2011

Developing a Hybrid Model to Predict Student First Year Retention and Academic Success in STEM Disciplines Using Neural Networks

Ruba Alkhasawneh
Virginia Commonwealth University

School of Engineering

Virginia Commonwealth University

This is to certify that is the dissertation prepared by Ruba Alkhasawneh entitled DEVELOPING A HYBRID MODEL TO PREDICT STUDENT FIRST YEAR RETENTION AND ACADEMIC SUCCESS IN STEM DISCIPLINES USING NEURAL NETWORKS has been approved by her committee as satisfactory completion of the Thesis Proposal requirement for the degree of Doctor of Philosophy

Rosalyn S. Hobson, Ph.D., Committee Chair, Department of Electrical and Computer Engineering

Stephanie G. Adams, Ph.D., Department of Mechanical Engineering

Lorraine M. Parker, Ph.D., Department of Computer Science

D'Arcy P. Mays, Ph.D., Department of Statistical Sciences and Operations Research

Jenny Jones, Ph.D., Department of Social Sciences

Developing a Hybrid Model to Predict Student First Year Retention and Academic Success in STEM Disciplines Using Neural Networks

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

by
Ruba Alkhasawneh

Director: Rosalyn Hobson
Associate Professor, Electrical and Computer Engineering

Virginia Commonwealth University
Richmond, VA
July, 2011

Table of Contents

List of Figures

List of Tables

Abstract

Developing a Hybrid Model to Predict Student First Year Retention and Academic Success in STEM Disciplines Using Neural Networks

By Ruba Alkhasawneh, PhD

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

Virginia Commonwealth University, 2011

Major Director: Rosalyn Hobson
Associate Professor, Electrical and Computer Engineering

Understanding the reasoning 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 many researchers for decades. Numerous studies have used traditional statistical methods to identify factors that affect and predict student retention. Recently, researchers have relied on using data mining techniques for modeling student retention in higher education [1].

This research has used neural networks for performance modeling in order to obtain an adequate understanding of factors related to first year academic success and retention of URM at Virginia Commonwealth University.

This research used feed forward back-propagation architecture for modeling. The student retention model was developed based on fall to fall retention in STEM majors. The overall freshman year GPA was used to model student academic success. Each model was built in two different ways: the first was built using all available student inputs, and the second using an optimized subset of student inputs. The optimized subset of the most relevant features that comes with the student, such as demographic attributes, high school rank, and SAT test scores was formed using genetic algorithms.

A further step towards understanding the retention of URM groups in STEM fields was taken by conducting a series of focus groups with participants of an intervention program at VCU. Focus groups were designed to elicit responses from participants for identifying factors that affect their retention the most and provide more knowledge about their first year experiences, academically and socially. Results of the genetic algorithm and focus groups were incorporated into building a hybrid model using the most relevant student inputs.

The developed hybrid model is shown to be a valuable tool in analyzing and predicting student academic success and retention. In particular, we have shown that identifying the most relevant student inputs from the student's perspective can be incorporated with quantitative methodologies to build a tool that can be used and interpreted effectively by people who are related to the field of STEM retention and education. Further, the hybrid model performed comparable to the model developed using the optimized set of inputs that resulted from the genetic algorithm. The GPA prediction hybrid model was tested to determine how well it would predict the GPA for

all students, majority students and URM students. The root mean squared error (RMSE) on a 4.0 scale was 0.45 for all students, 0.47 for majority students, and 0.45 for URM students. The hybrid retention model was able to predict student retention correctly for 74% of all students, 79% of majority students and 60% of URM students. The hybrid model's accuracy was increased 3% compared to the model which used the optimized set of inputs.

To my husband, my son, and the family

Acknowledgment

I feel very blessed to have the opportunity to work with and learn from so many incredible individuals during my Ph.D. study. It was a great experience to have been guided and supervised by Dr. Rosalyn Hobson who has offered me boundless support. I have benefited from her encouragement, and leadership. I also want to thank Dr. Stephanie Adams, Dr. Jenny Jones, Dr. Lorraine M. Parker, and Dr. D'Arcy Mays for their valuable feedback on my work. I am greatly thankful for Dr. Kayvan Najarian and Dr. Lisa Abrams for their generous support. I also thank Dr. Alan Sack the Director of Institutional Reporting and Analysis at VCU for his valuable collaboration.

My gratitude goes to my colleagues in the LSAMP program and VCU School of Engineering staff who have helped me to accomplish and provided me with assistance repeatedly. Special thanks to the focus groups participants for their involvement and feedback.

I want to express my great thanks to my husband, Fadi Obeidat, who has backed me up and inspired my thinking in all crucial moments. I could not have accomplished this without your support, patience, and love. I also cannot forget to thank my wonderful son, Majd Obeidat, for being so kind and understanding when I was unable to play all the time. Finally, I thank my parents for their endless support and encouragement during my Ph.D.

Chapter 1: Introduction

1.1 Problem Statement

Increasing student retention and academic success in STEM disciplines have been among the goals of higher education institutions for a long time. Significant efforts have been made to predict student retention in higher education and to understand the process of dropping out of college [2-4] by developing theoretical models of student retention using associated factors. The following studies used traditional methods of statistical analysis to validate these models and investigate student persistence/dropout in higher education [5]. Retention in higher education is defined in [6] as "staying in school until completion of a degree." The study argued that although retention and dropout in higher education are complicated processes, exploring their complexity provides researchers with better knowledge regarding student progress [6]. Seymour [7] reported that both enrollment and retention rates in STEM disciplines have declined. More specifically, Tinto [8] reported that freshmen year has the highest dropout rate especially in the first six weeks of the first semester.

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 [9-11]. These groups tend to enroll in STEM majors in small numbers and leave in higher numbers [12-13]. Tan [14] claimed that "although it is true that freshman STEM majors have indeed grown in numbers in the last decade or so, women and ethnic minorities (with the

exception of Asian Americans) are still underrepresented in STEM disciplines. Compounding the problem are the lower persistence and graduation rates among underrepresented minorities and women."

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 major and receive a bachelors degree in a STEM discipline nationwide [15].

A recent study claimed that the population of white non-Hispanic males will decline by about 11% in the period of 1995 – 2050, while the population of African Americans (AA) and Hispanic Americans (HA) in the workforce will increase by 2% and 14% over the same period, respectively [16]. By 2042, it is predicted that minority groups will be the majority in the US [17-18]. The need to diversify the STEM workforce is of utmost importance, not only because of changing population demographics, but also because workplace diversity has a great impact on increasing worker recruitment, retention, and productivity [19].

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 [20-22]. Retention is of significant interest because of its positive impact on college reputation and workforce demographics [23].

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 [24-25]. Zhang [26] reported 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 [24].

Studies varied in identifying factors that affect student retention the most in their freshmen year. Zhang [26] claimed that high school GPA and placement tests scores, in addition to grades in math, chemistry, and physics, are all strong predictors for engineering student retention. Gaskin [27] determined that pre-defined variables combined with environmental variables, such as living on campus or off-campus, and involvement in first year programs, such as a residential living learning community, are best predictors for student success.

Traditional methods of statistical analysis have been used to predict student retention, such as logistic regression [27]. Recently, research has focused on data mining techniques to study student retention in higher education [1]. 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. From this, it is possible to develop a prediction model, classify or cluster the model, validate it, and implement the developed model.

Data mining research uses several methods to study student retention in the first year in engineering, such as neural networks and structural equation modeling [25-26]. This research has used the neural network technique which is commonly employed for modeling and machine learning. Two models were developed to predict student academic success and retention. Each model used two input sets: the first used all available student inputs and the second used an optimized subset of inputs, which was obtained using genetic algorithms. Moreover, this research used qualitative methods (focus groups) to provide better understanding of first year academic success and retention among minority students. The results of genetic algorithms and qualitative methods were incorporated into modeling freshman year academic success and retention. The 10 fold cross-validation method was used to validate the developed neural networks models. In addition to using qualitative methods to assist in identifying the most relevant student inputs, they were also used to provide an understanding of minority students' freshman year experiences, academically and socially. The neural network technique and genetic algorithm are described in detail in chapter three. To our knowledge, this method is original and has never been developed before.

1.2 Purpose and Research questions

The purpose of this research is to develop a hybrid framework to model first year student academic success and 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. Both models used first-time first year students of 2007-2009 cohorts majoring in STEM. The focus groups participants were former Summer Transition Program (STP) students over a three year period of time, 2008-2010. VCU offered its first STP in summer 2008. The STP is a residential four week program for entering URM freshmen (African American Hispanic American, and Native American) targeting fourteen STEM majors including engineering, natural sciences, and mathematical sciences. More details about the STP and the selection criteria for participants can be found in section 4.5.2.2.

The examined research questions of this dissertation are:

1. Which student inputs impact first year student academic success in STEM disciplines the most?
2. Which student inputs impact first year student retention (from first fall of enrollment to the beginning of the second fall) in STEM disciplines the most?
3. 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 academic success and 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.

1.3 Contribution

This research contributes to the field of engineering by utilizing engineering techniques to develop a tool that is able to predict URM student academic success and retention in STEM disciplines. The developed tool aims to improve URM student freshman year academic success and retention in order to attract talented minds and prepare better engineering workforce. This tool is meant to be used not only by experts in the field of engineering, but also by people who are related to the field of STEM education in general.

This model was built by incorporating quantitative (genetic algorithm) and qualitative (focus groups) results. Further, this research focuses on analyzing freshman year experiences of URM students at VCU in order to build a full image of different dropout/persistence behaviors and their causes.

Obtaining an adequate understanding of URM student retention and academic success and modeling their performance and retention during freshman year, serves institutions

by identifying at-risk students in STEM fields. This study paves the way for advisors and instructors to better advise and direct students to benefit from available resources and assist them to achieve their goals.

Chapter 2: Background and Literature Review

2.1 Introduction

For many years, researchers have tried to understand and model student persistence/dropout in higher education. They have investigated associated factors and models that describe and explain student retention. The most comprehensively studied model was Tinto's theoretical model for student dropouts [2]. This model was followed by multiple studies that used statistical methods to test it. Tinto's and other models are discussed in this section, in addition to studies that are based on his theoretical model and other related works.

2.2 Predictive models of student retention

2.2.1 Tinto's model

Tinto in his model [2] 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 [28] 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, precollege 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. Levels of college commitment and different forms of behaviors were addressed in the study below:

"(a) Students with solid academic competence but moderately low commitment to college completion tended to withdraw voluntarily from college, often to transfer to another institution or reenroll at the same institution at a later date (i.e., stopout). (b) Students with poor academic qualifications but moderately high commitment tended to persist in college until completion or until forced to withdraw for academic reasons (i.e., academic dismissal). (c) Students with both low commitment to college completion and moderately low academic competence tended to withdraw from college and not transfer to another institution or reenroll at a later date (i.e., permanent dropout)."

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 dropout, and how these processes are affecting their academic and social integration [2].

2.2.2 Astin's Input-Environment-Output model

Astin in his book "Assessment for Excellence: The Philosophy and Practice of Assessment and Evaluation" [29] 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 reported that "even if we have good longitudinal input and student output data, our understanding of the educational process will still be limited if we lack information on the college environment." [29]. As an example, the author argued that relying on college GPA to evaluate student success and progress is not enough since it tells us little about the amount of knowledge that students gain during college courses.

Astin defined student inputs as precollege characteristics (i.e., race/ethnicity, gender, and family background), college admission tests and high school GPA, and student self-reported data (i.e., goals and college expectations). He addressed the importance of input data because it influences student output data and most likely influences the educational environment [29]. 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 argued that the lack of involvement in college environment was a significant cause of student withdrawal from college [30]. Educational outcomes refer to the college impact on student.

Both Tinto [2] and Astin [29] highlighted the importance of college experiences in order to understand student retention in higher education. Tinto in his study analyzed extensively the process of student persistence/dropout decisions, while Astin highlighted the importance of educational programs and assessment of practices in studying student success and progress, which are considered major factors influencing student dropout decision.

2.2.3 Terenzini and Pascarella

Terenzini and Pascarella's [5] study was developed based on Tinto's [2] 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.
- Precollege 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.

2.2.4 Other studies

Reason [31] reported that specific student features such as race/ethnicity, GPA, and gender, and institutional features such as selectivity and student integration into academic life are the main factors that affect retention. At Arizona State University, a survey was administered each semester to first year students in the School of Engineering in the Fall of 1995. Employment demands, financial problems, and family issues were reported as the three main causes of student drop out at school of engineering during the semester [32]. Several reasons were correlated with college retention with specific focus on the fields of science and engineering. Related studies [33] pointed to the following factors as significant in affecting student retention: "lack of adequate high school preparation; difficulty in adjusting to college life; lack of engineering community atmosphere; disappointment in not being exposed to engineering related courses and activities during the first two years; and financial."

In [17] the first year was described as a critical period for engineering students when they are not identified as engineers yet. It also reported that the attrition rate for women and minorities in engineering is on average 30% nationally [17]. Tinto in his speech "Taking Student Retention Seriously" believed that there are five conditions that support retention, "namely expectation, advice, support, involvement, and learning." He emphasized that it is also important to continue with students through the academic year to achieve a real impact on student retention [34].

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" [35]. Heywood [36] identified that the first few weeks play a significant role in shaping student motivation and attitude towards college life. He further added that the transition from high school to college is culturally challenging for minority students [36].

Furthermore, the literature review identifies first year college success as a significant impact on student retention [31], [37], [38], [20], [33]. For about two decades, research has shown that student performance and GPA in first and second semesters are crucial predictors of student retention [36], [16],[14].

2.3 Data Mining Models in predicting student retention and academic success

Research has shown that tracking students who transfer from STEM disciplines to a non-STEM disciplines is an increasingly difficult process [39]. 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 [26, 40]. As described in section 2.1, studies have varied in identifying factors that affect student retention the most, especially in their freshmen year. In [39], it was 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.

Gaskin [27] 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) at Bowling Green State University (BGSU) 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 freshmen year.

Besides traditional statistical analysis methods, data mining methods are becoming more popular and accurate in modeling student retention. 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 [25], were used. Each model used several precollege 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 its superiority among the other

four methods used in terms of prediction accuracy. A similar study that used a database of 39,277 engineering students from 9 different institutions found that high school GPA and standardized test scores were significant predictors of engineering freshmen retention [26]. The study also added ethnicity, gender and citizenship as influential factors but they were inconsistent among all included institutions.

Herzog [24] conducted two studies; one focused on studying student retention, which used forty variables, and the other focused on time to degree, which used seventy nine 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.

In Thailand, researchers were interested in applying data mining methods for predicting student performance as well [41]. Their research compared the accuracy of Decision Tree and Bayesian Network algorithms for predicting both undergraduate and graduate student academic performance at two different institutes. In the first institution, Can Tho University (CTU) in Viet-nam, the study used records and GPA of 20,492 students

admitted from 1995 to 2002 to predict their performance in the third year based upon their second year performance. At the second institution, the Asian Institute of Technology (AIT) in Thailand, data of 936 students was used to predict first year academic performance (GPA). The study showed the superiority of 3-class decision tree method with an overall prediction accuracy of 86% (CTU) and 74% (AIT).

In [39], data of 1,884 STEM freshmen who enrolled at ASU in the 1999-2000 academic year was collected. The study focused on 6 out of 18 available variables, which were: gender, ethnicity, citizenship, high school GPA, SAT- quantitative, and SAT- verbal. Classification trees and random forests were leading methodologies in studying STEM student retention compared to traditional statistical methods. In another study, European researchers were interested in identifying "at-risk" students before the freshmen year examination session started [42]. The study used 533 students registered in Belgian universities during the academic year 2003-2004. It classified students into three categories: low-risk, medium risk, and high-risk using several data mining methods such as neural networks, random forests and decision trees. The study found that 60% of its students dropped out of Belgian universities and discriminant analysis methodology performance was slightly better than the other two methods [42].

In the electrical engineering department of Eindhoven University of Technology, a study was conducted to identify factors that affect electrical engineering student retention. Data of all students who were enrolled in electrical engineering over the period 2000 – 2009

was collected. Several data mining methods were used such as decision trees, random forests, and Bayesian classifiers. That data set containing 648 students showed that simple and intuitive decision tree classifiers were the best methods for prediction with accuracy between 75% and 80% [40].

A study that included 48 students who were enrolled in a minority engineering program at the University of Akron investigated the significance of high school GPA and ACT score as predictors of minority student success in engineering programs [43]. This study used correlation and multiple linear regression. High school GPA and ACT scores were found to be correlated and high school GPA was a significant predictor of minority engineering student success.

2.4 Conclusion

In the past, researchers attempted to develop comprehensive theoretical models to analyze and predict student retention in higher education. Most of these models focused on the importance of precollege factors and academic and social integration in college life in impacting student persistence in or dropout from college. Further steps have been taken to validate these theoretical models, and advance research on student attrition using statistical analysis methods. Recently, data mining methods have proven to provide robust models that accurately predict student retention. This research uses the strength of data mining by incorporating its results with qualitative methodologies results to build and validate an effective framework to model freshman year student academic success

and retention. Further, this research provides an insight into freshman year experiences of URM students in STEM disciplines.

Chapter 3: Methodology Background Theory

3.1 Introduction

This research investigated the use of neural networks as a tool to model first year student academic success and retention. Neural networks have been shown to handle complex data sets and in some cases have performed better than traditional statistical methods. Besides the neural networks model, genetic algorithm has been used for feature subset selection to identify the most relevant factors in each developed model. Results obtained were incorporated with the qualitative methodology afterwards to build a comprehensive model that has better performance and better interpretability by end-users. In this chapter a background on neural networks, genetic algorithm, and qualitative research methods is presented.

3.2 Neural Networks

Neural networks are a mechanism that mimics the human brain's biological process of learning. Neural networks, first introduced in 1943 by Warren S. McCulloch and Walter Pitts [44], are a parallel processing computing technology comprised of interconnected processing elements or "neurons" that interact with each other mathematically to learn from the external environment based upon inputs to and outputs from the system. Neural networks have been applied in a variety of areas such as business, manufacturing, biology, engineering, and education. Although there are still some arguments that neural

network computations are an extension of regression analysis methodology, it is a proven technology for classification and prediction [45].

3.2.1 Neural network architecture

Neural networks are a parallel processing mechanism formed of multiple layers of processing element(s) or neurons. There are three classifications of architectures: feedforward, recurrent, and topological maps. This research will utilize the feedforward architecture which consists of an input layer, one or more hidden layers(s), and an output layer. The layers are made up of simple processing elements. The k^{th} processing element, shown in figure 1, consists of p input signals/values, x_p, each of which is multiplied by a synaptic weight/value, w_{kp}. These values are all summed together over j resulting in an output, $v_{k,}$ which serves as an input to an activation function, φ_k, which generates the output, y_k, for the processing elements. A common function in neural networks applications is the sigmoid function $(1/(1+e^{-1)})$. The output of the networks is calculated based on the following equations, where b_k represents the biases b_k represents the biases.

$$v_k(i) = \sum w_k(j)X(i,j) + b_k \quad (1)$$

$$y_k(i) = \Phi(v_k(i)) \quad (2)$$

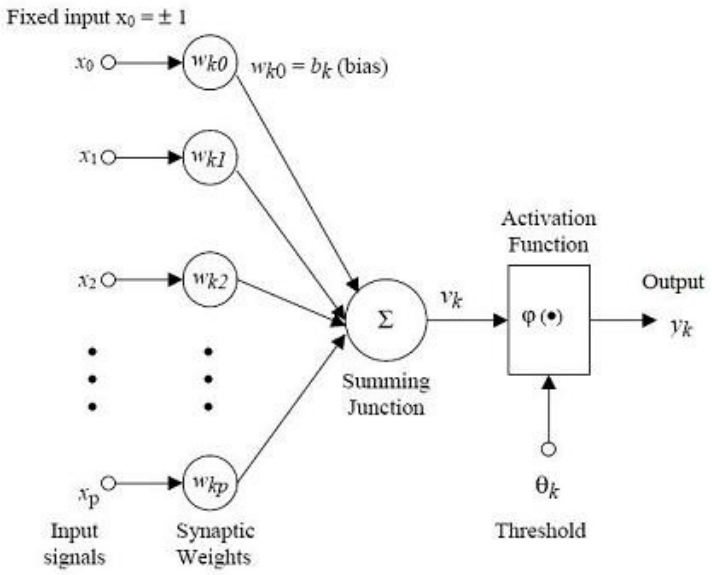

Figure 1 Neural Networks Learning Process

3.2.2 Feedforward backpropagation learning process

The neural networks learning process occurs when acquiring knowledge by the network. Synaptic weights, $w_{kp,}$ are used to store knowledge by the continuously iterating data and updating these weights to make predictions. Weights are initialized randomly in MATLAB and updated based on the type of learning (adaptation). The learning process is classified as either 1) supervised learning which has a desired output for every input, or 2) unsupervised learning where the training data has only inputs and the network learns via experience while training data. In this research, supervised learning is used.

The training algorithm adjusts the weights to minimize the difference between the desired output, $d_i(n)$, and the network output, $y_i(n)$, by calculating the error signal, e(n), as in equation 3. The Levenberg-Marquardt training algorithm is used in this research.

$$e_i(n) = d_i(n) - y_i(n) \qquad (3)$$

Where 'i' is the iteration and '*n*' is the number of inputs.

The weights vector is updated based upon the actual response and the desired response using the equation below.

$$w(n+1) = w(n) + \eta\,[d(n) - y(n)]\,x(n) \qquad (4)$$

Where

η – learning rate parameter, $0< \eta<1$

$w(n)$ – weight vector at time n or current weights.

x(n) – input vector at time n or current inputs.

$w(n+1)$ – vector of the new weights.

The backpropagation learning is based on an error-correction rule. Inputs are applied to the network and the output of each layer is calculated and passed forward to the following layer until the actual network output is calculated in the final layer. The error signal is calculated and then propagated back through the network and weights are updated until the minimum error is reached.

3.3 Genetic algorithm for optimization

Genetic algorithm is a powerful evolutionary computing technique which is widely used for optimization processes. Genetic algorithms are used in modeling to improve accuracy and performance of the developed model by selecting a subset of the most relevant input variables. Figure 2 represents the genetic algorithm implementation flow chart.

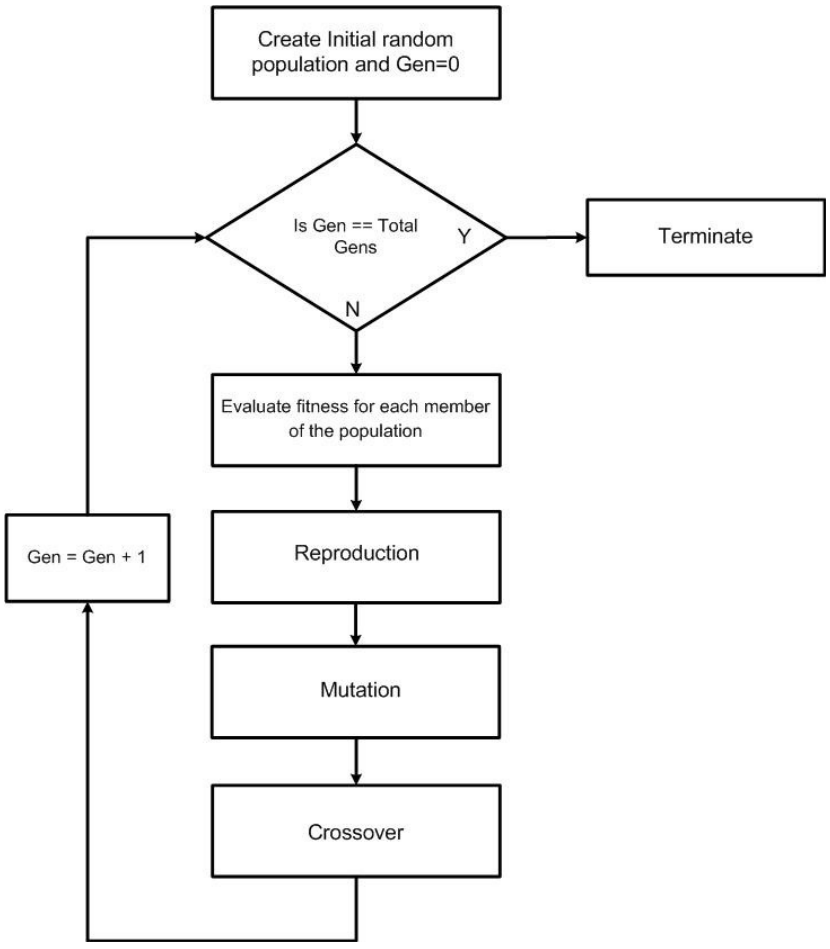

Figure 2 GA implementation flow chart

The algorithm consists of chromosomes which represent a vector of weights for each student input. In reality, they store genetic information that determines the specific characteristics of each organism; and the length of each vector is equal to the number of features. A random population of weight vectors are initialized and passed from generation to generation by selecting two parent chromosomes depending on their fitness. Then, the weight vectors go through crossover and mutation phases. In the mutation phase one or more bit(s) are randomly selected to be inverted in the chromosome. As an example, if we have the following binary number: 01001**1**01 the resulting string will be 01001**0**01. This step is to guarantee that variety of resulting features is achieved. The crossover phase is basically swiping features between selected parents to generate two new offspring with good features retained. The crossover process is shown in figure 3.

Parent X	**1010**	**11001100**
Parent Y	**1101**	**01101100**
Offspring 1	**1010**	**01101100**
Offspring2	**1101**	**11001100**

Figure 3 Crossover Phase

Finally, each weight vector represents a candidate subset of features. The selection process creates random combinations of input vectors and then each solution is evaluated by a fitness function. Vectors with good grades are passed from generation to generation until the optimal solution is obtained. The output obtained by the genetic algorithm is a binary vector with best subset of features. 1 represents that the feature was selected and 0 represents its absence.

3.4 Qualitative Research Design

Qualitative research methodologies are effective way in terms of analyzing non-quantitative data or data in the form of text rather than numbers. Researchers defined qualitative research as an " important modes of inquiry for the social sciences and applied fields, such as education, regional planning, health sciences, social work, community development, and management." [46]. The strength of qualitative research comes from three main points 1) "exploration and discovery" in which it aims to learn about a specific group of people, 2) "context and depth" by providing an insight into people's behaviors and experiences, 3) "interpretation" where it gives an understanding of the reasoning behind people's behaviors [47]. Qualitative research includes several strategies for data collection such as observations, content analysis of existing sources, interviews, and focus groups. This study will employ focus groups as a strategy for collecting data to gain insights into the STEM students' experience at VCU.

3.4.1 Focus groups

The focus group technique is becoming widely used as a "face-to-face interview" with a group of individuals to evaluate programs. This method encourages participants to give their responses regarding their freshmen year college experience and identify factors that could affect their academic success and integration in college life. Denton and McDonagh [48] define focus groups as follows:

"Focus group is an umbrella term. It centres on a gathering of target users brought together for a relatively informal discussion on a specific topic or issue. A

chairperson (moderator), using a flexible schedule of questions (the moderator's draft), promotes discussion, while carefully ensuring not to direct, but guide the group through issues which emerge as important to them. A variety of techniques can be used to promote discussion."

Literature shows an interest in focus group methodology has increased over the years due to its efficacy in collecting participant opinions and comments better than any other data collection method. Focus groups have been used not only to evaluate programs, but also to identify participants characteristics about a particular issue/concern [49]. For the purpose of this study, focus groups are being used to identify participant characteristics that may prevent him/her from continuing in a STEM discipline.

In [49] it was reported that the major factors that affect student success are: weak mathematics preparation in high school due to poor instruction; valuable study skills, such as critical thinking, and talking about what to expect in college were poorly addressed by high school teachers. Besides the poor preparation in high school, first year inexperienced mathematics teachers are strong impediments to freshmen persistence in STEM majors [49].

The focus groups technique used in this research to get a deep insight on major academic and environmental factors that impact URM student accomplishments the most and elicit responses regarding their freshman year experiences.

Chapter 4: Methodology

4.1 Introduction

This study develops a hybrid model by identifying the most significant student inputs that affect freshmen academic success and retention in STEM disciplines while focusing on URM groups at VCU. Retention in this study 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 and in which it focuses on fall to fall retention at VCU; all students included in this study were enrolled in the fall semester of their sophomore year. The model uses precollege and college characteristics such as admission test scores, high school percentile rank, number of attempted or earned credit hours, and demographic attributes of students to identify significant factors that impact the decision of persistence/dropout from a STEM discipline. Identifying significant factors that affect student academic success and retention in STEM disciplines is performed in two phases as described later in section 4.5. As an extension of Tinto's student dropout model [2], this research analyzes freshmen year experiences of URM groups and pre-freshman and freshman year factors that influence the re-defining of student goals and intention to graduate with a STEM degree. The population of this study is incoming freshmen at VCU in STEM disciplines, and the sub-group is URM students who are African American, Hispanic American, and Native American freshmen. The

VCU 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. This chapter is organized as follows: section 4.2 introduces the research questions. Section 4.3, population and sample sizes of both the quantitative and qualitative methods. In addition, discussion of the study's data collection and major variables used with a detailed analysis is included in section 4.4. Finally, a detailed description of the research design of the neural network models, genetic algorithm, focus group procedures, and the hybrid model is provided in section 4.5.

4.2 Research Questions

The developed framework used two input sets: the first used all available student features and the second used an optimized set of the most relevant factors. The examined research questions are:

1. Which student inputs impact first year student academic success in STEM disciplines the most?
2. Which student inputs impact first year student retention in STEM disciplines the most?

3. 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 the best inputs that contribute to student academic success and retention provides significant information for institutions to know what student needs are, how to support student academic success, and increase retention in STEM fields. Institutions can also rely on examining freshman year experiences to build a solid base of knowledge on different student retention behaviors from a diverse population. Based on this knowledge, better programs and student services can be developed for students.

4.3 Population and Sample

VCU is a large public research institution located in Richmond, Virginia. More than 32,000 students enroll at VCU. In fall 2006, University College was established to enhance the undergraduate student college experience especially in their freshmen year. A set of services, learning opportunities, and programs are offered to undergraduate students such as academic advising, tutoring, orientation, and group studies so as to motivate them to achieve higher levels of academic success. The goal of University College is to enhance the quality of undergraduate education at VCU by encouraging integration into college life and getting students involved in their own freshmen year experiences. Supplemental Instruction (SI) is one of the most significant services available for VCU students. SI is a peer-assisted study session which was designed to

assist students in courses that had proven to be difficult, and it is open to all students in these classes. SI sessions are conducted each week by students who have previously taken the courses, and currently attending the same class, taking notes, and reading the text.

Participants of this study fall into two groups:

1) The first group comprised of STEM fulltime first year students from the 2007-2009 academic years. Data was obtained from the VCU office of Institutional Research. The sample size consisted of 1966 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. At VCU, 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, majority student cohort that includes a total of 1468 students, and second, 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 the VCU LSAMP summer transition program 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' precollege 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 of whom two students were non-STP participants. These two students responded to an invitation for non-participants to get an insight into other freshman year experiences for students who did not have a chance to participate in the program. Participants' demographic and other characteristics are described in section 5.3.

This particular group was included because they were exposed to a variety of activities and programs prior to and during their freshmen 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 VCU population.

4.4 Data Collection

Literature has focused on the importance of precollege variables in impacting student retention in higher education, associated with college academic and social experience [2,

5, 29]. Pascarella and Terenzini in their study validating Tinto's model concluded that precollege factors were important in student persistence/dropout in a way how they interact with college experiences. This study will develop two hybrid models that predict URM student academic success and retention. The models incorporated both relevant factors that are determined using genetic algorithms and qualitative method via focus groups conducted to understand student first year experience.

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, precollege, and college variables. Table 1 includes a detailed description of student input variables and response variables. The demographic variables included in this study are: race/ethnicity, residency, and gender. The precollege variables are: honors, SATM, SATV, SATC, high school percentile rank, and first math course. The college variables are: term credits attempted in the first fall, term credits earned in the first fall, credits attempted in the first fall, credits earned in the first fall (this variable gives an indication of the student transfer credits, if any), term credits attempted in the first spring, term credits earned in the first spring, credits attempted in the first spring, credits earned in the first spring, first mathematics course grade, fall term GPA, and spring term GPA. 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.

Table 1 Summary of Variables

Variable	Abbreviation	Description
Race/Ethnicity	-	URM/Majority
Residency	-	In-state/Out-of-state
Gender	-	Male/Female
Honors	-	Student accepted into the Honors College (Yes/No)
Math SAT score	SATM	Math standardized test score
Verbal SAT score	SATV	Verbal standardized test score
Combined SAT score	SATC	Combined standardized test score
Percentile Rank	Rank	Student actual high school percentile rank (%)
Math course1	CourseM	Student's first math course (gives an indication of student's math placement test score and AP credits) – Algebra, Pre-calculus, Calculus I, Calculus II, Differential equations, and Other math courses

Term credits attempted in fall1	TCAF	Number of college credits student attempted to take in the first fall in which student enrolled
Term credits earned in fall1	TCEF	Number of college credits student earned by the end of first fall in which student enrolled
Credits attempted in fall1	CAF	Total number of college credits student attempted to take in the first fall in which student enrolled
Credits earned in fall1	CEF	Total number of college credits student earned in the first fall in which student enrolled
Term credits attempted in spring1	TCAS	Number of college credits student attempted to take in the first spring in which student enrolled
Term credits earned in spring1	TCES	Number of college credits student earned in the first spring in which student enrolled
Credits attempted in spring1	CAS	Total number of college credits student attempted to take in the first spring in which student enrolled
Credits earned in spring1	CES	Total number of college

		credits student earned since the first spring in which student enrolled
Math grade1	GradeM	Grade of the first math course that student took
Term GPA in fall	TGPAF	Fall semester GPA (out of 4.0)
Term GPA in spring	TGPAS	Spring semester GPA (out of 4.0)
GPA	-	Overall cumulative GPA of freshman year (out of 4.0)
Retention	-	Fall to fall retention in a STEM discipline

2) The qualitative data was obtained by conducting focus groups for VCU LSAMP summer transition program 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 (VCU IRB#: HM12908). A copy of the IRB Approval form can be found in Appendix A. Sixteen students participated in the three sessions: 9 in the first session, 5 in the second session, and 2 in the third session. More details are included in section 5.3.

4.4.1 Student Features Analysis

In this study of 1966 students it was observed that females represent higher a proportion 0.52 (1019 samples) compared to males (see table 2). The proportion of underrepresented URM students is 0.25 (498 samples). Most of the students in the samples were in-state residents with a proportion of 0.93 (1820 samples).

Table 2 All students' demographic variables

Variable	Level	Overall N (%)
Gender		
	Male	947(48%)
	Female	1019(52%)
Race		
	Majority	1468(75%)
	URM	498(25%)
Residency		
	In-State	1820(93%)
	Out-of-State	146(7%)

Further, it was observed that most of the students were not honors students, with a proportion of 0.9 (1775 samples), see table 3. The average SAT score was approximately 1124 and the average high school rank was 77% which is considered good since the VCU average is in the top 75%. The proportion of students in pre-calculus was 0.3 (600 samples) and the proportion of students who received an A in their first math course was 0.37 (732 samples). On average 13 college credits were earned per semester. The average overall GPA is 2.86. In order to be consistent with the focus groups sample, the mathematics courses were classified into six categories as shown in table 3. The "other"

category includes Introduction to Contemporary Mathematics, Mathematical Structures, Mathematics in Civilization, Introduction to Computational Mathematics, Introduction to Mathematical Reasoning, Multivariate Calculus, and Linear Algebra.

Table 3 All students precollege & college variables

Variable	Level	Overall N (%)
Honors		
	Yes	191(10%)
	No	1775(90%)
Course		
	Algebra	446(22.7%)
	Pre-calculus	600(30.5%)
	Calculus I	501(25.5%)
	Calculus II	157(7.98%)
	Differential Equations	23(1.16%)
	Other	239(12.16%)
Grade		
	A	732(37.2%)
	B	556(28.3%)
	C	358(18.2%)
	D	122(6.2%)
	F	71(3.61%)
	W	127(6.4%)

4.4.1.1 Majority Student Features Analysis

Regarding the 1468 majority students, it was observed that males represent higher proportion of 0.55 (809 samples) as compared to females with a proportion. Most of the students were in-state residents with a proportion of 0.94 (1386 samples) as shown in table 4.

Table 4 Majority students' demographic variables

Variable	Level	Overall N (%)
Gender		
	Male	809(55.1%)
	Female	659(44.9%)
Residency		
	In-State	1386(94.4%)
	Out-of-State	82(5.6%)

The proportion of honors students was 0.13 (186 samples); the average SAT score was approximately 1158 and the average high school rank was 77%. The majority of students, 0.57 (840 samples), were placed into pre-calculus or calculus I. In addition, the proportion of students who received an A in their first math course is 0.4 (591 samples) as shown in table 5. The average college credits earned by the end of freshman year were 38 hours while the average college credits earned per semester were 14 credit hours. The difference between the actual earned credits and the expected credits earned, gives an implication of the total transfer credits a student earn prior college starts. The average overall GPA was 2.93.

Table 5 Majority students precollege & college variables

Variable	Level	Overall N (%)
Honors		
	Yes	186(13%)
	No	1282(87%)
Course		
	Algebra	251(17.1%)
	Pre-calculus	421(28.7%)
	Calculus I	419(28.6%)
	Calculus II	150(10.2%)
	Differential Equations	21(1.4%)
	Other	206(14%)
Grade		
	A	591(40.3%)
	B	417(28.4%)
	C	239(16.3%)
	D	81(5.5%)
	F	45(3%)
	W	95(6.5%)

4.4.1.2 URM Student Features Analysis

A total of 498 URM students were included. It was observed from the demographics characteristics that females represented a higher proportion, 0.72 (360 samples), as shown in table 6. It was also observed that the majority male percentage is 27.1% higher compared to URM male percentage (see tables 5 and 6). The proportion of 0.13 (64 samples) students were out-of-state residents.

Table 6 URM students' demographic variables

Variable	Level	Overall N (%)
Gender		
	Male	138(28%)
	Female	360(72%)
Residency		
	In-State	434(87%)
	Out-of-State	64(13%)

A proportion of 0.99 (493 samples) were not honors students. The average SAT score was 1020 and the average high school rank was 78% and the total proportion of students in either algebra or pre-calculus was 0.75 (374 samples). Also, the proportion of students who received an A in their first math course was 0.28 (141 samples) as shown in table 7. The average was 13 college credits per semester, and a total of 32 college credits by the end of the freshman year. The average overall GPA was 2.7.

Table 7 URM students precollege & college variables

Variable	Level	Overall N (%)
Honors		
	Yes	5(1%)
	No	493(99%)
Course		
	Algebra	195(39.2%)
	Pre-calculus	179(35.9%)
	Calculus I	82(16.5%)
	Calculus II	7(1.4%)
	Differential Equations	2(0.4%)
	Other	33(6.6%)
Grade		
	A	141(28.3%)
	B	139(28%)
	C	119(24%)
	D	40(8.1%)
	F	26(5.2%)
	W	32(6.4%)

4.4.2 Student retention analysis

As shown in table 8, although females represented a higher percentage in the student population, their retention rate was 10% less than that of males. There was no difference in retention rate between in-state and out-of-state residents. 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.

Table 8 Summary of Student Retention by Factor

Variable	Level	Retained N (%)	Not-Retained N (%)
Gender			
	Male	818(86%)	129(14%)
	Female	774(76%)	245(24%)
Race			
	Majority	1229(84%)	239(16%)
	URM	363(73%)	153(27%)
Residency			
	In-State	1474(81%)	346(19%)
	Out-of-State	118(81%)	28(19%)
Honors			
	Yes	1414(80%)	361(20%)
	No	178(93%)	13(7%)
Course			
	Algebra	296(66%)	150(34%)
	Pre-calculus	472(79%)	128(21%)
	Calculus I	464(93%)	37(7%)
	Calculus II	153(97%)	4(3%)
	Differential Equations	22(96%)	1(4%)
	Other	185(77%)	54(23%)
Grade			
	A	626(86%)	106(14%)
	B	448(81%)	108(19%)
	C	286(80%)	72(20%)
	D	96(79%)	26(21%)
	F	52(73%)	19(27%)
	W	84(66%)	43(34%)

For the majority student group, student retention rate increased as their level of mathematics increased; and students who earned an A had the best retention rate of 87% (516 samples) while students who earned B, C, and D had almost the same retention rate as shown in table 9. The average SAT score was 1168 and 1109 for retained and non-retained students, respectively. The average high school rank was 78% for retained students and 75% for non-retained. The overall freshman year GPA was 3.05 for retained students and 2.8 for non-retained. The average total college credits earned by the end of

the freshman year were 39 for retained students and 34 for non-retained students. For overall, majority and URM, the retention rate was lowest for those who withdraw from first mathematics course. Even lower than those who failed.

Table 9 Summary of Majority Student Retention by Factor

Variable	Level	Retained N (%)	Not-Retained
Gender			
	Male	706(87%)	103(13%)
	Female	523(79%)	136(21%)
Residency			
	In-State	1154(83%)	232(17%)
	Out-of-State	75(91%)	7(9%)
Honors			
	Yes	173(93%)	13(7%)
	No	1056(82%)	226(18%)
Course			
	Algebra	175(70%)	76(30%)
	Pre-calculus	334(79%)	87(21%)
	Calculus I	389(93%)	30(7%)
	Calculus II	146(97%)	4(3%)
	Differential Equations	20(95%)	1(5%)
	Other	165(80%)	41(20%)
Grade			
	A	516(87%)	75(13%)
	B	348(83%)	69(17%)
	C	196(82%)	43(18%)
	D	67(83%)	14(17%)
	F	36(80%)	9(20%)
	W	66(69%)	29(31%)

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 as shown in table 10.

Table 10 Summary of URM Student Retention by Factor

Variable	Level	Retained N (%)	Not-Retained
Gender			
	Male	112(81%)	26(19%)
	Female	251(70%)	109(30%)
Residency			
	In-State	320(74%)	114(26%)
	Out-of-State	43(67%)	21(33%)
Honors			
	Yes	5(100%)	0(0%)
	No	358(73%)	135(27%)
Course			
	Algebra	121(62%)	74(38%)
	Pre-calculus	138(77%)	41(23%)
	Calculus I	75(91%)	7(9%)
	Calculus II	7(100%)	0(0%)
	Differential Equations	2(100%)	0(0%)
	Other	20(61%)	13(39%)
Grade			
	A	110(78%)	31(21%)
	B	100(72%)	39(28%)
	C	90(76%)	29(24%)
	D	29(72.5%)	11(27.5%)
	F	16(62%)	10(38%)
	W	18(56%)	14(44%)

4.5 Focus group instrumentation

The focus group protocol was designed for this study to elicit responses from participants about their freshmen year college experiences and determine which variables have the most impact on student academic success and retention (see Appendix B). 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.

Three focus groups were conducted with a total of sixteen participants; two of them were not former STP participants. The first group had nine participants, the second group had five, and the third group had two. The duration of each meeting ranged between 20-50 minutes based on the number of participants. All sessions were tape recorded (audio only) and later transcribed. In qualitative research, the richness and quality of collected data is not dependent on the sample size. Thus, a total of 16 out of 63 participants considered enough to reach a sufficient depth of information regarding the purpose of conducting focus groups. Prior to conducting each session, a demographic survey was

administered to each participant in order to get an insight into participants' diverse backgrounds.

Data Analysis:

The analysis approach used is content analysis which is a very effective method in analyzing data in textual context. This approach is used to describe, analyze, and summarize patterns and trends observed from the collected data [50]. It also analyzes what do participants talk about the most and how trends are related to each other. Trends and patterns were analyzed within and among groups.

4.5.1 The model test group selection criteria

A total of sixty-three former VCU STP participants were included in this study. VCU offered its first STP in summer 2008. The four week residential program's participants were incoming URM in STEM disciplines. The goal of the program is to ensure first year academic success.

The VCU STP was funded by the National Science Foundation (NSF) as part of the VA-NC LSAMP. The program focused on developing essential skills such as communication skills and critical thinking, enhancing mathematics and science study skills, and facilitating a smooth transition to the university community.

This group of students was selected for the study due to its diverse representation of the VCU population in STEM fields. Students were primarily selected to participate in the program based on their high school GPA, SAT test scores, math placement test scores, gender, race/ethnicity, and intended major. Academic and demographic variation among selected groups was obtained each year. Based on that, STP participants are considered a valuable data source for this model. Students are diverse, had a precollege experience, and were exposed to various services and activities during freshmen year. This rich college experience will provide the study with a better understanding of freshmen college experience and factors that impact their retention.

4.6 Research Design

4.6.1 Neural network models design

The feedforwrd backpropagation network used to model first year student academic success and retention at VCU in STEM disciplines. 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. Section 4.4 includes student inputs used in each model. 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 two comprehensive models that are able to predict URM students' first year academic success and 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.

4.6.1.2 Neural network framework performance

Two response variables were used in this study, overall first year GPA for the academic success model, and retention. GPA is a numeric variable ranging from 0 – 4 while retention is a categorical variable of two values, retained or not. To compare prediction models, several error measurements could be used such as mean square error (MSE), mean absolute error, mean absolute percentage error, mean error, and mean percentage error. For mathematical convenient the root mean square error (RMSE) is used instead of

the MSE to compare the academic success model's performance. The RMSE is measured with the same units as the data and represents the size of the typical error. The RMSE is the square root of the average of the total squared error between the predicted and actual values as in the following equation:

$$RMSE = ([\Sigma(\hat{Y}i - Yi)^2]/n)^{1/2} \qquad (5)$$

where $\hat{Y}i$ and Yi are the predicted and actual values, and n is the total number of records. Small RMSE values give an indication of good prediction of the actual values. Generally, if there was no significant difference between the compared models, the simpler and easier model to interpret is preferred

Mean error, maximum error, minimum error, standard deviation, and the GPA error were calculated as well to give a better indication of the GPA model's performance.

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:

$$ACC= (TP+TN)/N \qquad (6)$$

The Receiver Operating Characteristics (ROC) curve was also reported for the retention model. The ROC curve is a plot of true positive rate (TP divided by the total number of retained students) vs. false positive rate (number of incorrectly predicted retained students divided by N). It describes the relationship between correctly predicted and

incorrectly predicted retained students. If the curve is following the left axis and the top of the plot, the model is a good predictor. 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.

4.6.2 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 algorithm (GA) technique gives good results for feature selection [51].

The feature subset selection was used to provide a deep insight into freshmen academic success and 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 genetic algorithm implementation is described in section 3.3. 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 was used and the population (chromosome) size is chosen to be 20. The algorithm accepts a vector of student inputs and returns a bit string that indicates whether the feature was selected or not. 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.

4.6.3 Hybrid Model Design

The hybrid framework is developed to model first year student academic success and retention for URM. This model used results obtained from the quantitative methods (genetic algorithms) and qualitative methods (focus groups) were incorporated to develop two comprehensive models that are able to predict URM students' first year academic success and retention accurately. 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 feedforwrd 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 (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 10 fold cross-validation used to validate the neural networks models.

The accuracy of the developed models was measured using RMSE to compare the GPA model's performance with the GPA model that used only the genetic algorithms to generate an optimized set of student features. Mean error, maximum error, minimum error, standard deviation, and the GPA error were calculated as well to give a better indication of the GPA model's performance. For the retention model the ACC and the

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

Chapter 5: Results

5.1 Introduction

This chapter is organized as follows: sections. Section 5.2 contains results of the student academic success model for three different datasets: all students regardless of their ethnic origin, majority students, and URM students. Section 5.3 presents results of the retention model using the same three datasets used for the academic success model. Next, section 5.4 contains the results of the student academic success model using an optimum student features for the three ethnic datasets included in the previous models. Section 5.5 has the results of the student retention model using an optimum student features for the three ethnic groups as well. The final section (5.6) has the results obtained from the qualitative methodology.

5.2 Student Academic Success Model

This section describes the results obtained by using a neural network to model student academic success for three dataset cohorts- all students as a general model, majority students, and URM student cohorts. These observations are explained in detail in the sections that follows. 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.

5.2.1 Model Performance Results– All students

In the overall GPA model using all features, the RMSE value is 0.45 (on scale of 4.0), as shown in table 11. In analyzing the model's accuracy, it was revealed that the mean error for predicting the overall freshman year GPA was approximately 0.35. Also, it was noticed that the model which used all student inputs was able to predict the GPA of 76% of students within an error of less than or equal to ± 0.5 on scale of 4.0 (which is equivalent to an absolute error of less than or equal to12.5%). Table 11 also shows that only 3.2% of students had a prediction error of greater than ± 1.0 (an absolute error of greater than 25%) for the model that used all student features

Table 11 Summary results of the GPA absolute error analysis

Variable	Output
RMSE	0.45
Max	2.875
Min	0.0007
Mean	0. 35
Std	0. 28
GPAerr≤0.25	44.9%
0.25<GPAerr≤0.50	31.1%
0.50<GPAerr≤0.75	15.7%
0.75<GPAerr≤1.00	5.1%
GPAerr>1.00	3.2%

5.2.2 Model Performance Results- Majority Students

As shown in table 12, the RMSE value is 0.47. The model's accuracy analysis showed that the mean error for predicting the overall GPA is approximately 0.37 and the model was able to predict the GPA of 72.7% of students within an error of less than or equal to ± 0.5 on scale of 4.0 (an absolute error of less than or equal to12.5%). In addition, it was noticed that only 2.4% of students had a prediction range of error of greater than ± 1.0 (an absolute error of greater than 25%).

Table 12 Summary results of the GPA absolute error analysis for majority students

Variable	Output
RMSE	0.47
Max	1.687
Min	0.0045
Mean	0.37
std	0.28
GPAerr≤ 0.25	39.6%
0.25< GPAerr≤0.50	33.1%
0.50< GPAerr≤0.75	16.9%
0.75< GPAerr≤1.00	8%
GPAerr>1.00	2.4%

5.2.3 Model Performance- URM Students

The RMSE for the URM student model was 0.45 and the mean error for predicting the overall GPA is approximately 0.35. It was indicated that 76.6% of students had an error of less than or equal ± 0.5 (an absolute error of less than or equal to12.5%). Further, it was noticed that 3.2% of students had a prediction range of error of greater than ± 1.0 (an absolute error of greater than 25%). The summary results are shown in table 13.

Table 13 Summary results of the absolute GPA error analysis for URM students

Variable	Output
RMSE	0.45
Max	3.143
Min	0.0000
Mean	0.35
std	0.29
GPAerr≤0.25	44.6%
0.25< GPAerr≤0.50	32%
0.50< GPAerr≤0.75	15.2%
0.75< GPAerr≤1.00	5%
GPAerr>1.00	3.2%

The neural networks models for all students cohort and URM students performance were similar at predicting the freshman year overall GPA with a 0.45 value of the RMSE. Both models performed slightly better when compared with the majority student model which had a RMSE value of 0.47. Overall, the performance of all models shows that the networks are very good at predicting the absolute freshman year overall GPA.

5.3 Retention Model

As in the student academic success model, this section describes the results obtained by using retention model using three dataset cohorts- all students, majority and URM cohorts. Results obtained are explained in the following sections. 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.

5.3.1 Model Performance Results– All students

Of the 1966 samples, 74% cases were predicted correctly which considered good prediction accuracy. Figure 4 represents the ROC curve of the developed model.

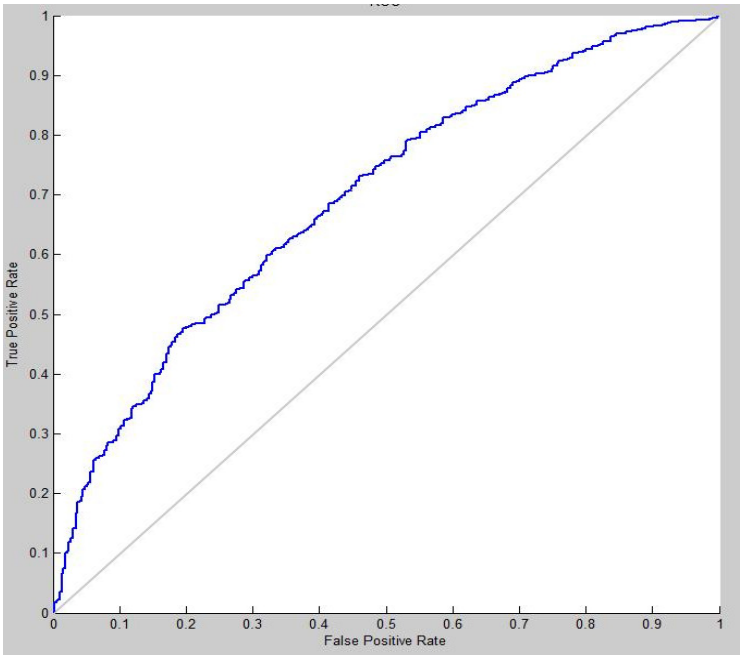

Figure 4 ROC curve of all students using all inputs

5.3.2 Model Performance Results- Majority Students

The majority student model's accuracy was 79% which also considered a good performance in predicting majority student retention in STEM disciplines. The ROC curve is shown in figure 5.

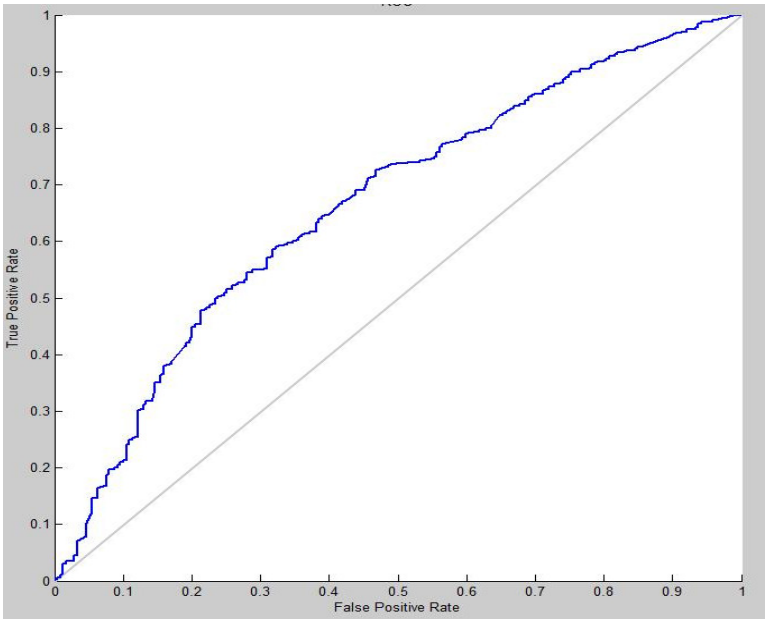

Figure 5 ROC curve of majority students using all inputs

5.3.3 Model Performance Results- URM Students

As for the URM student retention model, the model performed not as good as the previous models with an accuracy of 60%. Figure 6 shows the ROC curve for the model.

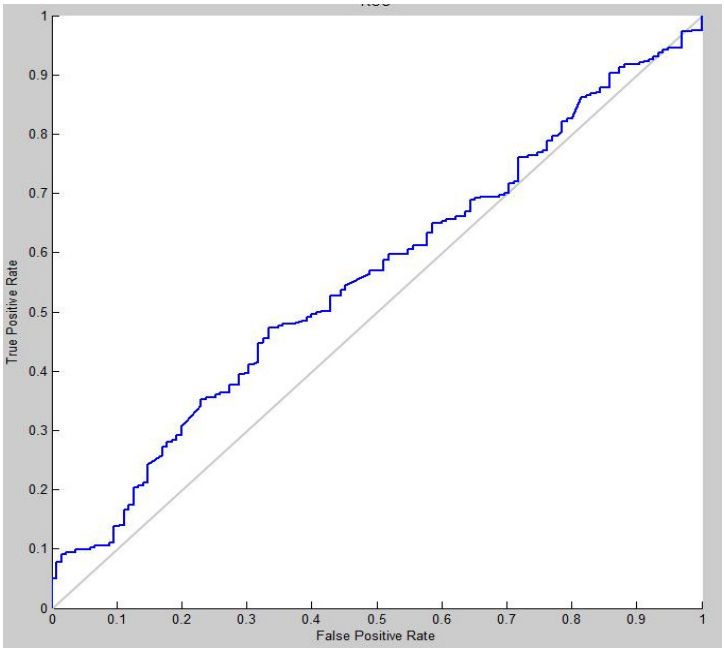

Figure 6 ROC curve of URM students using all inputs

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 good 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.

5.4 Student Feature Optimization Results

Genetic algorithms 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 might improve performance. In this study, the genetic algorithms technique was used to identify the most relevant features which impact student academic success and retention the most.

5.4.1 Optimized Student Academic Success Model

This section describes the results obtained by using a neural network to model student academic success using an optimized set of student inputs that was generated by the genetic algorithm. The results were obtained for three dataset cohorts- all students as a general model, majority students, and URM student cohorts. A detailed explanation of the genetic algorithms results and the neural networks performance is provided in the sections that follow.

5.4.1.1 Feature subset selection

The output of the genetic algorithm is the most relevant student features to the freshman year performance (overall first year GPA). Results showed that there were similarities and differences between groups as shown in table 14. The table represents binary vectors with the value of 1 if the feature is selected; and 0 otherwise. It was observed that six relevant features were common among the three groups (all students, majority students and URM students); they are: total credits earned in fall semester (TCEF), SAT math

score (SATM), total credits attempted in spring semester (TCAS), total credits earned in spring semester (TCES), first mathematics course (CourseM), and grade of first mathematics course (GradeM). This gives an indication of the importance of mathematics for student academic success. Gender was selected for the URM and majority groups as a relevant feature, but not for all students which could be related to the variation of included features within each group. In addition, the Honors variable was not selected for the URM group but was selected for the majority group. SATV was not selected for any of the three groups, while SATC was selected for all students and the majority groups. Residency was not selected for any of the groups. In addition, Rank was not selected neither for the majority not the URM groups. The total college credits attempted and earned in the spring semester was selected for both the majority and URM groups. However, the total college credits attempted and earned were selected only for the majority group.

Table 14 Output of GPA model feature subset selection by group

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

5.4.1.2 Model Performance – All students

This section describes the results obtained by using the neural network to model first year student academic success using the ten features selected by the genetic algorithm. The RMSE value is 0.44 which is approximately the same values that was obtained when all student features used to build the network. See table 15.

In analyzing the model's accuracy, it was revealed that the mean error for predicting the overall GPA is approximately 0.34 and the model was able to predict the GPA of 76.9% of students within an error of less than or equal to ± 0.5 on scale of 4.0 (which is equivalent to an absolute error of less than or equal to12.5%). The summary results are shown in table 15. The table also shows that only 2.5% of students had a prediction error of greater than ± 1.0 (an absolute error of greater than 25%).

Table 15 Summary results of the GPA absolute error analysis

Variable	Output
RMSE	0.44
Max	2.829
Min	0.0000
Mean	0.34
Std	0.28
GPAerr≤0.25	45.6%
0.25<GPAerr≤0.50	31.3%
0.50<GPAerr≤0.75	15.3%
0.75<GPAerr≤1.00	5.3%
GPAerr>1.00	2.5%

5.4.1.3 Model Performance - Majority Students

It was observed that the majority student's model gave similar results when it used all features as inputs compared to the model that used the selected subset of features. The RMSE value is 0.46 for the model which used an optimum student features as shown in table 16.

The model's accuracy analysis showed that the mean error for predicting the overall GPA is approximately 0.37. In addition, the model predicted the GPA of 73.1% of students within an error of ± 0.5 and it was noticed that 3% of students had a prediction error of greater than ± 1.0 (an absolute error of greater than 25%), see table 16.

Table 16 Summary results of the GPA absolute error analysis for majority students

Variable	All with Feature Selection
RMSE	0.46
Max	1.644
Min	0.000
Mean	0.37
std	0.27
GPAerr≤ 0.25	38.6%
0.25< GPAerr≤0.50	34.5%
0.50< GPAerr≤0.75	18.3%
0.75< GPAerr≤1.00	5.6%
GPAerr>1.00	3%

5.4.1.4 Model Performance- URM Students

The RMSE for the URM student model was 0.45. It was observed that the performance was the same regardless of the input set used to build the model. The mean error for predicting the overall GPA is approximately 0.35 for the model with the two different input sets. It was indicated that when an optimized subset of features used, the model predicted the GPA of 75.9% of students within an error of ± 0.5 (an absolute error of less than or equal to12.5%). Further, it was noticed that only 2.8% of students had a prediction error of greater than ± 1.0 (an absolute error of greater than 25%). The summary results are shown in table 17.

Table 17 Summary results of the absolute GPA error analysis for URM students

Variable	All with Feature Selection
RMSE	0.45
Max	2.872
Min	0.0001
Mean	0.35
std	0.29
GPAerr≤0.25	44.4%
0.25< GPAerr≤0.50	31.5%
0.50< GPAerr≤0.75	15.4%
0.75< GPAerr≤1.00	5.9%
GPAerr>1.00	2.8%

The neural networks models for all students, majority students, and URM students' performance was similar at predicting the freshman year overall GPA using an optimum set of inputs. In comparing the network's accuracy when all student features used and when an optimum set of student features used, no significant difference was observed. However, using the genetic algorithm provided a simplified and interpretable model that uses the most relevant student inputs with a slight improvement in the networks accuracy and an increase in the network's learning time.

5.5 Retention Model

As in the student academic success model, this section describes the results obtained for the retention model using an optimized set of inputs which was generated by the genetic algorithm. The model used three dataset cohorts- all students, majority and URM cohorts. Results obtained are explained in the following sections.

5.5.1 Feature subset selection

The results of this section 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 18. It was observed that seven features were common among the three groups, and they were: 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. In addition, as in the academic success section, 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. Unlike the academic success model, 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.

Table 18 Output of Retention model feature subset selection by group

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

5.5.2 Modeling freshman Retention

This section describes the results obtained from modeling student fall to fall retention using a selected subset of most relevant features. The results of the model's performance were compared within the groups that used the optimum set of student features and between the groups that used all available student features.

5.5.2.1 Model Performance – All students

Of the 1966 samples, it was indicated that the accuracy of the model slightly improved when feature selection set was used. The model's accuracy was 74% when all student inputs were used and 75% when the optimized set was used. Figure 7 represents the ROC curve of this model.

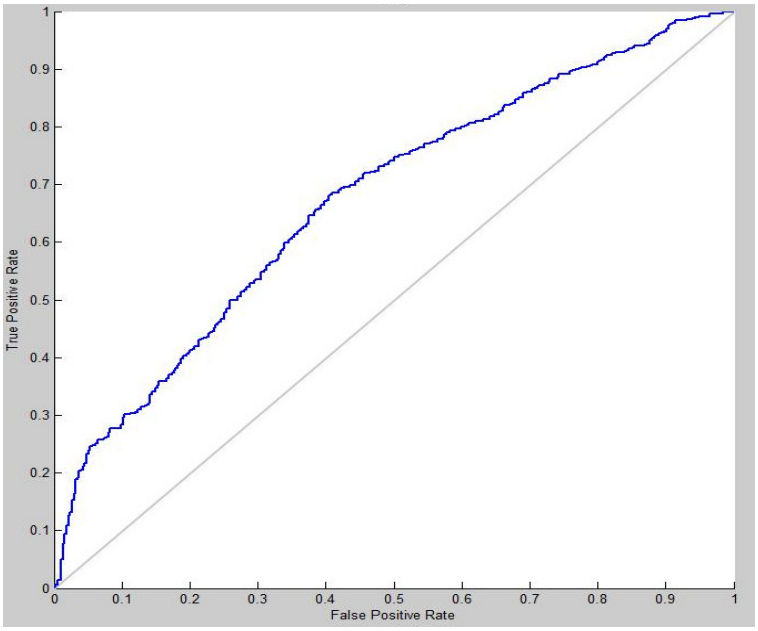

Figure 7 ROC curve of all students using optimized inputs

5.5.2.2 Model Performance- Majority Students

It was observed that the majority student model's accuracy increased approximately 2% when the selected subset of features was used as input; it was 79% without feature selection and 81% with feature selection. The ROC curve is shown in figure 8.

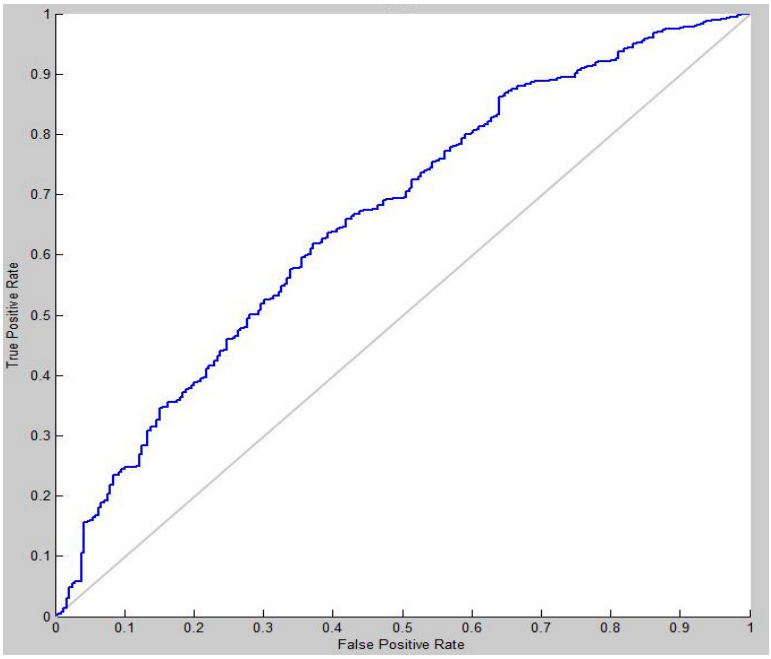

Figure 8 ROC curve of majority students using optimized inputs

5.5.2.3 Model Performance- URM Students

As for the URM student retention model, the accuracy of the model increased 3% when the selected subset of features was used. The model's accuracy in predicting non-retained students was 60% when using all student inputs and 63% when using an optimized subset of student inputs. Figure 9 shows the ROC plot for the model using the optimized set of inputs.

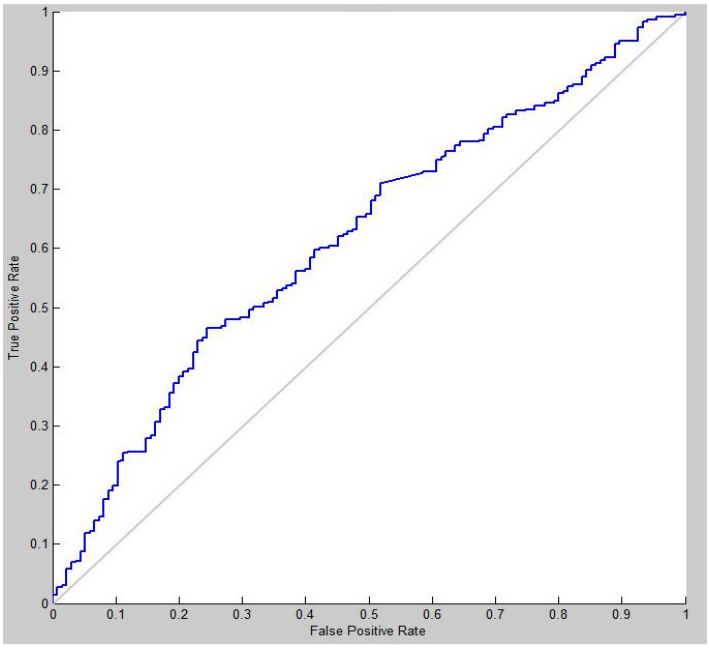

Figure 9 ROC curve of URM students using optimized inputs

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.

5.6 Qualitative Analysis

This section provides a further insight into URM students' academic and social life. In addition, the focus group questions shed light on the most relevant features that affect student retention behavior. Results of focus group sessions analysis and surveys are shown in the following sections. Methods and procedures are described in section 4.5.

5.6.1 Focus Group Sessions Analysis

Table 19 summarizes the background information on all participants. It was observed that a total of (n=16) 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. Eight students were placed into Calculus I (Math 200) and the average SAT score was 1620. The average high school GPA was 3.6 and the average study hours were 3.7 hours. Seven students indicated that they were the first generation to go to college. Only three students declared that they work during the academic year. Students' responses varied on how this affects their college life and participation in university activities. One student responded that she still had time to participate in organization's activities because she only works for a couple of hours per week; another responded that she managed her time between work and university activities; and one responded that he had no free time at all.

Table 19 Focus Groups Survey Summary Results

Gender	Race/Ethnicity	Major	Math Level	Work during academic year	First Generation Student
Female 12	African American 15	Biology 5	Algebra 0	Yes 3	Yes 7
Male 4	Hispanic American 1	Forensic Science 2	Pre-calculus 3	No 13	No 9
		Biomedical Engineering 2	Calculus I 8		
		Electrical and Computer Engineering 4	Calculus II 3		
		Mechanical Engineering 2	DE 1		
		Business Administration 1	Other 1		

Q1. Motivation for majoring in STEM discipline: In analyzing the students' responses, it was indicated that parents played a significant role in inspiring students to consider majoring in STEM. A majority of students believed that their parents motivated their decision to major in STEM in the first place. Some students saw their parents as role models and tried to follow their steps and pursue a career in STEM fields. Relatives and friends could be a good source of motivation as well.

Some students developed their mathematics and science skills since their high school period and they realized that STEM fields are commensurate with their career goals and abilities. Some stated that they were interested in a specific field of study in high school. A few students pointed to the importance of participating in a science or engineering program in high school. They were exposed to some college courses such as computer programming, biology, and environmental sciences. They indicated that these programs introduced them to science and engineering and to hands-on experiments. Some mentioned the effect of high school teacher, and that the main reason for majoring in engineering was a TV show that was an inspiration since childhood.

Q2. Freshman year experiences: 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, participation in science and engineering programs, and their participation in the STP as factors that helped in making first year introductory courses easier and smooth. The mathematics and

chemistry courses of the STP got high credits from students, although they were just a review for some of them. Socially, all students showed their concern for adjusting to college life and the new environment but it was easy for them because of their prior experience in high school programs and the STP.

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 course level had more pressure to be a freshman in a sophomore class level. From the social perspective, students found the STP very helpful for adjusting to VCU and meeting new friends, especially engineering students where they got familiar with the engineering buildings and labs. In addition, some found that joining student organizations such as NSBE (National Society of Blacks Engineers) was very helpful to get involved in college social life.

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 STP described her freshman year experience as “lonely”. The reason was that she did not know anybody at the beginning and later she joined NSBE to build relationships and find the support she needed to continue.

Q3. Difficulties in STEM during freshman year & how they were handled: Getting more specific about freshman year difficulties, students' responses, among all groups, were mostly from an academic perspective. It was observed that many students had difficulties in their first chemistry class. Even though most of them took the chemistry class during the STP, it was hard for them to keep up with such a demanding course and grasp any new material. The chemistry class was not added to the STP until the second year of the program. Due to this, a few students stated that it would be helpful if it was available at the beginning of the program; one mentioned that the last chemistry class she took was in 10^{th} grade, which is considered a big gap. Students revealed that they had to work harder, get tutoring, join SI sessions, and attend other chemistry classes taught by different instructors. Besides chemistry, a student expressed that her difficulty was in the introduction to engineering class because of the professor who expected that all students should know the basic material already, and moved forward from there. The student stated she had to put double effort and grasp the material quickly to improve her performance.

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 more workload than expected for a freshman especially if all three were taken at the same time and if the freshman never took physics in high school. Students handled this difficulty by attending SI sessions, going to the library and working with classmates. Online courses were a problem for freshmen as well. A student revealed that he was not ready for that

kind of classes which puts more responsibility to check homework and due dates online without having someone reminding him about the class duties.

Socially, 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. A student from the second group stated that the whole new teaching environment while another said that the campus life were not as they expected them to be when they came to the STP.

Q4. Indicators of freshman year performance 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. 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 it.

A couple of students stated that their good academic preparation in high school was due to their participation in mathematics or science programs. One student in the second group said that he did not have enough preparation in high school for college due to his school environment (small classrooms) but he emphasized that his father was the most

influential factor for encouraging him to major in engineering. Also, 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.

Advanced Placement (AP) classes were among the significant indicators of good performance in freshman year. Some students from all groups claimed that these advanced classes gave students an insight into college classes with regard to work-load and hard work. 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 emphasized that when they were self-motivated, they worked hard to achieve their goals and maintain academic success.

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 Black and now he is the only Black in his major. The student added that "I felt like I want to prove that only I am successful among all Blacks as I am the only Black graduating in

this major." One student pointed to the safe and diverse environment as a good indicator of freshman year performance.

Q5. Factors that impacted student academic performance: 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. A student, who dropped out of the engineering school, revealed that he did not do as well in the sophomore year as he did in the freshman year. The student added that after that he lost his self-motivation and started thinking about leaving engineering.

One more engineering student thought of leaving engineering when she got bad grades at the beginning of her freshman year. She had to re-motivate herself since she was the first to graduate from high school in her family, the first to go to college, and all of her family members were looking forward to seeing her graduating with an engineering degree. Another engineering student made the point that switching to a different major meant one more year in college, and the decision should be made in the freshman year to avoid more delay in graduation.

Some students indicated that even though they did not get good grades, or their GPA was not what they expected, they moved forward because of their self-motivation and their

family's support. One student added that this was his only option and he realized that this is what he wanted to do.

Q6. Environmental Factors that affected student academic performance and retention: Family and friends had the most influence on student retention decision. A majority of students revealed that their family member(s) formed a big support system. They tended to check on how they were doing in college and tried to push them towards academic success wisely. As for friends, most students stated they played a significant role in their adjustment to the college environment and improvement in academic performance especially for the STP students. They started their freshman year knowing many friends and attended the same freshman year classes together. Some 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.

Some students pointed out that their advisors did not help at all; one student said that her freshman year advisor was really helpful, while the rest of the students did not mention the role of their advisors in their freshman year at all. In addition, a couple of students stated that their professors did not influence them at all; some of them stated that it depends on the professor; and few stated that their professors were very helpful whenever they needed their assistance and that they were acting more nicely and supportive in their

offices than in class. In addition, a student claimed that teaching assistants sometimes could be more helpful than professors themselves. A couple of students highlighted the role of high school teachers in motivating them to do better in college. Money and roommates too were on the list as good and not very good influential factors, respectively.

As for the STP, some students emphasized on the influence of having friends from the program and how it made them more comfortable and they could adjust easily to VCU. Moreover, a few students highlighted the role of the mentoring program in maintaining academic success and getting the advice they need when they had any issue.

Q7. The STP experience: The STP 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, and did not get lost in the fall semester; adjusted to being away from home before fall started; learned time management because in high school they did not have free time as in college; 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.

Some biology and forensic science students from the first group stated that the program was more beneficial socially than academically because there were lots of mathematics and engineering activities.

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.

Chapter 6: Discussion

The purpose of this dissertation was to examine student features which have the most impact on first year student academic success and retention in STEM disciplines focusing on URM groups at VCU. The study utilized genetic algorithms and neural network techniques to model student academic success and retention using the most relevant student inputs. Further, this study employed a qualitative approach for examining more deeply student freshman year experiences and identifying major factors which affect their retention the most, both academically and socially. This chapter presents a detailed discussion of the results obtained from genetic algorithms, neural networks, and focus groups.

The genetic algorithm showed attractive results in terms of feature subset selection. Results were also comparable with responses obtained from focus groups. In general, student academic and social adjustment to college significantly reflects student first year academic performance and retention. It is even more influential for students in STEM disciplines due to the demanding nature of courses, and dependency on prior student preparation in mathematics and science.

In examining the results obtained, it was indicated that high school preparation has a great impact on student adjustment and performance in college. High school GPA; percentile rank; high school STEM programs; honors; mathematics and science teachers;

personal interests in mathematics and science; and mathematics placement test scores are all indicators of student high school preparation.

Students who had good preparation in high school tended to have better first year experiences due to their solid preparation, especially in their first semester classes. Usually, students with high interest in the fields of mathematics and science tended to major in STEM. The first mathematics course is a good predictor of student performance and retention, as well as a good indicator of student mathematics skills. Students who were placed into higher math level based on their placement test and had good mathematics skills tended to persist in their major besides earning more college credits. In the previous chapter, it was observed that approximately 75% of URM students started with algebra or pre-calculus in their freshman year. On the other hand, only about 46% of Majority students started with algebra or pre-calculus.

Interestingly, SAT scores were most likely selected by the genetic algorithm as strong predictors of student performance and retention. Conversely, it was considered an irrelevant indicator by participants of focus groups. In this study, SATM score was selected to build the hybrid model in which it could be a sign of student mathematics skills.

AP classes had a significant impact on student performance and retention in STEM as well. Students started with upper level courses and they did not have to worry about

being behind their peers in the freshman year. Not only this, it gave students an insight into college classes, and they learned about the demanding nature of this type of classes. Although they found taking sophomore courses while being freshmen slightly difficult to handle, they nonetheless showed great interest in being ahead of their peers and passing more college classes (i.e. earning more college credits) which allowed them to finish college earlier. These students expressed their interest in being independent, learning to handle advanced level college classes, and being more confident. Some students understood the cost of switching to a different major and the nature of college credits and load before they came to college. They knew that they needed to work hard, make wise decisions, and keep going. Being unaware of all the above could cause difficulties in freshman year.

Percentile rank was selected as a predictor variable for the URM retention model but not for the GPA model or the two majority models, although the average percentile rank for both majority and URM groups was approximately the same. This factor was reconsidered as an input for the comprehensive model because it is a good indicator of student high school preparation especially in the absence of high school GPA. Being accepted into the Honors College (honors variable) is another precollege preparation related factor. Students who got accepted into the Honors College tended to perform better in college due to their strong prior preparation the special curriculum they are taking. Nevertheless, this factor was not selected for the URM student group as relevant to student academic success and retention, while it was selected for the majority group.

This could be justified by the fact that only 5% of URM students got accepted into the Honors College over a three year period of time while this figure was 13% for majority students. Honors factor was not included in the comprehensive model inputs list because it could lower the model's performance.

Whether a student was in-state or out-of-state did not influence student decision about staying or leaving a STEM discipline. Instead, living on-campus or off-campus had a significant impact on student adjustment to VCU. What mattered to students the most was getting familiar with the VCU campus and the city.

Gender was a strong predictor for both URM and majority groups. As for race/ethnicity, URM engineering students, particularly females, tended to have more concerns for being a URM in a STEM major as compared to their peers from other non-engineering majors. It is believed that the reasoning behind race and gender being significant factors for engineering students has to do with the nature of engineering courses, laboratories, and projects. Engineering courses are more demanding with regard to team work in and outside classroom. Due to this, it is very important for these students to maintain high expectations, and put in double the effort to achieve good performance. This also explains why most URM students joined race related student organizations looking for support and advice. Some revealed that the diverse nature of VCU had a positive impact on their first year adjustment. Mostly, students with higher self-motivation had strong commitment to succeed and graduate with a STEM degree.

Freshman year academic performance plays a significant role in student dropout/persistence decisions. All students expressed their concerns about freshman year classes, both introductory and upper level. The overwhelming workload and lack of knowledge about college credits or course load made students put in double the effort for success. Responses of students on handling freshman year difficulties highlighted the variation in student retention behaviors. Some students worked hard to maintain good grades while others decided to withdraw from the class so that the overall GPA would not be affected. Getting good grades in the first semester has a great impact in motivating students and encouraging them to maintain high performance from the very beginning. Retention behavior for students who did not perform well in their freshman year differs from one student to another. Some start thinking that this is not what they should be doing and that they could achieve better in other non-STEM majors especially if they had difficulties in mathematics, which is a core subject in most STEM majors. Dropping out of STEM fields could happen unless students got self-motivated by other external factors such as family members or an advisor; this is discussed later in this section. However, some students would continue regardless of their performance. This could be influenced by the number of college credits they passed and the fact that they still could graduate with their peers without taking more year(s) to graduate by switching to a different major.

During freshman year students are exposed to several factors while trying to adjust to college life and handling college level classes. Gaskin [27] reported that “Direct and

indirect factors contributing to student success vary from student to student, between institutions, and even within a college or university. Students learn in different ways and through different experiences. Some students are better prepared for college than others." Due to this, student retention behaviors are varied even though students are exposed to the same factors and came from similar backgrounds. However, delving into student freshman year experience associated with academic and demographic factors could provide an insight into possible student behaviors, and ways of utilizing university resources and programs to increase retention in STEM fields.

Students who are better involved socially have higher self-motivation which is positively correlated with student first year performance and retention. Student self-motivation could be empowered by numerous factors. Family background and influence comes in first place for motivating students to major in STEM by keeping up with them until graduation. Even if parents did not go to college, it does not mean that their role ends once their kids enter college. They still can play an important role in providing inspiration and motivation that a student needs.

Another evidence of differences in students' behaviors is the learning process inside and outside the classroom. Each student follows different strategies to learn although similarities would be found. Those strategies are usually influenced by the class type and student major. Engineering students look for teamwork and support from upper class students by joining student organizations or hanging out in the engineering laboratory.

On the other hand, for non-engineering students, individual support, such as getting SI sessions, would be enough to be satisfied with their performance in the class.

Incoming freshman intervention programs, for the most part, positively impacted participants both academically and socially. Such programs helped students to adjust to college and gave an insight into what to expect. Participants in such programs were well prepared for college life and learned how to be independent and well-organized. On the other hand, some participants needed much more than that in order to succeed in college and persist in a STEM field afterwards. Mostly, URM student programs are self-selecting programs; in other words, students choose whether to participate or not. Usually, URM students do not prefer to be identified as minorities and to participate in programs labeled just for minorities. Thus, analyzing URM student characteristics and experiences allows institutions to employ available programs, resources, and activities to meet different students' needs. Besides, exploring student experiences and characteristics, and predicting student performance and retention would have a great impact on student advising process.

In conclusion, the results presented in this study indicated that the genetic algorithm is an attractive approach in selecting an adequate subset of relevant factors to build neural network predictive models. Genetic algorithms used in this study for optimizing student inputs set mathematically turned out to be a fast and easy method to exploit solutions. Focus groups were used to identify the most significant student inputs that greatly

impacted student academic success and retention. We are not comparing the results of genetic algorithms and focus groups in this study. The goal is to incorporate results obtained in order to develop a comprehensive model which is intelligible to the end-user and accurately predicts academic success and retention of URM students in STEM disciplines. From the results in the previous chapter, there was no significant difference observed in the model's performance when used with two different sets of inputs. This could be regarding to the total number of student features used in this dissertation. However, a better improvement of the model performance might be achieved by using additional student features.

In analyzing both the results, it was indicated that although the genetic algorithms ignored such factors as are redundant and irrelevant to student academic performance and retention, some of these factors were considered significant regarding URM student accomplishment. The genetic algorithm selected ten out of eighteen student inputs. Three student inputs were reselected, namely credits attempted in fall semester (CAF), credits earned in fall semester (CEF), and Rank. Both credits attempted in fall semester (CAF) and credits earned in fall semester (CEF) give an indication of the number of college credits earned, AP credits, and the extent of student academic performance, while rank gives an indication of student high school preparation. Four were discarded from the list, total credits earned in fall semester (TCEF), total credits attempted in spring semester (TCAS), total credits earned in spring semester (TCES), and combined SAT scores (SATC), to avoid redundancy since CAF and CEF are both selected. Students' response

was that SAT is not an indicator of how well they are expected to do in their freshman year. However, SATM is believed to give an indication of student mathematical skills and it was already selected by the genetic algorithm. For the retention model, SATM, CES, and term GPA of fall semester (TGPAF) were reselected while verbal SAT score (SATV) was discarded from the list.

Chapter 7: Hybrid Model

7.1 Introduction

This chapter discusses results obtained from using neural networks to model URM student academic success and retention by incorporating results obtained from the genetic algorithm and focus groups. The selected set of inputs represents student features which impact student performance and retention the most. Models were developed and validated using the 10 fold cross-validation method. Sections 7.2 and 7.3 include the results of neural networks model for predicting student performance and retention, respectively.

7.2 URM student academic success hybrid model performance

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 10.

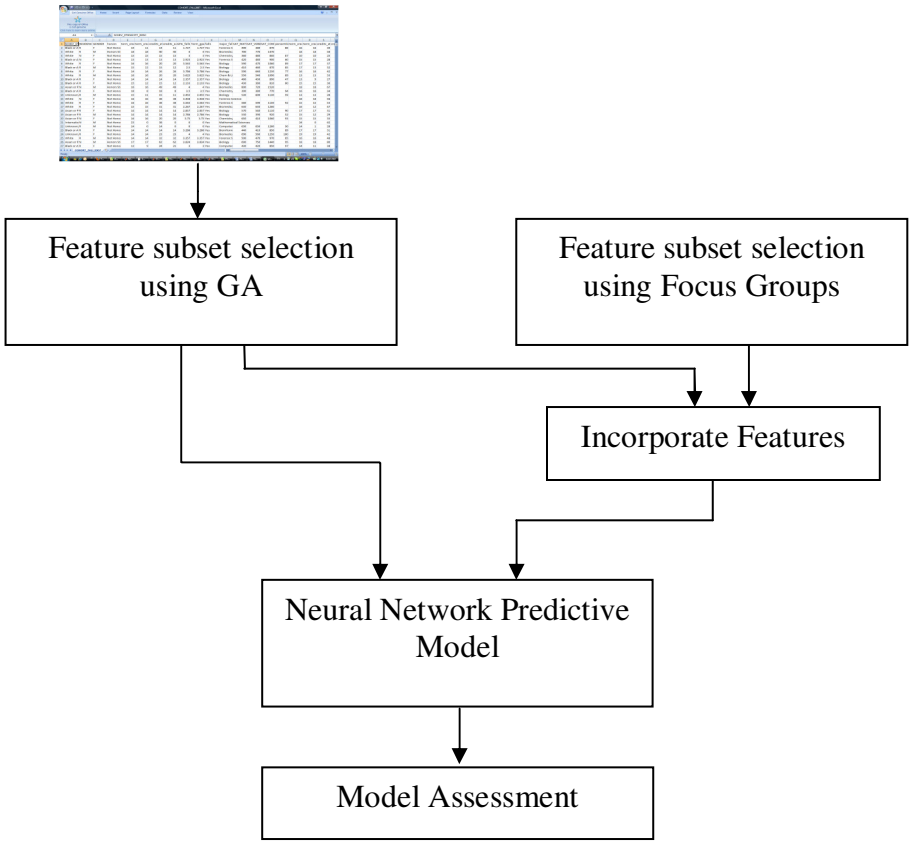

Figure10 Hybrid Model Block Diagram

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. The RMSE value is 0.45 and the mean error value obtained is 0.35 as shown in table 20.

In analyzing the model's accuracy, it was revealed that the model of all student features was able to predict the GPA of 72.1% of students within an error of less than or equal to ± 0.5 on scale of 4.0 (which is equivalent to an absolute error of less than or equal to12.5%). Also, it was observed that 23.9% of students had a prediction error greater than ± 0.5 and less than ± 1.0. Only 4% of students had a prediction error of greater than ± 1.0 (an absolute error of greater than 25%). The summary results are shown in table 20.

Table 20 Summary results of the absolute GPA error analysis for URM student-hybrid model

Variable	Error
RMSE	0.45
Max	2.409
Min	0.0000
Mean	0.35
std	0.28
GPAerr≤0.25	39%
0.25< GPAerr≤0.50	33.1%
0.50< GPAerr≤0.75	17.1%
0.75< GPAerr≤1.00	6.8%
GPAerr>1.00	4%

7.3 URM student retention hybrid model performance

As for the URM student retention model, the accuracy of the model increased to 66% compared to the model's accuracy when genetic algorithm optimized set of inputs was used. The ROC curve is shown in figure 11.

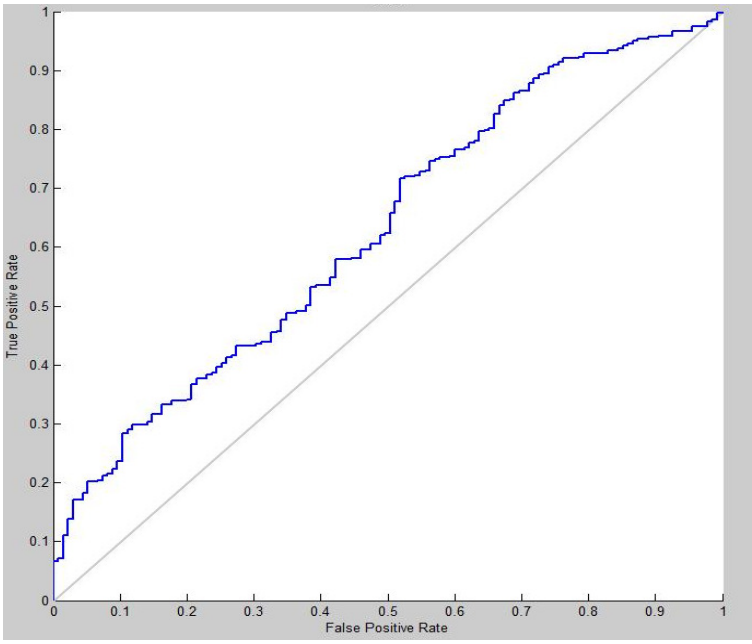

Figure 10 ROC curve of URM students-Hybrid model

7.4 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 academic success and retention without any direct interaction with students. On the other hand, 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 basically to incorporate results, and not to compare them, in order to develop a hybrid predictive model and validate it using a 10 fold cross-validation.

The model includes key student inputs in a way that it provides a deep understanding of significant factors and predicts student accomplishment in STEM disciplines at the same time. In other words, 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. Besides, the qualitative analysis of URM student freshman year experiences would play a positive role in analyzing student performance and retention behaviors.

At VCU, approximately 25% of freshman population in STEM fields is URM students. Consequently, it is useful to use the hybrid framework to model academic success and 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.

In predicting URM student academic success, it was observed that the hybrid model performed comparable to the model developed using the optimized set of inputs that resulted from the genetic algorithm with an RMSE value of 0.45. As for the retention model, the hybrid model's accuracy was increased 3% compared to the model which used the optimized set of inputs. 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 desired optimum. Instead they might only be a partial optimal value. Due to 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. Therefore, the model used in this study focused on incorporating results obtained from neural networks and focus groups so as to understand deeply student academic success and retention in STEM fields.

Chapter 8: Conclusions and Recommendations

8.1 Introduction

Modeling academic success and 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. This chapter will discuss conclusions and recommendations based on this research.

8.2 Conclusions

The process of developing the hybrid predictive model has three major phases as follows:

- Identifying the most relevant factors to student academic success and retention at VCU in STEM fields using genetic algorithms for all student groups- as a general inputs set, majority student group, and URM student group. Once the optimized set of inputs was generated for each group, a neural network model was developed and validated using two different sets of inputs: 1) all student academic success predictors, and 2) an optimized set of student academic success predictors. For student academic success, the overall freshman year GPA was used as the response variable, while student fall to fall retention was used for the retention model. Results obtained from the neural networks models were analyzed and compared within each group to construct an idea of the model's performance

with different sets of predictors. It was found that there were common predictors between groups.

- Identifying the most relevant factors to URM student academic success and retention at VCU in STEM fields using qualitative analysis (focus groups). As mentioned earlier, the focus group questions were divided into two major parts: 1) Determining significant precollege, academic, and environmental student features and 2) Analyzing URM freshmen college experiences and providing an adequate understanding of different dropout behaviors.
- Results of phases one and two were incorporated to develop the hybrid model that has those student inputs which impact URM student academic success and retention the most. The model can be easily applied using available university data to predict and analyze targeted students' accomplishments.

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 academic success and retention. A major part of this study was to construct an adequate understanding of URM student persistence/dropout behaviors in STEM fields.

VCU 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 it 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 academic success and 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. For the student academic success model that used all student inputs, the RMSE for all students, majority, and URM students were 0.45, 0.47, and 0.45,

respectively. When an optimized subset of student inputs used for the same model, the RMSE for all students, majority, and URM students were 0.44, 0.46, and 0.45, respectively. The model provided a good accuracy of modeling student performance using the overall freshman year GPA. The hybrid model performance was the same as the academic success model's performance for URM students.

As for the retention model, 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 models 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 interpreted model.

The related research work presented in [24-27, 39-43] 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 dissertation 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 academic success and 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 [24] 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 of time, and achieved an 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 academic success and retention for students in STEM disciplines, especially for URM students.

8.3 Future work and recommendations

Future studies could use the same procedures and models with a larger dataset and more participants for focus group sessions in order to provide a better reflection of URM STEM student first year experiences and retention behaviors. This study focused on identifying student inputs which have the most impact on academic success and retention. Analyzing these student inputs assists in studying students as individuals, each with particular characteristics, different behaviors, and specific needs. Freshman year college experiences could be narrowed down and categorized in order to be used as inputs to the model. It would be effective to apply the developed models during first and second semesters of freshman year to provide institutions with a supportive base of knowledge about targeted students' accomplishments.

This study could be extended by adding more precollege inputs to the model such as family background, high school GPA, and financial aid background. It was a challenge in this study to focus only on the limited available precollege factors especially when most of them were selected as irrelevant, such as residency, honors, and SATV. Predicting student academic success and retention at an early stage (i.e. at the beginning of freshman year) would be effective to enhance targeted students' performance and to increase retention rates in STEM fields. Also, it would assist in gearing intervention programs towards particular groups of students in order to address their needs.

As discussed in chapter six, high school preparation has a great impact on STEM student first year academic performance. STEM majors first year curriculum focuses on the areas of mathematics and science. Therefore, students with better high school preparation tend to do well in their freshman year which in turn is directly correlated with student retention decision. Nevertheless, targeted students should get the help they need, once they get into college based on their prior preparation, to lead them to success. College performance is as important as high school preparation. Future studies could include more factors reflecting student freshman year performance such as science course grades and chemistry placement test scores.

At Virginia Commonwealth University, the University College and STEM schools around the campus are running numerous valuable programs and activities for freshmen to support their transition to college and improve their academic performance. Since most of these are self-selecting programs, some students either do not step-in to ask for help or do not participate in the right program that addresses their needs. Sometimes, even if they do participate in the right program, they might not realize how to benefit from that program at the right time. Continuous follow up with students is essential to keep them on the right track to success, and to retain them in STEM fields. As concluded from this study, the role of the academic advisor in the educational system could be expanded and empowered to continuously follow up with individuals.

In most of related literature the neural networks proved its superiority among other techniques. It is also preferable regarding its good generalization ability and its capability of handing non-linearity and missing values among variables. However, different data mining techniques could be considered in future research such as decision trees, random forests, and Bayesian classifiers, especially for modeling URM student retention.

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 during the first or second semester (i.e. 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 precollege programs or from taking AP classes other than a general idea by comparing the total freshman year college credits earned 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 the STP since 2008. The participants represented all 63 former STP students in order to get a deep understanding of URM student first year experiences and retention behaviors. However, participants of the focus groups were diverse and comparable to the larger group of STP participants. Further, the study

collected a survey prior conducting the focus group sessions which might cause results to be skewed regarding self-reporting issues such as the SAT scores.

A final limitation was that this study focused on first year academic success and retention in STEM disciplines. Accordingly, an extension of this work could be to track students beyond freshman year. Also, determine graduation in STEM, non-STEM, and dropout of college including transfer students.

References

References:

[1] J. Brown, "Neural Network Prediction of Math and Reading Proficiency as Reported in the Educational Longitudinal Study 2002 Based on Non-Curricular Variables," School of Education, Duquesne University, 2007.

[2] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of educational research,* vol. 45, p. 89, 1975.

[3] W. Spady, "Dropouts from higher education: An interdisciplinary review and synthesis," *Interchange,* vol. 1, pp. 64-85, 1970.

[4] J. Bean, "Dropouts and turnover: The synthesis and test of a causal model of student attrition," *Research in Higher Education,* vol. 12, pp. 155-187, 1980.

[5] P. Terenzini and E. Pascarella, "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, 1980.

[6] L. Hagedorn, "How to define retention: A new look at an old problem," *College student retention: Formula for student success,* pp. 89-105, 2005.

[7] E. Seymour, "Tracking the processes of change in US undergraduate education in science, mathematics, engineering, and technology," *Science Education,* vol. 86, pp. 79-105, 2002.

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

[9] M. Besterfield-Sacre*, et al.*, "Characteristics of Freshman Engineering Students: Models for Determing Student Attrition in Engineering," *JOURNAL OF ENGINEERING EDUCATION-WASHINGTON-,* vol. 86, pp. 139-150, 1997.

[10] T. Mitchell and A. Daniel, "A Year-Long Entry-Level College Course Sequence for Enhancing Engineering Student Success."

[11] L. Fleming*, et al.*, "AC 2008-1039: ENGINEERING STUDENTS DEFINE DIVERSITY: AN UNCOMMON THREAD," 2008.

[12] J. Urban*, et al.*, "Minority engineering program computer basics with a vision," 2002, pp. S3C1-5.

[13] R. Hobson and R. Alkhasawneh, "SUMMER TRANSITION PROGRAM: A MODEL FOR IMPACTING FIRST-YEAR RETENTION RATES FOR UNDERREPRESENTED GROUPS," in *ASEE conference & exposition*, Austin, TX, 2009.

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

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

[16] G. May and D. Chubin, "A retrospective on undergraduate engineering success for underrepresented minority students," *JOURNAL OF ENGINEERING EDUCATION-WASHINGTON-,* vol. 92, pp. 27-40, 2003.

[17] K. B. Stephen Roberts, Vikram Shishodia, Jeff Citty, Deborah Mayhew, James Ogles, Angela Lindner, "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, 2009.
[18] R. Hackett and G. Martin, "Faculty Support for Minority Engineering Programs," 1997.
[19] M. Reichert and M. Ahsher, "Taking another look at educating African American engineers: The importance of undergraduate retention," *JOURNAL OF ENGINEERING EDUCATION-WASHINGTON-,* vol. 86, pp. 241-254, 1997.
[20] M. Sidle and J. McReynolds, "The freshman year experience: Student retention and student success," *NASPA JOURNAL,* vol. 36, pp. 288-300, 1999.
[21] F. Nave*, et al.*, "Prairie View A&M University: Assessing the impact of the STEM-Enrichment Program on women of color," 2006.
[22] S. Hargrove and L. Burge, "Developing a six sigma methodology for improving retention in engineering education," 2002, pp. S3C20-24.
[23] A. Williford and J. Schaller, "All retention all the time: How institutional research can synthesize information and influence retention practices," 2005.
[24] S. Herzog, "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, 2006.
[25] J. Lin*, et al.*, "Student Retention Modelling: An Evaluation of Different Methods and their Impact on Prediction Results."
[26] G. Zhang*, et al.*, "Identifying factors influencing engineering student graduation and retention: A longitudinal and cross-institutional study," 2002.
[27] B. Gaskins, "A Ten-Year Study of the Conditional Effects on Student Success in the First Year of College," Bowling Green State University, 2009.
[28] E. Durkheim, "Suicide: A study in sociology (JA Spaulding & G. Simpson, Trans.)," *Glencoe, IL: Free Press.(Original work published 1897),* 1951.
[29] A. Astin, *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*: Oryx Pr, 1991.
[30] A. Astin, "Student involvement: A developmental theory for higher education," *Journal of College Student Development,* vol. 40, pp. 518-529, 1999.
[31] R. Reason, "Student variables that predict retention: Recent research and new developments," *NASPA JOURNAL,* vol. 40, pp. 172-191, 2003.
[32] M. Anderson-Rowland, "A first year engineering student survey to assist recruitment and retention," 1996, pp. 372-376.
[33] J. Nicklow*, et al.*, "A short-term assessment of a multi-faceted engineering retention program," 2009, pp. 424-429.
[34] V. Tinto, "Taking student retention seriously," *Syracuse University, Syracuse, NY,* 1995.
[35] K. Maton*, et al.*, "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,* 2007.

[36] J. Heywood, *Engineering education: research and development in curriculum and instruction*: IEEE Press, 2005.
[37] E. Crawley*, et al.*, *Rethinking engineering education: The CDIO approach*: Springer Verlag, 2007.
[38] A. Persaud and A. Freeman, "Creating a Successful Model for Minority Students' Success in Engineering: The PREF Summer Bridge Program."
[39] G. Mendez*, et al.*, "Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests," *JOURNAL OF ENGINEERING EDUCATION-WASHINGTON-,* vol. 97, p. 57, 2008.
[40] G. Dekker*, et al.*, "Predicting Students Drop Out: A Case Study," 2009, pp. 41–50.
[41] N. Nghe*, et al.*, "A comparative analysis of techniques for predicting academic performance," 2007.
[42] J. Superby*, et al.*, "Determination of factors influencing the achievement of the first-year university students using data mining methods," 2006, pp. 37-44.
[43] P. Lam*, et al.*, "Predicting success in a minority engineering program," *Journal of Engineering Education,* vol. 88, pp. 265–267, 1999.
[44] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of mathematical biology,* vol. 5, pp. 115-133, 1943.
[45] I. Basheer and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," *Journal of Microbiological Methods,* vol. 43, pp. 3-31, 2000.
[46] C. Marshall and G. B. Rossman, *Designing qualitative research*: Sage Publications, Inc, 2010.
[47] D. L. Morgan, *The focus group guidebook* vol. 1: Sage Publications, Inc, 1998.
[48] H. Denton and D. Mcdonagh, "Using focus group methods to improve students' design project research in schools: Drawing parallels from action research at undergraduate level," *International Journal of Technology and Design Education,* vol. 13, pp. 129-144, 2003.
[49] V. Galarza*, et al.*, "IDENTIFYING IMPROVEMENT OPPORTUNITIES IN THE HIGH SCHOOL–COLLEGE BRIDGE FOR ENGINEERING STUDENTS: A FOCUS GROUP APPROACH," presented at the American Society for Engineering Education, Honolulu, HI, 2007.
[50] S. Stemler. (2001, *An overview of content analysis. Practical Assessment, Research & Evaluation*. Available: http://PAREonline.net/getvn.asp?v=7&n=17
[51] R. Flake and N. Touba, "Measuring Program Similarity for Efficient Benchmarking and Performance Analysis of Computer Systems."

Appendix A: IRB Approval Form

VCU Memo

Virginia Commonwealth University

Office of Research Subjects Protection
BioTechnology Research Park
BioTech One, 800 E. Leigh Street, #114
P.O. Box 980568
Richmond, Virginia 23298-0568
(804) 828-0868, fax (804) 827-1448

DATE: July 6, 2010

TO: Rosalyn S. Hobson, PhD
Electrical and Computer Engineering
Box 843068

FROM: Lloyd H. Byrd, MS
Chairperson, VCU IRB Panel E
Box 980568

RE: **VCU IRB #: HM12908**
Title: Impact of Pre-College Preparation for Incoming Freshmen in STEM Disciplines

On June 27, 2010 the following research study ***qualified for exemption*** according to 45 CFR 46.101(b) Category 2. This approval reflects the revisions received in the Office of Research Subjects Protection on. June 27, 2010. This approval includes the following items reviewed by this Panel:

RESEARCH APPLICATION/PROPOSAL: NONE

PROTOCOL: Impact of Pre-College Preparation for Incoming Freshmen in STEM Disciplines, version 4/7/10, received 4/9/10

CONSENT/ASSENT:
- Research Subject Information, version 6/27/10, 3 pages, received 6/27/10

ADDITIONAL DOCUMENTS:
- Invitation to attend Focus Groups, version 2-4/9/10, received 4/9/10

The Primary Reviewer assigned to your research study is Lloyd H. Byrd, MS. If you have any questions, please contact Mr. Byrd at lhbyrd@vcu.org; or you may contact Donna Gross, IRB Coordinator, VCU Office of Research Subjects Protection, at dsgross@vcu.edu or 827-2261.

Attachment – Conditions of Approval

Page 1 of 2

Conditions of Approval:

In order to comply with federal regulations, industry standards, and the terms of this approval, the investigator must *(as applicable)*:

1. Conduct the research as described in and required by the Protocol.

2. Obtain informed consent from all subjects without coercion or undue influence, and provide the potential subject sufficient opportunity to consider whether or not to participate (unless Waiver of Consent is specifically approved or research is exempt).

3. Document informed consent using only the most recently dated consent form bearing the VCU IRB "APPROVED" stamp (unless Waiver of Consent is specifically approved).

4. Provide non-English speaking patients with a translation of the approved Consent Form in the research participant's first language. The Panel must approve the translated version.

5. Obtain prior approval from VCU IRB before implementing any changes whatsoever in the approved protocol or consent form, unless such changes are necessary to protect the safety of human research participants (e.g., permanent/temporary change of PI, addition of performance/collaborative sites, request to include newly incarcerated participants or participants that are wards of the state, addition/deletion of participant groups, etc.). Any departure from these approved documents must be reported to the VCU IRB immediately as an Unanticipated Problem (see #7).

6. Monitor all problems (anticipated and unanticipated) associated with risk to research participants or others.

7. Report Unanticipated Problems (UPs), including protocol deviations, following the VCU IRB requirements and timelines detailed in VCU IRB WPP VIII-7):

8. Obtain prior approval from the VCU IRB before use of any advertisement or other material for recruitment of research participants.

9. Promptly report and/or respond to all inquiries by the VCU IRB concerning the conduct of the approved research when so requested.

10. All protocols that administer acute medical treatment to human research participants must have an emergency preparedness plan. Please refer to VCU guidance on http:/www.research.vcu.edu/irb/guidance.htm.

11. The VCU IRBs operate under the regulatory authorities as described within:
    a) U.S. Department of Health and Human Services Title 45 CFR 46, Subparts A, B, C, and D (for all research, regardless of source of funding) and related guidance documents.
    b) U.S. Food and Drug Administration Chapter I of Title 21 CFR 50 and 56 (for FDA regulated research only) and related guidance documents.
    c) Commonwealth of Virginia Code of Virginia 32.1 Chapter 5.1 Human Research (for all research).

Appendix B: Focus Group Protocol

One hour

1. What was the main thing that motivated you to major in STEM?

2. How do you describe your freshmen year experience? Would you say it was…easy, moderately easy, hard, or very hard? Why?

3. Can you talk about at least one thing (academic or social) that made it difficult for you to be successful in your STEM major during your freshmen year? And how did you handle it?

4. What are the most important factors that you believe indicated how well you would do in your freshman year? Such as SAT scores, gender, math placement test scores, high school performance.

5. Do you think that your freshmen year academic performance was a significant influence of your decision whether to pursue in STEM or switch into a non-STEM major? Which one of the following could impact your decision the most: your first/second semester GPA, first year cumulative GPA, Math courses performance, or college credits earned?

6. What about environmental factors such as professors, roommates, friends, advisors if any, family members, resources? Were they significant in influencing your persistence/dropout decision as well?

7. For the summer transition program participants, do you feel it was helpful to participate in a precollege preparation program? How was it helpful?

Focus Group Survey Questions

1) What is your gender? __Male __Female

2) Are you of Hispanic origin? __Yes __No

3) What is your race?

☐African American
☐Native American/ Alaskan Native
☐Other

4) What is your current major? _____

5) What were your SAT scores? Math_____

Verbal _____ Writing_____

6) What was your high school GPA?_____out of_____

7) What math did you place into at VCU? _____

8) Do you work during the academic year?
☐Yes
☐No

9) If yes, how does this affect your college life and participation in university activities? Please specify.

10) Are you a first generation student?
☐Yes
☐No

On average, how many hours do you study per day?

Appendix C

C 1. Neural Networks Source Code-Academic Success Model

```
%backpropagation network training with Levenberg-Marquardt algorithm

[data_rows,data_cols]=size(data);

p=data(:,1:data_cols-1)';
t=data(:,data_cols)';

hiddenLayerSize = 3; % hidden layer size changes based on dataset & input set used

%Cross-validation implementation
indices = crossvalind('Kfold',t,10);
dlmwrite('indices',indices ,'-append');
load indices;

% Network Creation, training, & validation
for i = 1:10
val=(indices==i);trn= ~val; %Preparing validation & testing data
net=newff(p(:,trn),t(:,trn),hiddenLayerSize,{'tansig','purelin'},'trainlm');

net.divideParam.trainRatio=0.8;
net.divideParam.valRatio=0.2;
net.divideParam.testRatio=0;

[net,tr]=train(net,p(:,trn),t(:,trn));

%Network testing
out = sim(net,p(:,val));
valerr=out-t(:,val);

dlmwrite('outall',out' ,'-append');
dlmwrite('tall',t(:,val)' ,'-append');
dlmwrite('valerrall',valerr' ,'-append');
end

%Model's Accuracy Calculations
load valerrall;
perfval=mse(valerrall);
RMSE=sqrt(perfval)
```

```
mean=mean(abs(valerrall))
max=max(abs(valerrall))
min=min(abs(valerrall))
std=std(abs(valerrall))

err_output = abs(valerrall);
for i=1:data_rows
 if  abs(err_output(i,1)) <= 0.25
         lt_25(i,1)=1;
      else
         lt_25(i,1)=0;

          if  abs(err_output(i,1)) <= 0.50
          lt_50(i,1)=1;
          else
          lt_50(i,1)=0;

             if  abs(err_output(i,1)) <= 0.75
             lt_75(i,1)=1;
             else
             lt_75(i,1)=0;

                if  abs(err_output(i,1)) <= 1
                lt_100(i,1)=1;
                else
                lt_100(i,1)=0;

                   if  abs(err_output(i,1)) > 1
                   gt_100(i,1)=1;
                   else
                   gt_100(i,1)=0;
                   end

                end

             end

          end

 end
end

total_lt_25= sum(lt_25)/data_rows
total_lt_50= sum(lt_50)/data_rows
```

```
total_lt_75= sum(lt_75)/data_rows
total_lt_100= sum(lt_100)/data_rows
total_gt_100= sum(gt_100)/data_rows
total2 = total_lt_25+total_lt_50+total_lt_75+total_lt_100+total_gt_100

clear;
```

C 2. Neural Networks Code-Retention Model

```
%backpropagation network training with Levenberg-Marquardt algorithm

[data_rows,data_cols]=size(data);

p=data(:,1:data_cols-1)';
t=data(:,data_cols)';

hiddenLayerSize = 3; % hidden layer size changes based on dataset & input set used

%Cross-validation implementation
 indices = crossvalind('Kfold',t,10);
dlmwrite('indices',indices ,'-append');
load indices;

% Network Creation, training, & validation
 for i = 1:10
val=(indices==i);trn= ~val; %Preparing validation & testing data
net=newff(p(:,trn),t(:,trn),hiddenLayerSize,{'tansig', 'tansig'},'trainlm');

net.divideParam.trainRatio=0.8;
net.divideParam.valRatio=0.2;
net.divideParam.testRatio=0;

[net,tr]=train(net,p(:,trn),t(:,trn));

%Network testing
out = sim(net,p(:,val));
valerr=out-t(:,val);
dlmwrite('outall',out' ,'-append');
dlmwrite('tall',t(:,val)' ,'-append');
dlmwrite('valerrall',valerr' ,'-append');

end
```

```
load valerrall;
perfval=mse(valerrall);
RMSE=sqrt(perfval)

clear;

%ROC plotting and Confusion Matrix Calculation
[tpr,fpr,thresholds] = roc(tAll,outAll)
[c,cm,ind,per] = confusion(tAll,outAll)
```

Note: Same indices generated from the cross-validation are used for each group (all, majority, URM) when different input sets used.

B 3. Genetic Algorithm Objective Function

```
function avg_error=ruba_objfunpred_nn(indiv)

%load input file

dataset = Data_College_noBlanks_GPA;
[no_rows,no_col]=size(dataset);

inputs = dataset(:,1:no_col-1); %input data
output = dataset(:,no_col);

global predictee;

    if predictee==1
        subsetinputs = inputs(2:no_rows,:);
        subsetoutput = output(2:no_rows,:);
    elseif predictee~=no_rows
        subsetinputs = inputs([1:predictee-1 predictee+1:no_rows],:);
        subsetoutput = output([1:predictee-1 predictee+1:no_rows],:);
    else
        subsetinputs = inputs(1:no_rows-1,:);
        subsetoutput = output(1:no_rows-1,:);
    end

[chromo_size_row,chromo_size_column] = size(indiv);
[rows,columns]=size(subsetinputs);
```

```
weighted_data = zeros(rows,columns);

for i=1:rows
   for k=1:columns
      weighted_data(i,k)= subsetinputs(i,k)*indiv(1,k);
   end
end

inputs=weighted_data';
targets=subsetoutput';

% Create a backpropagation Network
hiddenLayerSize = 4;

net=newff(inputs,targets,hiddenLayerSize,{'tansig','purelin'},'trainlm'); %the activation
function is tansig for the retention model

% Train the Network
net = init(net);
[net,tr] = train(net,inputs,targets);

% Test the Network
outsim=sim(net,inputs);

for i=1:rows
errors(i,1) = abs(targets(i)-outsim(i));
end

 RMSE = sqrt((sum( (targets(:)-outsim(:)).^2) / no_rows) )
```