 0.01, and the selection function used was roulette wheel, which is a commonly used function for feature selection. This selection function makes a random selection similar to the rotation of the roulette wheel to select the best fit. 100 generations were used and the population (chromosome) size was chosen to be 20. The algorithm accepts a vector of student inputs and returns a bit string that indicates whether the feature was selected. If the feature is selected, it gets a value of 1; otherwise it gets a 0 value. The dataset was divided into two groups based on student race/ethnicity (URM or majority) to compare and contrast the two resulting vectors.

3.3 Focus group instrumentation

Qualitative research methodologies are effective in terms of analyzing non-quantitative data or data in the form of text rather than numbers. Researchers defined qualitative research as "important modes of inquiry for the social sciences and applied fields, such as education, regional planning, health sciences, social work, community development, and management" (Marshall & Rossman, 2010).

The focus group protocol was designed for this study to elicit responses from participants about their freshman year college experiences and determine which variables have the most impact on student academic success and retention. Seven open-ended questions were asked of each group, and students were informed about the confidentiality of all the sessions. The first question discussed reasons behind students' motivation to major in STEM fields. The second and third questions focused on analyzing freshman year experiences, the difficulties participants had, and how they handled them. The fourth, fifth, and sixth questions determined which academic, demographic, and social variables have the most impact on student academic success and retention. The final question examined the extent to which precollege intervention programs could affect student retention in a STEM discipline. The analysis approach used is content analysis, which is a very effective method in analyzing data in textual context.

3.4 Hybrid Model Design

The hybrid framework is developed to model first-year student retention for URM. In this model, the results obtained from the quantitative methods (genetic algorithms) and qualitative methods (focus groups) were incorporated to develop a comprehensive model capable of predicting URM students' first-year retention. The main goal of incorporating the results is to build a simple and interpretative tool that could be used effectively to impact URM students' accomplishments during their freshman year.

The FeedForward backpropagation network architecture used to develop this model and the number of hidden layer neurons used was 3 where each neuron has a hyperbolic tangent activation function. The network's output layer for predicting student retention has the hyperbolic tangent activation function. The training function used is Levenberg-Marquardt. The 10 fold cross-validation used to validate the neural networks models.

Features	All Students	Majority Students	Minority Students
Race	0	-	-
Residency	0	1	0
Gender	1	1	1
Honors	0	1	0
TCAF	0	1	0
TCEF	1	1	0
CAF	1	0	1
CEF	1	1	1
TGPAF	1	0	0
SATM	1	1	0
SATV	0	1	1
SATC	1	1	0
RANK	1	0	1
TCAS	1	1	1
TCES	0	1	0
CAS	1	1	1
CES	1	1	0
TGPAS	1	1	1
GPA	1	1	1
CourseM	1	0	1
GradeM	1	1	1

Table 4. Output of Retention model feature subset selection by group

For the retention model the accuracy is used to interpret the model's performance and compare it with the other retention models that used genetic algorithms output as an input set.

4. Results

4.1 Model Performance

This section describes the results obtained by using retention model using three dataset cohorts—all students, majority and URM cohorts. The performance of the neural network of the different datasets is compared and the network's number of hidden layer neurons is selected by trial and error until the best performance achieved.

Of the 1,966 samples, 74% cases were predicted correctly which considered good prediction accuracy. The majority student model's accuracy was 79%, which also considered a good performance in predicting majority student retention in STEM disciplines. As for the URM student retention model, the model did not perform as well as the previous models with an accuracy of 60%.

The neural networks model for predicting majority student retention achieved better accuracy compared to models of all students and URM students. However, the URM student model did not perform as well as the other two models. In general, the accuracy between 70%-80% is categorized as good, while the accuracy between 60%-70% is categorized as fair.

4.2 Student Feature Optimization Results

Genetic algorithms are considered a very efficient way for feature subset selection. It is usually used to reduce the model's complexity, reduce the learning time of the network, enhance generalization, and improve performance. In this study, the genetic algorithms technique was used to identify the most relevant features that impact student retention the most.

4.2.1 Feature subset selection

The results show the output of the genetic algorithm to select the most influential student features in student retention behavior in a STEM discipline as shown in table 4. It was observed that seven features were common among the three groups: gender, credits earned in fall semester (CEF), total credits attempted in spring semester (TCAS), credits attempted in spring semester (CAS), term GPA of spring semester (TGPAS), overall freshman year GPA (GPA), and grade of first mathematics course (GradeM). This gives an indication of the influence of the overall GPA on student retention decision. Rank was not selected for majority students, although it was selected for the URM group. The Honors feature was not selected for the URM group. The SATM and the SATC were not selected for the URM group, while the SATV was selected for the majority and the URM groups. The total college credits attempted in the first fall and the first mathematics course were not selected for the majority group. The total college credits earned in the fall semester was not selected for the URM group as well.

4.3 Modeling freshman Retention with Reduced Feature Set

Of the 1,966 samples, it was indicated that the accuracy of the model slightly improved when feature selection set was used. The model's accuracy increased from 74% to 75% when the optimized set was used. It was observed that the majority student model's accuracy increased approximately 2% when the selected subset of features was used as input, from 79% to 81%. As for the URM student retention model, the accuracy of the model increased 3% when the selected subset of features was used, from 60% to 63%.

The neural networks model for predicting majority student retention achieved better accuracy compared to models of all students and URM students. In general, the network's accuracy was improved for the three groups when an optimum input features used. A significant improvement was observed for the majority students group when the optimum set of features was used. The model achieved an accuracy of 81%, which is considered a very good model in predicting student retention.

4.4 Focus Group Sessions Analysis

In this study, a total of sixteen students participated in the three sessions: twelve females and four males. Fifteen participants were African Americans, and one was Hispanic American. All participants were majoring in STEM disciplines, except one student who switched from STEM to a major in business administration. Seven students indicated that they were the first generation to go to college, and only three students declared that they work during the academic year.

The first question discussed reasons behind students' motivation to major in STEM fields. The major themes that arose from the students responses included the critical role family members, mainly parents, played in influencing their decision to consider majoring in STEM. Some students saw their parents as role models and tried to follow their steps and pursue a career in STEM fields. Relatives and friends were a good source of motivation as well. A second theme arose, high competency in mathematics and science; students realized that STEM fields are commensurate with their career goals and abilities. To a lesser degree, students identified the impact of high school teachesr and the media (inspirational TV shows).

The second and third questions focused on analyzing freshman year experiences, the difficulties participants had, and how they handled them. Students' responses varied when they were asked to evaluate their freshman year experience. A majority of students responded that it was easy. Academically, students referred to their high school preparation and participation in science and engineering programs such as the summer transition program as factors that helped in making first year introductory courses easier.

A few students described their freshman year experience as moderate. These students mostly had difficulties in academic adjustment. For example, students who came with AP credits and were placed in the advanced courses had more pressure as a freshman in a sophomore-level class. Three students said that their freshman year experience was difficult, but overall they enjoyed it. Being away from home and taking all the responsibility of being placed in upper-level classes was the difficult part of the experience. Also, some found it hard to balance between priorities. One of the students in the first group who did not participate in the summer transition program described her freshman year experience as "lonely."

One of the major difficulties student faced during their freshman year was their first chemistry class. Even though most of them took a chemistry class during the summer transition program, it was hard for them to keep up with such a demanding course and grasp any new material. Students revealed that they had to work harder, get tutoring, and attend other chemistry classes taught by different instructors. Upper level classes such as differential equations, physics, and programming were on the list of difficult courses as well. It was observed that these courses required a larger workload than expected for a freshman, especially if all three were taken at the same time and if the freshman did not take physics in high school. Online courses were a problem for freshmen as well. A student revealed that he was not ready for the kind of classes that put more responsibility on himself to check homework and due dates online without having someone remind him about the class duties.

Socially, all students showed their concern for adjusting to college life and the new environment. Students from the three groups agreed that distractions and peer pressure were difficult things to handle in freshman year. Students came to college, lived with roommates, and had no curfews as they used to have in high school. It was hard to take the full responsibility to avoid these distractions and maintain academic success. Students from the second group stated that the whole new teaching environment and the campus life were not as they expected them to be when they came to the summer transition program.

The fourth, fifth, and sixth questions determined which academic, demographic, and environmental variables have the most impact on student academic success and retention. High school preparation was a significant indicator of freshman year performance for almost all the students. A majority of students revealed that their high school mathematics and science background helped them to get good grades in their first semester's introductory courses. This was unlike what we observed from the previous questions, as some students complained about their weak chemistry and physics preparation and how difficult it was for them to handle those classes. A few students highlighted the impact of their strong support system, family, and friends on their freshman year performance. Usually family members keep up with the students and try to push them to achieve academic success._

AP classes were among the significant indicators of good performance in freshman year and none of the students said that SAT scores were an indicator of their freshman year performance even when they were asked about it. Self-motivation and the ability to be independent were among the top freshman year performance indicators as well. Most students in the three groups said that they had not thought of switching to another major because they did well in their classes or got a good GPA, especially in their first semester. They also added that this increased their self-motivation that they can do even better if they worked harder, in spite of facing any possible difficulties.

From a demographic perspective, it was observed that gender was not an issue for any male student and non-engineering female student. However, almost all female students in engineering indicated that it was challenging and motivating at the same time for them to be "a minority within a minority," referring to gender and race. One engineering male student stated that he came from a high school where 90% of the population was African American, and now he is the only African American in his major. The student added that he feels pressure to be successful as one of the only African Americans graduating with this major. One student pointed to the safe and diverse environment as a good indicator of freshman year performance.

As for environmental factors that affected their academic performance and retention, students identified family and friends as most influencing their retention decision. Several students revealed that their classmates were very helpful, too, especially in large classes where it was hard to build a relationship with the professor; some students stated that they usually refer to upper-class students because they know the material, study habits, and the best teachers and can give the best advice.

Student advisors and professors were among the influential factors on student retention. Some students pointed out that their advisors did not help, while others stated that their professors did not influence them; some stated that it depends on the professor, and few stated that their professors were very helpful whenever they needed their assistance. Some noted that professors are nicer and more supportive in their offices than in class. In addition, a student claimed that teaching assistants sometimes could be more helpful than professors themselves. Money and roommates also were on the list as both positive and negative influential factors.

The final question examined the extent to which precollege intervention programs could affect student retention in a STEM discipline. The summer transition program impact on participants' pre-college preparation was divided clearly into academic and social. A majority of participants in both groups stated that the program was more helpful from the social perspective. Students said that they made new friends with diverse experiences, became familiar with the college environment, did not get lost in the fall semester, adjusted to being away from home before fall started, learned time management, got used to campus and city life, gained good dorm experience, especially when they had a roommate with the same major, and found the study skills class to be good. One student, however, from the first group stated she did not utilize it well.

Academically, students from the first group stated that they learned how college classes are and realized that they need to work harder, boosted their self-esteem when they got good grades during the program, got more confident in freshman year classes, and found a study buddy. The second and third groups agreed that the mathematics and chemistry classes served as a good review before the beginning of fall semester. Some students from the second group stated that they knew what to expect in college, and the science class helped in learning how to write laboratory reports. The third group's students stated that the study skills class was good in teaching them time management.

4.5 The Hybrid Model

The most relevant student features obtained from the genetic algorithm and focus groups were selected to build the hybrid model. The model was developed using neural networks and then validated using the 10 fold cross-validation. The hybrid model diagram is shown in Figure 1.

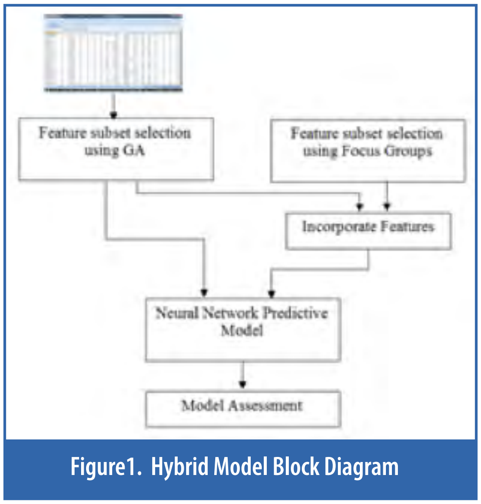
Feature subset selection using GA
Feature subset selection using Focus Groups
Incorporate Features
Neural Network Predictive Model
Model Assessment

Figure1. Hybrid Model Block Diagram

It was observed that the integrated majority student's model performed the same as the model that used genetic algorithm optimized set of student inputs. When attempting to predict URM student retention, the hybrid model improved prediction by 3%, increasing in accuracy from 63% to 66%.

5. Discussion

Incorporating results obtained from the genetic algorithm and focus groups was a challenge in this study. The genetic algorithm relies on mathematical calculations to determine which student inputs are most relevant to student retention without any direct interaction with students. Focus groups are used to elicit direct responses from participants regarding their college experiences and key factors that have a significant impact on their achievements. Thus, the goal was to incorporate results—not compare them—and to develop a hybrid predictive model and validate it using a 10 fold cross-validation.

The developed model presents students as interactive entities in the system instead of just numbers. The student could be identified and his/her inputs could be analyzed to build a profound knowledge of different performance and retention behaviors. The qualitative analysis of URM student freshman year experiences would play a positive role in analyzing student performance and retention behaviors.

Approximately 25% of freshman population in the studied institution in STEM fields is made up of URM students. Consequently, it is useful to use the hybrid framework to model student retention as well as analyze significant factors of targeted students in the prediction process and gear available resources and intervention programs based on student needs. The process of putting all the pieces together would be completed by analyzing student college experiences to address differences in human behavior.

Genetic algorithms select the chromosomes at random from the design space and might not select all possible chromosomes. Due to this, the optimized values of the parameters might not be the global optimum. Instead they might only be localized optimal value. Because of this randomization used in genetic algorithms, results obtained from the hybrid model were either the same or slightly different compared to results obtained when feature selection was used.

The model's performance could be improved by increasing the sample size and increasing the number of included features.

6. Conclusions and Limitations

Modeling retention for URM students in STEM disciplines and analyzing key factors that impact student accomplishment, in addition to understanding student first year educational experience, can effectively build a learning environment and strategies that lead targeted students to the right path to success.

High school academic mathematics and science preparation has a great impact on student freshman year accomplishments. High school rank, SAT mathematics scores, and mathematics placement test scores were considered strong predictors of retention. A major part of this study was to construct an adequate understanding of URM student persistence/dropout behaviors in STEM fields.

The institution included in this study has several intervention programs and activities to support students. Usually, these programs and activities are self-selecting where students choose whether to participate or not. Many students who need help are left behind because they do not know where to go or they participate in a different program that cannot address the students' needs for succeeding in a STEM major. Therefore, leading targeted students to success and retaining them in a STEM field is not just the responsibility of students themselves but is as much a responsibility of their family, friends, advisors, teachers, and the surrounding environment. It was found that URM students come to college with high self-motivation and commitment to graduate with a degree in STEM. Once college starts, many factors impact student self-motivation either positively or negatively. Empowering a student with self-motivation has a great influence on the student's decision to continue in STEM fields.

It was revealed that freshman intervention programs improve student performance and increase retention in STEM fields. Such programs were effective academically and socially. Participants gained more self-confidence and reviewed essential material of gateway classes. Also, it was found that student learning is a continuous process where students seek assistance outside the classroom to improve their performance. Overall, high school mathematics and science preparation, race, gender, and freshman year grades are strong predictors of student retention. In addition, freshman year cumulative GPA is a strong predictor of student retention.

Overall, the neural networks model performance results were similar when different input sets were used. The model's accuracy when all student inputs used was 74%, 79%, and 60% for all students, majority students, and URM students, respectively. The model's accuracy slightly improved 1%, 2%, and 3% for the three different datasets (all students regardless to their ethnicity, majority students, and URM students). A 3% increase (66%) of the model's accuracy was observed for the developed hybrid model.

Overall, the network's accuracy was improved using an optimum set of student inputs for the three groups. However, the network's accuracy of the majority group was much higher than the URM network's accuracy. This could be due to the difference in the size of both samples for the majority (N=1468) and URM (N=498) groups linked with the student inputs used as well. Thus, this research paves the way for future research to use additional significant inputs that identified by URM students point of view in order to increase the model's accuracy. However, the resulted hybrid system is a simplified and easier-to-interpret model.

The related research work presented (Herzog, 2006; Lin, Imbrie, Reid, 2009; Zhang, Anderson, Ohland, Carter, & Thorndyke, 2002; Gaskins, 2009; Mendez, Buskirk, Lohr, & Haag, 2008; Dekker, Pechenizkiy, & Vleeshouwers, 2009; Thai-Nghe, Janecek, & Haddawy, 2007; Superby, Vandamme, Meskens, 2006; Lam, Doverspike, & Mawasha, 1999) targeted different student populations, different input features, and different methodologies. Hence, it is hard to make a direct comparison between the accuracy of the developed framework presented in this research and accuracy of the other developed frameworks. Moreover, this research incorporated results obtained from the genetic algorithm and focus groups to build a model that includes the most relevant student features in order not only to model student retention but also provide a deep insight into student freshman year experiences and different retention behaviors. To our knowledge, the presented method of incorporating results of genetic algorithm and focus groups is new to the field of modeling student performance and retention, especially for URM students.

However, these models were comparable to those of other studies results. In one study (Herzog, 2006) the neural networks model accuracy for predicting student retention was between 77% and 84%. The study used different data sets and different input sets of student features. Another study used different data mining techniques to predict retention of electrical engineering students over 10-year period and achieved accuracy between 75% and 80%. Thus considering the sample size and student features used in this study, the developed model performance is effective in modeling retention for students in STEM disciplines, especially for URM students.

Several limitations existed for this study. The first set of limitations was in the sample of the quantitative part which was limited to first-time first year students in STEM. Thus, transfer students were not included. In addition, students who dropped out or suspended their studies during the first or second semester (e.g., did not register for the fall semester) were excluded since the study focused on fall to fall retention. Also, there was no indication whether the student gained any college credits regarding to participation in a pre-college programs or from taking AP classes other than a general idea by comparing the total freshman year college credits earned per semester with the college credits earned in the fall semester. A limited number of URM students (N=498) were included in this study, which affected the network's accuracy.

The second set of limitations was in the focus groups sample. Data were collected from 16 URM students who participated in an intervention program for incoming freshmen in STEM disciplines since 2008. Participants of the focus groups were diverse and comparable to the larger group of the program participants. However, the study collected a survey prior to conducting the focus group sessions, which might cause results to be skewed regarding self-reporting issues.

References

Alkhasawneh, R &Hobson, R. (2009). Summer Transition Program: A Model for Impacting First-Year Retention Rates for Underrepresented Groups. P*aper presented at the 2009 American Society for Engineering Education Annual Conference, Austin, TX.*

Anderson-Rowland, M. (1996). A first year engineering student survey to assist recruitment and retention. pp. 372-376.

Astin, A. (1991). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education:* Oryx Pr.

Astin, A. (1999). Student involvement: A developmental theory for higher education. *Journal of College Student Development*, vol. 40, pp. 518-529.

Besterfield-Sacre, M., Atman, C., & Shuman, L. (1997). Characteristics of Freshman Engineering Students: Models for Determining Student Attrition in Engineering. *Journal of Engineering Education-Washington-, vol. 86, pp. 139-150.*

Brown, J. D. (2007). Neural Network Prediction of Math and Reading Proficiency as Reported in the Educational Longitudinal Study 2002 Based on Non-Curricular Variables. *(Doctoral dissertation, Duquesne University).*

Crawley, E., Malmqvist, J., Ostlund, S., & Brodeur, D. (2007) *Rethinking engineering education: The CDIO approach: Springer Verlag.*

Dekker, G., Pechenizkiy, M. & Vleeshouwers, J. (2009). Predicting Students Drop Out: a Case Study, *In Proceedings of the 2nd International Conference on Educational Data Mining (EDM'09), pp. 41-50.*

Durkheim, E. (1951). Suicide: A study in sociology (JA Spaulding & G. Simpson, Trans. *Glencoe, IL: Free Press.(Original work published 1897).*

Fleming, L., Ledbetter, S., Williams, D., & McCain, J. (2008). ENGINEERING STUDENTS DEFINE DIVERSITY: AN UNCOMMON THREAD. *Paper presented at the 2008 American Society for Engineering Education Annual Conference, Pittsburgh, PA.*

Gaskins, B. (2009). A Ten-Year Study of the Conditional Effects on Student Success in the First Year of College. (Doctoral dissertation, Bowling Green State University).

Hargrove, S. & Burge, L. (2002). Developing a six sigma methodology for improving retention in engineering education. *ASEE/IEEE Fronties in Education Conference. pp. S3C20-24.*

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. New Directions for Institutional Research, vol. 2006, p. 17.

Heywood, J. (2005). *Engineering education: research and development in curriculum and instruction: IEEE Press.*

Jody Markley, C.F.F. (2005). Integrating a Faculty Directed Research Experiences into a High School Bridge Program. in *WEPAN/NAMEPA Joint Conference*.

Lam, P.C., Doverspike, D. & Mawasha, R.P. (1999). Predicting success in a minority engineering program. *Journal of Engineering Education, 265-267*.

Lin, J., Imbrie, P.K., Reid, K.J. (2009). Student Retention Modelling: An Evaluation of Different Methods and their Impact on Prediction Results. *In proceedings of the Research in Engineering Education Symposium, Palm Cove, QLD.*

Marshall, C. & Rossman, G.B. (2010). Designing qualitative research. *Sage Publications, Inc.*

Maton, K., Ozdemir, M. (2007). Opening an African American STEM Program to talented students of all races: evaluation of the Meyerhoff Scholars Program, 1991–2005. *Charting the Future of College Affirmative Action: Legal Victories, Continuing Attacks, and New Research, ed. G. Orfield, P. Marin, SM Flores, and LM Garces, Los Angeles, CA: The Civil Rights Project, UCLA.*

May G. & Chubin, D. (2003). A retrospective on undergraduate engineering success for underrepresented minority students. *Journal Of Engineering Education-Washington, vol. 92, pp. 27-40.*

Mendez, G., Buskirk, T. D., Lohr, S., & Haag, S. (2008). Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. *Journal of Engineering Education 97(1), 57–70.*

Mitchell, T. & Daniel, A. (2007). A Year-Long Entry-Level College Course Sequence for Enhancing Engineering Student Success. *ICEE, Coimbra, Portugal*

Nave, F., Frizell, S., Obiomon, P., Cui, S., & Perkins, J. (2006). Prairie View A&M University: Assessing the Impact of the STEM-Enrichment Program on Women of Color. In B. Bogue & R. Marra (Eds.). *In proceedings of 2006 Women in Engineering Programs & Advocates Network (WEPAN) Conference. Pittsburgh, PA.*

Nicklow, J., Kowalchuk, R., Gupta, L., Tezcan, J.,Mathias, J. (2009). A short-term assessment of a multi-faceted engineering retention program. *39th ASEE/ IEEE Frontiers in Education Conference. San Antonio, TX. pp. 424-429.*

Persaud, A. & Freeman, A. (2005). Creating a Successful Model for Minority Students' Success in Engineering: The PREF Summer Bridge Program. *In Proceedings of the WEPAN/NAMEPA Joint Conference.*

Reason, R. (2003). Student variables that predict retention: Recent research and new developments. *NASPA JOURNAL*, vol. 40, pp. 172-191.

Roberts,S., et al. (2009). Evaluation of Retention and Other Benefits of a Fifteen-Year Residential Bridge Program for Underrepresented Engineering Students. *Presented at the American Society for Engineering Education, Austin, TX*

Sidle, M. & McReynolds, J. (1999). The freshman year experience: Student retention and student success. *NASPA Journal, vol. 36, pp. 288-300.*

Superby, J.F., Vandamme, J-P., Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. *Workshop on Educational Data Mining. pp. 37-44.*

Tan, D. (2002). Majors in science, technology, engineering, and mathematics: Gender and ethnic differences in persistence and graduation. *Norman, Okla: Department of Educational Leadership and Policy Studies.*

Terenzini, P. and Pascarella, E. (1980) Toward the validation of Tinto's model of college student attrition: A review of recent studies. *Research in Higher Education*, vol. 12, pp. 271-282.

Thai-Nghe, N., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. *In Proceedings of 37th ASEE/ IEEE Frontiers in Education (FIE 2007), IEEE Xplore, pp. T2G-7-T2G-12.*

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, vol. 45, p. 89, 1975.

Tinto, V. (1988). Stages of student departure: Reflections on the longitudinal character of student leaving. *The Journal of Higher Education, vol. 59, pp. 438-455.*

Tinto, V. (1995). Taking student retention seriously. *Syracuse University, Syracuse, NY.*

Urban, J.E., Reyes, M.A., & Anderson-Rowland, M.R. (2002). Minority engineering program computer basics with a vision. *fie, vol. 2, pp.S3C1-5, 32nd Annual Frontiers in Education (FIE'02)*

Williford, A. M., & Schaller, J. Y. (2005). All retention all the time: How institutional research can synthesize information and influence retention practices. *In proceedings of the 45th Annual Forum of the Association for Institutional Research, San Diego, 29 May—1 June*

Zhang, G., Anderson, T., Ohland, M.,Carter,R., & Thorndyke, B. (2002). Identifying Factors Influencing Engineering Student Graduation and Retention: A Longitudinal and Cross-Institutional Study. *In proceedings of the American Society of Engineering Education Southeast., Gainesville, FL, April 2002, Session 2793.*

Dr. Ruba Alkhasawneh is currently a System Validation Engineer at Intel Corporation in Austin. She received her B.S., M.S. and Ph.D. degrees in Computer Engineering from Jordan University of Science and Technology (JUST), Yarmouk University, and Virginia Commonwealth University (VCU) respectively. Her research interests are in STEM education, women in engineering, and machine learning. Dr. Alkhasawneh worked on promoting STEM education through the Virginia-North Carolina Louis Stock Alliance for Minority Participation grant (VA-NC LSAMP) at VCU -Funded by National Science Foundation (NSF).

Dr. Rosalyn Hobson Hargraves holds a joint appointment in the Schools of Education and Engineering as Associate Professor of Teaching and Learning and Associate Professor of Electrical Engineering at Virginia Commonwealth University. She received her B.S., M.S., and Ph.D. degrees in Electrical Engineering from the University of Virginia. Her research interests are in STEM education, biomedical signal and image processing, and machine learning. She has been awarded the Dominion Strong Men & Women Excellence in Leadership Award, Richmond Joint Engineers Council Engineer of the Year, AAAS Diplomacy Fellowship, and the NSBE Janice Lumpkin Educator of the Year Award.