Transactions on Computational Science and Computational Intelligence

Hamid R. Arabnia · Ken Ferens
David de la Fuente · Elena B. Kozerenko
José Angel Olivas Varela
Fernando G. Tinetti *Editors*

Advances in Artificial Intelligence and Applied Cognitive Computing

Proceedings from ICAI'20 and ACC'20

Springer

Transactions on Computational Science and Computational Intelligence

Series Editor

Hamid R. Arabnia
Department of Computer Science
The University of Georgia
Athens, GA, USA

Computational Science (CS) and Computational Intelligence (CI) both share the same objective: finding solutions to difficult problems. However, the methods to the solutions are different. The main objective of this book series, "Transactions on Computational Science and Computational Intelligence", is to facilitate increased opportunities for cross-fertilization across CS and CI. This book series will publish monographs, professional books, contributed volumes, and textbooks in Computational Science and Computational Intelligence. Book proposals are solicited for consideration in all topics in CS and CI including, but not limited to, Pattern recognition applications; Machine vision; Brain-machine interface; Embodied robotics; Biometrics; Computational biology; Bioinformatics; Image and signal processing; Information mining and forecasting; Sensor networks; Information processing; Internet and multimedia; DNA computing; Machine learning applications; Multi-agent systems applications; Telecommunications; Transportation systems; Intrusion detection and fault diagnosis; Game technologies; Material sciences; Space, weather, climate systems, and global changes; Computational ocean and earth sciences; Combustion system simulation; Computational chemistry and biochemistry; Computational physics; Medical applications; Transportation systems and simulations; Structural engineering; Computational electro-magnetic; Computer graphics and multimedia; Face recognition; Semiconductor technology, electronic circuits, and system design; Dynamic systems; Computational finance; Information mining and applications; Astrophysics; Biometric modeling; Geology and geophysics; Nuclear physics; Computational journalism; Geographical Information Systems (GIS) and remote sensing; Military and defense related applications; Ubiquitous computing; Virtual reality; Agent-based modeling; Computational psychometrics; Affective computing; Computational economics; Computational statistics; and Emerging applications. For further information, please contact Mary James, Senior Editor, Springer, mary.james@springer.com.

More information about this series at http://www.springer.com/series/11769

Hamid R. Arabnia • Ken Ferens • David de la Fuente
Elena B. Kozerenko • José Angel Olivas Varela
Fernando G. Tinetti
Editors

Advances in Artificial Intelligence and Applied Cognitive Computing

Proceedings from ICAI'20 and ACC'20

Springer

Editors
Hamid R. Arabnia
Department of Computer Science
University of Georgia
Athens, GA, USA

Ken Ferens
Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, MB, Canada

David de la Fuente
Business Administration
University of Oviedo
Oviedo, Asturias, Spain

Elena B. Kozerenko
Institute of Informatics Problems
The Russian Academy of Sciences
Moscow, Russia

José Angel Olivas Varela
Technology and Information systems
Universidad de Castilla La Mancha
Ciudad Real, Ciudad Real, Spain

Fernando G. Tinetti
Facultad de Informática - CIC PBA
Universidad Nacional de La Plata
La Plata, Argentina

ISSN 2569-7072 ISSN 2569-7080 (electronic)
Transactions on Computational Science and Computational Intelligence
ISBN 978-3-030-70295-3 ISBN 978-3-030-70296-0 (eBook)
https://doi.org/10.1007/978-3-030-70296-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

It gives us great pleasure to introduce this collection of papers that were presented at the following international conferences: Artificial Intelligence (ICAI 2020) and Applied Cognitive Computing (ACC 2020). These two conferences were held simultaneously (same location and dates) at Luxor Hotel (MGM Resorts International), Las Vegas, USA, July 27–30, 2020. This international event was held using a hybrid approach, that is, "in-person" and "virtual/online" presentations and discussions.

This book is composed of nine parts. Parts 1 through 8 (composed of 78 chapters) include articles that address various challenges in the area of artificial intelligence (ICAI). Part 9 (composed of 12 chapters) includes a collection of research papers in the area of applied cognitive computing (ACC).

An important mission of the World Congress in Computer Science, Computer Engineering, and Applied Computing, CSCE (a federated congress to which this event is affiliated with), includes *"Providing a unique platform for a diverse community of constituents composed of scholars, researchers, developers, educators, and practitioners. The Congress makes concerted effort to reach out to participants affiliated with diverse entities (such as: universities, institutions, corporations, government agencies, and research centers/labs) from all over the world. The congress also attempts to connect participants from institutions that have **teaching** as their main mission with those who are affiliated with institutions that have **research** as their main mission. The congress uses a quota system to achieve its institution and geography diversity objectives."* By any definition of diversity, this congress is among the most diverse scientific meeting in the USA. We are proud to report that this federated congress had authors and participants from 54 different nations representing variety of personal and scientific experiences that arise from differences in culture and values.

The program committees (refer to subsequent pages for the list of the members of committees) would like to thank all those who submitted papers for consideration. About 50% of the submissions were from outside the USA. Each submitted paper was peer-reviewed by two experts in the field for originality, significance, clarity, impact, and soundness. In cases of contradictory recommendations, a member of the conference program committee was charged to make the final decision; often, this involved seeking help from additional referees. In addition, papers whose authors included a member of the conference program committee were evaluated using the double-blind review process. One exception to the above evaluation process was for papers that were submitted directly to chairs/organizers of pre-approved sessions/workshops; in these cases, the chairs/organizers were responsible for the evaluation of such submissions. The overall paper acceptance rate for regular papers was 20%; 18% of the remaining papers were accepted as short and/or poster papers.

We are grateful to the many colleagues who offered their services in preparing this book. In particular, we would like to thank the members of the Program Committees of individual research tracks as well as the members of the Steering Committees of ICAI 2020 and ACC 2020; their names appear in the subsequent pages. We would also like to extend our appreciation to over 500 referees.

As sponsors-at-large, partners, and/or organizers, each of the followings (separated by semicolons) provided help for at least one research track: Computer Science Research, Education, and Applications (CSREA); US Chapter of World Academy of Science; American Council on Science and Education & Federated Research council; and Colorado Engineering Inc. In addition, a number of university faculty members and their staff, several publishers of computer science and computer engineering books and journals, chapters and/or task forces of computer science associations/organizations from 3 regions, and developers of high-performance machines and systems provided significant help in organizing the event as well as providing some resources. We are grateful to them all.

We express our gratitude to all authors of the articles published in this book and the speakers who delivered their research results at the congress. We would also like to thank the following: UCMSS (Universal Conference Management Systems & Support, California, USA) for managing all aspects of the conference; Dr. Tim Field at APC for coordinating and managing the printing of the programs; the staff at Luxor Hotel (MGM Convention) for the professional service they provided; and Ashu M. G. Solo for his help in publicizing the congress. Last but not least, we would like to thank Ms. Mary James (Springer Senior Editor in New York) and Arun Pandian KJ (Springer Production Editor) for the excellent professional service they provided for this book project.

Book Co-editors and Chapter Co-editors: ICAI 2020 and ACC 2020

Athens, GA, USA	Hamid R. Arabnia
Winnipeg, MB, Canada	Ken Ferens
Oviedo, Asturias, Spain	David de la Fuente
Moscow, Russia	Elena B. Kozerenko
Ciudad Real, Ciudad Real, Spain	José Angel Olivas Varela
La Plata, Argentina	Fernando G. Tinetti
Seoul, South Korea	Charlie (Seungmin) Rho
Houston, TX, USA	Xiaokun Yang

Artificial Intelligence: ICAI 2020 – Program Committee

- Prof. Abbas M. Al-Bakry (Steering Committee); University of IT and Communications, Baghdad, Iraq
- Prof. Emeritus Nizar Al-Holou (Steering Committee); Electrical and Computer Engineering Department; Vice Chair, IEEE/SEM-Computer Chapter; University of Detroit Mercy, Detroit, Michigan, USA
- Prof. Emeritus Hamid R. Arabnia (Steering Committee); University of Georgia, USA; Editor-in-Chief, Journal of Supercomputing (Springer); Fellow, Center of Excellence in Terrorism, Resilience, Intelligence & Organized Crime Research (CENTRIC).
- Prof. Mehran Asadi; Department of Business and Entrepreneurial Studies, The Lincoln University, Pennsylvania, USA
- Prof. Juan Jose Martinez Castillo; Director, The Acantelys Alan Turing Nikola Tesla Research Group and GIPEB, Universidad Nacional Abierta, Venezuela
- Dr. Arianna D'Ulizia; Institute of Research on Population and Social Policies, National Research Council of Italy (IRPPS), Rome, Italy
- Prof. Emeritus Kevin Daimi (Steering Committee); Department of Mathematics, Computer Science and Software Engineering, University of Detroit Mercy, Detroit, Michigan, USA
- Prof. Zhangisina Gulnur Davletzhanovna; Vice-Rector of the Science, Central-Asian University, Kazakhstan, Almaty, Republic of Kazakhstan; Vice President of International Academy of Informatization, Kazskhstan, Almaty, Republic of Kazakhstan
- Prof. Leonidas Deligiannidis (Steering Committee); Department of Computer Information Systems, Wentworth Institute of Technology, Boston, Massachusetts, USA
- Dr. Roger Dziegiel; US Air Force Research Lab, AFRL/RIEA, USA
- Prof. Mary Mehrnoosh Eshaghian-Wilner (Steering Committee); Professor of Engineering Practice, University of Southern California, California, USA; Adjunct Professor, Electrical Engineering, University of California Los Angeles, Los Angeles (UCLA), California, USA

- Prof. Ken Ferens (Steering Committee); Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada
- Dr. David de la Fuente (Chapter Editor); University of Oviedo, Spain
- Hindenburgo Elvas Goncalves de Sa; Robertshaw Controls (Multi-National Company), System Analyst, Brazil; Information Technology Coordinator and Manager, Brazil
- Prof. George A. Gravvanis (Steering Committee); Director, Physics Laboratory & Head of Advanced Scientific Computing, Applied Math & Applications Research Group; Professor of Applied Mathematics and Numerical Computing and Department of ECE, School of Engineering, Democritus University of Thrace, Xanthi, Greece.
- Prof. George Jandieri (Steering Committee); Georgian Technical University, Tbilisi, Georgia; Chief Scientist, The Institute of Cybernetics, Georgian Academy of Science, Georgia; Ed. Member, International Journal of Microwaves and Optical Technology, The Open Atmospheric Science Journal, American Journal of Remote Sensing, Georgia
- Prof. Byung-Gyu Kim (Steering Committee); Multimedia Processing Communications Lab.(MPCL), Department of CSE, College of Engineering, SunMoon University, South Korea
- Prof. Tai-hoon Kim; School of Information and Computing Science, University of Tasmania, Australia
- Dr. Elena B. Kozerenko (Chapter Editor); Institute of Informatics Problems of the Russian Academy of Sciences, Moscow, Russia
- Prof. Louie Lolong Lacatan; Chairperson, Computer Engineerig Department, College of Engineering, Adamson University, Manila, Philippines; Senior Member, International Association of Computer Science and Information Technology (IACSIT), Singapore; Member, International Association of Online Engineering (IAOE), Austria
- Prof. Dr. Guoming Lai; Computer Science and Technology, Sun Yat-Sen University, Guangzhou, P. R. China
- Dr. Peter M. LaMonica; US Air Force Research Lab, AFRL/RIEBB, USA
- Prof. Hyo Jong Lee; Director, Center for Advanced Image and Information Technology, Division of Computer Science and Engineering, Chonbuk National University, South Korea
- Dr. Changyu Liu; College of Mathematics and Informatics, South China Agricultural University, Guangzhou, P. R. China and Visiting scientist, School of Computer Science, Carnegie Mellon University, USA
- Dr. Muhammad Naufal Bin Mansor; Faculty of Engineering Technology, Department of Electrical, Universiti Malaysia Perlis (UniMAP), Perlis, Malaysia
- Dr. Andrew Marsh (Steering Committee); CEO, HoIP Telecom Ltd (Healthcare over Internet Protocol), UK; Secretary General of World Academy of BioMedical Sciences and Technologies (WABT) a UNESCO NGO, The United Nations
- Dr. Mohamed Arezki Mellal; Faculty of Engineering Sciences (FSI), M'Hamed Bougara University, Boumerdes, Algeria

- Dr. Ali Mostafaeipour; Industrial Engineering Department, Yazd University, Yazd, Iran
- Dr. Houssem Eddine Nouri; Informatics Applied in Management, Institut Superieur de Gestion de Tunis, University of Tunis, Tunisia
- Prof. Dr., Eng. Robert Ehimen Okonigene (Steering Committee); Department of EEE, Faculty of Engineering and Technology, Ambrose Alli University, Edo State, Nigeria
- Dr. Jose A. Olivas (Chapter Editor); University of Castilla - La Mancha, Spain
- Prof. James J. (Jong Hyuk) Park (Steering Committee); DCSE, SeoulTech, Korea; President, FTRA, EiC, HCIS Springer, JoC, IJITCC; Head of DCSE, SeoulTech, Korea
- Dr. Xuewei Qi; Research Faculty & PI, Center for Environmental Research and Technology, University of California, Riverside, California, USA
- Dr. Charlie (Seungmin) Rho (Chapter Editor); Department Software, Sejong University, Gwangjin-gu, Seoul, Republic of Korea
- Prof. Abdel-Badeeh M. Salem; Head of Artificial Intelligence and Knowledge Engineering Research Labs and Professor of Computer Science, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt; Editor-In-Chief, Egyptian Computer Science Journal; Editor-In-Chief, International Journal of Bio-Medical Informatics and e-Health (IJBMIeH); Associate-Editor-In-Chief, International Journal of Applications of Fuzzy Sets and Artificial Intelligence (IJAFSAI)
- Dr. Akash Singh (Steering Committee); IBM Corporation, Sacramento, California, USA; Chartered Scientist, Science Council, UK; Fellow, British Computer Society; Member, Senior IEEE, AACR, AAAS, and AAAI; IBM Corporation, USA
- Ashu M. G. Solo (Publicity), Fellow of British Computer Society, Principal/R&D Engineer, Maverick Technologies America Inc.
- Dr. Tse Guan Tan; Faculty of Creative Technology and Heritage, Universiti Malaysia Kelantan, Malaysia
- Prof. Fernando G. Tinetti (Steering Committee); School of Computer Science, Universidad Nacional de La Plata, La Plata, Argentina; also at Comision Investigaciones Cientificas de la Prov. de Bs. As., Argentina
- Prof. Hahanov Vladimir (Steering Committee); Vice Rector, and Dean of the Computer Engineering Faculty, Kharkov National University of Radio Electronics, Ukraine and Professor of Design Automation Department, Computer Engineering Faculty, Kharkov; IEEE Computer Society Golden Core Member; National University of Radio Electronics, Ukraine
- Prof. Shiuh-Jeng Wang (Steering Committee); Director of Information Cryptology and Construction Laboratory (ICCL) and Director of Chinese Cryptology and Information Security Association (CCISA); Department of Information Management, Central Police University, Taoyuan, Taiwan; Guest Ed., IEEE Journal on Selected Areas in Communications.
- Dr. Todd Waskiewicz; US Air Force Research Lab, AFRL/RIED, USA

- Prof. Layne T. Watson (Steering Committee); Fellow of IEEE; Fellow of The National Institute of Aerospace; Professor of Computer Science, Mathematics, and Aerospace and Ocean Engineering, Virginia Polytechnic Institute & State University, Blacksburg, Virginia, USA
- Dr. Xiaokun Yang (Chapter Editor); College of Science and Engineering, University of Houston Clear Lake, Houston, Texas, USA
- Prof. Jane You (Steering Committee); Associate Head, Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong

Applied Cognitive Computing: ACC 2020 – Program Committee

- Prof. Emeritus Hamid R. Arabnia (Steering Committee); University of Georgia, USA; Editor-in-Chief, Journal of Supercomputing (Springer); Editor-in-Chief, Transactions of Computational Science & Computational Intelligence (Springer); Fellow, Center of Excellence in Terrorism, Resilience, Intelligence & Organized Crime Research (CENTRIC).
- Prof. Juan Jose Martinez Castillo; Director, The Acantelys Alan Turing Nikola Tesla Research Group and GIPEB, Universidad Nacional Abierta, Venezuela
- Prof. Leonidas Deligiannidis (Steering Committee); Department of Computer Information Systems, Wentworth Institute of Technology, Boston, Massachusetts, USA
- Dr. Ken Ferens (Co-Chair); Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB, Canada
- Prof. George A. Gravvanis (Steering Committee); Director, Physics Laboratory & Head of Advanced Scientific Computing, Applied Math & Applications Research Group; Professor of Applied Mathematics and Numerical Computing and Department of ECE, School of Engineering, Democritus University of Thrace, Xanthi, Greece.
- Prof. Byung-Gyu Kim (Steering Committee); Multimedia Processing Communications Lab.(MPCL), Department of CSE, College of Engineering, SunMoon University, South Korea
- Prof. Tai-hoon Kim; School of Information and Computing Science, University of Tasmania, Australia
- Prof. Louie Lolong Lacatan; Chairperson, Computer Engineerig Department, College of Engineering, Adamson University, Manila, Philippines; Senior Member, International Association of Computer Science and Information Technology (IACSIT), Singapore; Member, International Association of Online Engineering (IAOE), Austria
- Dr. Changyu Liu; College of Mathematics and Informatics, South China Agricultural University, Guangzhou, P. R. China and Visiting scientist, School of Computer Science, Carnegie Mellon University, USA

- Dr. Muhammad Naufal Bin Mansor; Faculty of Engineering Technology, Department of Electrical, Universiti Malaysia Perlis (UniMAP), Perlis, Malaysia
- Prof. Dr., Eng. Robert Ehimen Okonigene (Steering Committee); Department of Electrical & Electronics Engineering, Faculty of Engineering and Technology, Ambrose Alli University, Edo State, Nigeria
- Prof. James J. (Jong Hyuk) Park (Steering Committee); DCSE, SeoulTech, Korea; President, FTRA, EiC, HCIS Springer, JoC, IJITCC; Head of DCSE, SeoulTech, Korea
- Prof. Fernando G. Tinetti (Steering Committee); School of Computer Science, Universidad Nacional de La Plata, La Plata, Argentina; also at Comision Investigaciones Cientificas de la Prov. de Bs. As., Argentina
- Prof. Hahanov Vladimir (Steering Committee); Vice Rector, and Dean of the Computer Engineering Faculty, Kharkov National University of Radio Electronics, Ukraine and Professor of Design Automation Department, Computer Engineering Faculty, Kharkov; IEEE Computer Society Golden Core Member; National University of Radio Electronics, Ukraine
- Prof. Shiuh-Jeng Wang (Steering Committee); Director of Information Cryptology and Construction Laboratory (ICCL) and Director of Chinese Cryptology and Information Security Association (CCISA); Department of Information Management, Central Police University, Taoyuan, Taiwan; Guest Ed., IEEE Journal on Selected Areas in Communications.
- Prof. Layne T. Watson (Steering Committee); Fellow of IEEE; Fellow of The National Institute of Aerospace; Professor of Computer Science, Mathematics, and Aerospace and Ocean Engineering, Virginia Polytechnic Institute & State University, Blacksburg, Virginia, USA
- Prof. Jane You; Associate Head, Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong
- Dr. Farhana H. Zulkernine; Coordinator of the Cognitive Science Program, School of Computing, Queen's University, Kingston, ON, Canada

Contents

Part I
Deep Learning, Generative Adversarial Network, CNN, and Applications

Fine Tuning a Generative Adversarial Network's Discriminator for Student Attrition Prediction

Eric Stenton and Pablo Rivas

1 Introduction

Most colleges want to retain the number of freshman students enrolled and do what they can to prevent them from leaving within the first year. We will use the word "attrition" to describe students who have either dropped out or transferred to another college. A strong tool in lowering the amount of student attrition is the ability to predict who will leave as well as determine a trend or commonality between those who do leave. An inevitable problem with developing a good manner of prediction is the small amount of data that is available as a result of a typically small incoming class and the even smaller amount of those who leave. In other words, predicting student attrition in the first year can be proposed as an anomaly detection problem with a very limited amount of data to use in creating prediction models. In this paper, the freshman population of Marist College of years 2016 and 2017 will be examined using a GAN architecture in order to predict attrition in 2018. First, the neural network model learns the characteristics of a first-year student through adversarial learning. Second, the model is fine-tuned to classify students as either those who will stay or those who will leave. Third, the latent space of the layer directly before the final one that gives the final prediction is inspected for comparing three versions of the model. The versions are the following: The model traditionally trained without a GAN (the control), one adversarially trained without tuning, and one adversarially trained with tuning. The hypothesis is that the model that is adversarially trained with tuning will have a latent space more representative of the freshman population producing a higher accuracy when predicting student attrition.

E. Stenton (✉) · P. Rivas
Computer Science, Baylor University, Waco, TX, USA
e-mail: eric.stenton1@marist.edu; pablo.rivas@marist.edu

H. R. Arabnia et al. (eds.), *Advances in Artificial Intelligence and Applied Cognitive Computing*, Transactions on Computational Science and Computational Intelligence, https://doi.org/10.1007/978-3-030-70296-0_1

The following section will provide a brief background of the concepts in this paper. Following this section will be a description of the methodology used to test the models and how the models were built. The next section will be an overview of the three experiments performed, their accompanying diagrams, and a short explanation of the results. Finally, the last section will be a concluding paragraph on the findings of the experiments.

2 Background and Other Work

It is important to note this paper serves as an extension of research carried out by Dr. Eitel Lauria and colleagues in which the same population of students was used to predict attrition using multiple machine learning algorithms, the primary one being XGBoost [3]. Dr. Lauria's research produced models with accurate predictions of student attrition despite minimal amounts of data. This research extends the knowledge of neural models for student attrition introduced by E. Lauria et al. [8].

Current insights in GAN architectures originated in a paper by Dr. Ian Goodfellow et al. where the concept of a discriminator model and generator model playing a minimax game first arose [5]. Their paper shows the following value function for how the GAN operates:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \tag{1}$$

In the value function $V(D, G)$, G is a differential function representing the generator model that takes noise input $p_z(z)$ and maps it to a data space. This data space is meant to represent possible values that can mimic variables $p_{\text{data}}(\boldsymbol{x})$, real data, when inputted into another function represented by the discriminator model and denoted as D that outputs a prediction of whether the input was generated or not. D is trained to maximize the probability of correctly labeling generated and real samples while G is trained to minimize $\log(1 - D(G(z)))$, or lower the probability of D predicting correctly.

Shortly after Dr. Goodfellow's paper, the structure of the GAN training python code and the calculation of both the Wasserstein loss and gradient penalty for the training of the discriminator originated in an experiment from a paper by Martin Arjovsky, Soumith Chintala, and Léon Bottou [1]. The formula for the Wasserstein distance which is described in further detail in the referenced paper is the following:

$$W\left(\mathbb{P}_r, \mathbb{P}_g\right) = \inf_{\gamma \in \Pi(\mathbb{P}_r, P_g)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|]. \tag{2}$$

In the Wasserstein distance equation, $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ represents the set of all joint distributions $\gamma(x, y)$ with marginals $\mathbb{P}_r$ and $\mathbb{P}_g$, respectively. In order to transform distributions $\mathbb{P}_r$ into distribution $\mathbb{P}_g$, $\gamma(x, y)$ denotes the amount of "mass" to be transported from x to y while the Wasserstein distance describes the "cost" of the optimal method of transport.

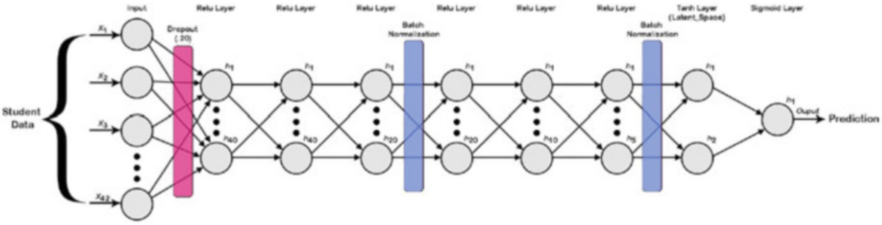

Fig. 1 Discriminator architecture diagram

The next section will describe the methodology for building the discriminator and generator models as well as how the Wasserstein distance equation will be utilized.

3 Methodology

The main pieces of GAN architectures are the discriminator and generator models as shown in Eq. 2. These models will be explained in this section in detail.

3.1 Discriminator

The discriminator is a neural model composed of 12 layers as shown in Fig. 1. These layers are: dropout, ReLU, batch normalization, tanh, and sigmoid. First, in order to prevent any one feature of the input data becoming heavily weighted, the dropout layer disconnects about 20% of the features randomly on each training step [9]. Second, batch normalization layers are placed intermittently to prevent the outputs of the ReLU layers from becoming too large and slowing or preventing convergence [4]. Third, the Python implementation of our model is based on Keras' functional model due to its ability to work with the tanh layer separately as this will serve as a view into the latent space of the model directly before an output is computed. Fourth, the discriminator's loss is based on weighing two Wasserstein loss calculations with a weight of one and a gradient penalty with a weight of ten.

3.2 Generator

The generator is a sequential model made up of 9 layers with a similar layout to the discriminator in which it has ReLU layers with intermittent batch normalization layers and an output consisting of a sigmoid layer as shown in Fig. 2. The most notable difference the generator has from the discriminator is the nature of its input

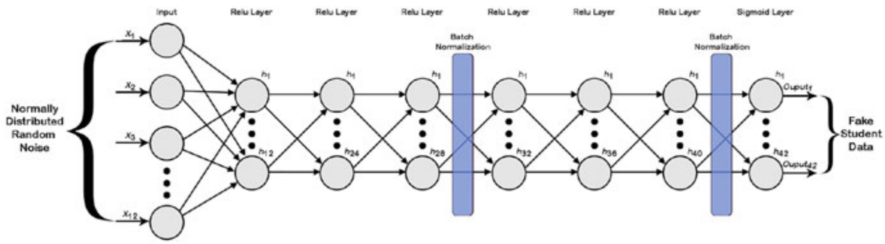

Fig. 2 Generator architecture diagram

which is 12 normally distributed random values between 0 and 1. These values are “noise” or values within a latent dimension defining different vectors that will eventually become generated data mimicking the input to the discriminator. This latent dimension should not be confused with the latent space referenced in this paper describing the output of the tanh layer in the discriminator. Furthermore, the generator’s loss function is simpler than the discriminator’s as it only consists of a single Wasserstein calculation.

The following section will present the three experiments conducted using the aforementioned models in detail as well as expound on the results of each.

4 Experiments and Results

Before getting into the details of the experiment, let us take a look at the input data. Table 1 describes the features and corresponding data types.

Some of the most noteworthy predictors in Table 1 are the following: “HSGPA,” “DistanceInMiles,” “MeritScholAmt,” and “APCourses.” “HSGPA” is a student’s GPA from high school measured with a 4.0 scale. “DistanceInMiles” is the distance from a student’s hometown to the college measured in miles. “MeritScholAmt” is the amount of money awarded to the student through a merit scholarship. Finally, the “APCourses” feature is a binary value where 1 means the student has taken AP courses and 0 means they have not. It is important to note that the majority of the aforementioned predictors relate to how well the student has done academically in high school. Furthermore, the “DistanceInMiles” predictor may indirectly relate to the student’s emotional well-being as a larger distance away from their hometown may limit visits home. However, due to the difficulty in measuring the importance of predictors in a neural network, the speculation on the impact each feature has on predicting student attrition is rooted in the work by E. Lauria et al. where many of the same predictors are used and measured based on their importance in multiple machine learning models [8].

Besides the most important predictors, it is also imperative to point out the most “noisy” predictors, or those that have a large number of null values, which are the following: “DistanceInMiles,” “OccupantsBuilding,” “OccupantsRoom,” and

Table 1 Description of predictors

Feature	Description	Data type
EarlyAction	Applied for early action	Binary (1/0)
EarlyDecision	Applied for early decision	Binary (1/0)
MeritScholAmt	Merit scholarship amount awarded	Binary (1/0)
FinAidRating	Financial aid rating	Categorical encoded as binary (1,0)
HSTier	High school tier	Categorical encoded as binary (1,0)
Foreign	Foreign student	Binary (1/0)
FAFSA	Applied for federal student aid	Binary (1/0)
APCourses	Took AP courses	Binary (1/0)
Sex	The sex of the student	Binary (1/0)
Athlete	Is a student athlete	Binary (1/0)
EarlyDeferral	Applied for early deferral	Binary (1/0)
WaitlistYN	Was waitlisted	Binary (1/0)
Commute	Is a commuter student	Binary (1/0)
HSGPA	High School GPA	Integer
DistanceInMiles	Distance from home (miles)	Integer
School	Member of a certain school, e.g., CC (ComSci & Math)	Categorical encoded as binary (1,0)
IsPellRecipient	Is recipient of Pell Grant	Binary (1/0)
IsDeansList	Joined Dean's List	Binary (1/0)
IsProbation	Is on probation	Binary (1/0)
OccupantsBuilding	Number of occupants in dorm	Integer
OccupantsRoom	Number of occupants in dorm room	Integer
IsSingleRoom	Uses a single room	Binary (1/0)
IsUnlimitedMealPlan	Has unlimited meal plan	Binary (1/0)
PercentHigherEd	Percent of those with higher education in home area	Float
GiniIndex	Gini Index value of home area	Float
MedianIncome	Median income of home area	Float
PercentWithInternet	Percent with internet in home area	Float
Attrited (Target)	Left the college	Binary (1/0)

"GiniIndex." As mentioned previously, "DistanceInMiles" is the amount of miles between the college and the student's hometown. "OccupantsBuilding" is the number of students that live in a student's dorm building. Similarly, "OccupantsRoom" is the number of students that live within the student's dorm room including themselves. Last, "GiniIndex" is the Gini coefficient of the student's hometown which is a measurement of income distribution in the area where a high value indicates greater inequality. In order to handle these features, the data is cleaned.

Our method of preprocessing the data includes removing any feature that is comprised of more than 30% of nulls and imputing the remaining features with missing values using K nearest neighbors (KNN). Additionally, the preprocessing step also included normalizing values between 0 and 1 for all integer and float type features. All categorical features mentioned in Table 1 are dummified.

After preprocessing the data, it is used to perform three experiments as described in the next few sections.

4.1 Experiment 1

In the first experiment, the GAN model was trained for 10,000 epochs. The weights were then transferred to two models, one that is tuned for 500 epochs to classify student attrition and the other that is left alone. This transference of weights is an example of transfer learning where the knowledge gained through adversarial training is applied to predicting student attrition (further details can be found in the referenced work) [6]. A control model was made from the same architecture as the GAN one, but trained separately on only the data previously used to tune for classification for 500 epochs. From the Receiver Operating Characteristic (ROC) diagrams, the control model performed marginally better with an accuracy of 0.68 than the tuned GAN model with only 0.64 accuracy. A ROC curve is a plot of the true positive rate against the false positive rate across various thresholds that determine the dividing line between classifications for a given model (more info in the provided reference) [2]. The accuracy of the GAN model, before tuning, is extremely low at 0.42. It is important to also note the discriminator and generator loss converging at about 10,000 epochs, or around the amount of epochs this experiment ran.

Directing our attention to the Cohen's kappa statistic, we observed that the relationship between the control and tuned model shows a kappa value of 0.5301 when a threshold resulting in about a 5% error rate is used. This value could be in the range of -1 to 1 and shows how close the model's outputs are where 1 is identical and anything 0 or below is akin to equivalent by chance [7]. The formula for the Cohen's kappa coefficient is the following:

$$\kappa = (p_o - p_e) \,/\, (1 - p_e)\,. \tag{3}$$

The p_o variable in the equation is the observed agreement of the labels applied to a sample by the models while p_e is the probability of chance agreement. The aforementioned value 0.5301 demonstrates that the control and tuned models for this experiment are outputting predictions that are similar but also having a good number of discrepancies. The fact that they are different suggests that the models are fundamentally different in their output distributions which is desired. In the second experiment, we will see how the Cohen's kappa coefficients change.

4.2 Experiment 2

In the second experiment, the GAN model trained for 15,000 epochs. We observed that the discriminator and generator losses converged and begun separating again though on inverse sides. The GAN model, before tuning, still demonstrates a low accuracy and a latent space with a similar linear relationship as in experiment 1. The control model's accuracy remains at about 0.68 with 500 epochs of training. It is here that we see an improvement in the accuracy of the tuned model boasting a 0.69 which is 0.05 higher than its previous. Last, the kappa value for the control and tuned model is 0.4426 which is lower than in the first experiment when ran with a threshold resulting in about a 5% false positive rate despite the overall accuracy of the two models being different by a 0.01 margin. This means that despite their close accuracies, the two models are providing differing outputs which suggests the two models are correctly classifying students the other is misidentifying. The third and final experiment will demonstrate what happens to the Cohen's kappa coefficient when the accuracy of the tuned model is higher than the control model.

4.3 Experiment 3

In the third experiment, the GAN model trained for 20,000 epochs. We chose 20,000 as the largest amount of epochs for an experiment due to the losses converging at about 10,000 epochs and to see how well the model performed with a large number of epochs at about double the point of loss convergence. The loss and kappa statistic results are shown in Fig. 3. As shown in (a), the discriminator and generator losses converged, separated, and continued to grow apart though on inverse sides to where they began. When we take a look at the kappa score in (b), where the control and tuned models are predicting at a threshold resulting in about a 5% false positive rate, it is higher than the previous two experiments. Here, we see that their output similarity is measured to be a 0.6358 kappa score. This increase in the kappa score is expected since both models have an increased accuracy from the previous two experiments which naturally leads to their outputs being similar as they both are making more correct predictions. While this value is higher than in experiment 1 and 2, it still demonstrates the predictions of the two models show a noteworthy degree of discrepancy and produce different output distributions.

Figure 4 shows the GAN model before tuning. In (a) observe a low accuracy though now with a noticeably different latent space, (b), that seems to still have some semblance of a linear relationship with a high amount of data clumping at the bottom left corner and some at the top right corner. This can be explained by the nature of hyperbolic tangent activation function which aims to pull separate classes into opposite sides of the quadrants.

Figure 5 shows the control model, which was able to reach an accuracy of 0.69 with 500 epochs of training (a). However, it is still 0.01 below the tuned model in