USING DEEP LEARNING IN THE EARLY PREDICTION OF AT-RISK LEARNERS IN SELF-PACED ONLINE EDUCATION

by

Hajra Waheed

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science

Examination Committee: Dr. Saeed Ul Hassan (Supervisor) Dr. Naif Radi Aljohani (Co-Supervisor)

External Evaluation Committee Prof. Dr. Miltiadis D. Lytras (Greece) Prof. Dr. Mohamed Gaber (United Kingdom)

Scholarship Donor: Information Technology University, Pakistan

Information Technology University Computer Science Department Pakistan January 2022
Acknowledgments

Amongst the many people that I would like to thank, my foremost gratitude goes to my supervisor, Dr. Saeed Ul Hassan, who played an instrumental role in formulating my research path. With his immense support, positive words of encouragement and invaluable expertise, he has been a constant source of inspiration throughout my PhD journey. His dedication and devotion is exemplary, inculcating a positive attitude.

My sincere thanks to the Information Technology University, for providing me the funding to conduct my research. I would also like to express my gratitude to my co-supervisor Dr. Naif Radi Aljohani and my collaborators Dr. Raheel Nawaz and Dr. Dragan Gasevic for their invaluable suggestions throughout my PhD journey, providing me with the opportunities to reach my research objective.

In addition, on a personal note, my sincere thanks to my parents for their wise suggestions, unwavering support and constructive counseling throughout my PhD journey. My brothers, for putting up with me and supporting me throughout this time, especially in my low phases. Finally, my sincere gratitude to my friends and lab members for their stimulating discussions, inspiring me to constructively focus on my goals.

I thank my Almighty ALLAH for providing me a guiding light in the form of these people, without whom this thesis would not have been possible.
Abstract

Learning Analytics is an emerging field of research motivated by the wide spectrum of educational information available for analysis to provide data-driven decisions on various learning problems. Exploring the research landscape to deliver a comprehensive understanding of this multidisciplinary field’s activities, this dissertation focuses on the various dimensions of the students academic performance prediction. To meet the demands of modern education and its increasing flexibility, many higher education institutions are considering online self-paced education programs. Especially recently, due to the pandemic, teaching and learning mechanisms have been revolutionized and adequate tools and techniques are essential; however, a widely recognized challenge in such education is student retention. The abundance of accessible educational data presents an opportunity to investigate students learning behavior, addressing their issues, optimizing the educational environment and enabling data-driven decision-making.

In contrast to predicting the academic performances of students at the course end using legacy data, its early prediction is an opportunity for educational stakeholders to establish early warning systems and instructional interventions. This dissertation analyses both these approaches using deep learning-based techniques with data from the Open University UK. Further, it explores the learner prediction problem in a temporal setting so that learners’ engagement patterns are transformed into a weekly and quarterly granularity. The results demonstrate the effectiveness of the deep LSTM approach, outperforming conventional machine learning approaches with an increase in accuracy of up to 4.66% in the first five weeks: on a quarterly basis, an increase of up to 3.46% is observed by applying artificial neural networks. To interpret the modeling methods, the shapely values-based Shapely Additive Explanation method is used to identify the factors impacting the academic performances of students. The interaction with assessments and quizzes illustrates a significant attribute in the model, relative to other activities.

The issue of class imbalance is prevalent in the learning analytics community, and its significance is accentuated in a temporal setting; however, a research gap exists in the up-sampling of data in such a setting where each up-sampled student’s behavior should be encapsulated in time sequences. The dissertation proposes an adversarial-based approach that up-samples those students who failed a course by encapsulating their behavior over the course’s duration. The results of the proposed adversarial-based method show its effectiveness, with an increase in accuracy of up to 5.33% over conventional adversarial networks. The novel approach can serve as a baseline for the learning analytics community to eliminate class imbalance in a
quarterly setting and to improve the performance of machine learning models.

The implications of this research are not restricted to the educational community; other disciplines dealing with temporal human behavior can adapt the proposed temporal adversarial method. Moreover, the early prediction of students at risk of failure can assist instructors to develop early warning systems, enabling them to provide support to students in-need, devising corrective strategies and offering interventions.
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Title Page</td>
<td>i</td>
</tr>
<tr>
<td></td>
<td>Acknowledgments</td>
<td>ii</td>
</tr>
<tr>
<td></td>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td></td>
<td>Table of Contents</td>
<td>v</td>
</tr>
<tr>
<td></td>
<td>List of Figures</td>
<td>viii</td>
</tr>
<tr>
<td></td>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Background</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Learning Analytics</td>
<td>2</td>
</tr>
<tr>
<td>1.2.1</td>
<td>Related Terms and Semantic Mapping of Learning Analytics</td>
<td>2</td>
</tr>
<tr>
<td>1.3</td>
<td>Evolution of Learning Analytics</td>
<td>5</td>
</tr>
<tr>
<td>1.3.1</td>
<td>Deep Learning and Learning Analytics</td>
<td>6</td>
</tr>
<tr>
<td>1.4</td>
<td>Thesis Focus</td>
<td>7</td>
</tr>
<tr>
<td>1.5</td>
<td>Statement of the Problem</td>
<td>9</td>
</tr>
<tr>
<td>1.6</td>
<td>Research Objectives</td>
<td>10</td>
</tr>
<tr>
<td>1.7</td>
<td>Contributions</td>
<td>11</td>
</tr>
<tr>
<td>1.7.1</td>
<td>Bibliometric Analysis on Learning Analytics</td>
<td>12</td>
</tr>
<tr>
<td>1.7.2</td>
<td>Leveraging Students’ Behavioral Characteristics from VLEs</td>
<td>12</td>
</tr>
<tr>
<td>1.7.3</td>
<td>Influential Attributes in Predicting Academic Performances during On-going Courses</td>
<td>13</td>
</tr>
<tr>
<td>1.7.4</td>
<td>Balancing the Award-gap Imbalance in Students’ Learning</td>
<td>14</td>
</tr>
<tr>
<td>1.8</td>
<td>Main Findings of this Thesis</td>
<td>14</td>
</tr>
<tr>
<td>1.9</td>
<td>Thesis Outline</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>Literature Review</td>
<td>18</td>
</tr>
<tr>
<td>2.1</td>
<td>Self-paced Online Learning</td>
<td>18</td>
</tr>
<tr>
<td>2.1.1</td>
<td>Predictive Modeling in Learning Analytics</td>
<td>19</td>
</tr>
<tr>
<td>2.2</td>
<td>At-risk Student Prediction at the End of a Course</td>
<td>20</td>
</tr>
<tr>
<td>2.3</td>
<td>Early Performance Prediction of Students At-risk</td>
<td>23</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Predictive Modeling Approaches in Self-paced Online Learning</td>
<td>25</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Important Predictors of Student Success in Self-paced Online Learning</td>
<td>27</td>
</tr>
<tr>
<td>2.4</td>
<td>Class Balancing of Students Logs in Self-paced Online Learning</td>
<td>28</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Review on Up-Sampling Techniques in Educational Data Science</td>
<td>28</td>
</tr>
<tr>
<td>2.4.2</td>
<td>GANs for Temporal Data Generation</td>
<td>30</td>
</tr>
</tbody>
</table>
Bibliometric Analysis of Learning Analytics

3.1 Bibliography Dataset for Learning Analytics
3.1.1 Data Pre-processing for Bibliometric Analysis
3.2 Bibliometric Indices
3.3 Results & Analysis
3.3.1 Publication Output of Countries and Institutions in Learning Analytics
3.3.2 Discussion on Collaboration Network among Authors and Countries
3.3.3 Discussion on Citation Network among Sources
3.3.4 Discussion in Terms of Co-occurrence Network
3.3.5 Discussion on Recent and Most Significant Papers in Learning Analytics
3.3.6 Overall Discussion on the Bibliometric Study
3.4 Summary

Academic Performance Prediction of Students and their Behavior

4.1 Dataset for Academic Performance Prediction of Students
4.1.1 Descriptive Analysis of the Dataset
4.2 Data Pre-processing
4.2.1 Demographics and VLE Portal Data
4.3 Artificial Neural Networks (ANN)
4.4 Results: Experimental Setup
4.4.1 Results using Demographics and VLE Portal Data
4.5 Summary

Early Performance Prediction of Students in Self-paced Online Learning

5.1 Dataset for Students’ Early Performance Prediction
5.1.1 Data Processing for Quarterly VLE Clickstream Data
5.1.2 Data Processing for Week-wise Student Engagements
5.2 Handling Class Imbalance
5.3 Modeling Approaches for Students Performance Prediction
5.3.1 Conventional Machine Learning Models
5.3.2 Time-series Data Modeling
5.3.3 Evaluation of Feature Significance for Performance Prediction
5.3.4 Evaluation
5.4 Results for Students’ Early Performance Prediction
5.4.1 Academic Performance Prediction at the End of a Course
5.4.2 Early Performance Prediction in a Quarterly Setting
5.4.3 Early Performance Prediction in a Week-wise Setting
5.4.4 Identifying Features impacting Students’ Academic Performance
5.5 Summary
6 Balancing the Award-Gap Imbalance in the Academic Performances of Students

6.1 Up-Sampling Time-Series Data
6.2 Dataset for Balancing Students Logs
6.3 ED-GAN for Temporal Synthetic Data Generation
   6.3.1 Conditional Vector
   6.3.2 Generator Loss
6.4 Results: Experimental Setup
   6.4.1 Baseline Methods
   6.4.2 Proposed Methods for Up-Sampling Temporal Data
   6.4.3 Experimental Setup for Proposed Temporal Models
   6.4.4 Evaluation of Proposed Temporal Models using OULAD
   6.4.5 Results & Discussion
6.5 Summary

7 Conclusion and Recommendations

7.1 Value Added
7.2 Limitations
7.3 Future Directions

References
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Semantic Mapping of Educational Analytics with Learning Analytics</td>
<td>4</td>
</tr>
<tr>
<td>1.2</td>
<td>Deep Learning for Educational Data Science</td>
<td>8</td>
</tr>
<tr>
<td>3.1</td>
<td>Document Types</td>
<td>35</td>
</tr>
<tr>
<td>3.2</td>
<td>Bibliographic Coupling and Co-citation Method</td>
<td>37</td>
</tr>
<tr>
<td>3.3</td>
<td>Publication and Citation Counts of Countries in Learning Analytics</td>
<td>40</td>
</tr>
<tr>
<td>3.4</td>
<td>Evolution of Learning Analytics Research</td>
<td>42</td>
</tr>
<tr>
<td>3.5</td>
<td>Institute-wise Publications in Learning Analytics from 2000 to 2017</td>
<td>43</td>
</tr>
<tr>
<td>3.6</td>
<td>Publication Output of Countries in Learning Analytics from 2008 to 2017</td>
<td>46</td>
</tr>
<tr>
<td>3.7</td>
<td>Publication/Citation Counts of Authors and Co-authorship Network in Learning Analytics</td>
<td>47</td>
</tr>
<tr>
<td>3.8</td>
<td>Co-authorship Network of Learning Analytics Among Nations from 2008 to 2017</td>
<td>49</td>
</tr>
<tr>
<td>3.9</td>
<td>Citation Network Among Source Titles</td>
<td>51</td>
</tr>
<tr>
<td>3.10</td>
<td>Title-based Thematic Analysis on Learning Analytics Keywords</td>
<td>52</td>
</tr>
<tr>
<td>3.11</td>
<td>Author-defined Keyword Evolution in Learning Analytics</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>OULA Dataset Descriptives for Age, Gender, Highest Education &amp; Imd-Band</td>
<td>61</td>
</tr>
<tr>
<td>4.2</td>
<td>Heat-map of all Feature Activities with their Feature Codes</td>
<td>63</td>
</tr>
<tr>
<td>4.3</td>
<td>Proposed Architecture using Artificial Neural Network for Predicting Student Performance at the End of a Course</td>
<td>71</td>
</tr>
<tr>
<td>4.4</td>
<td>Correlation Heat-maps among Features and Class Categories</td>
<td>73</td>
</tr>
<tr>
<td>5.1</td>
<td>Data Processing Steps for the Early Prediction Model using LSTM</td>
<td>81</td>
</tr>
<tr>
<td>5.2</td>
<td>Dataset Distribution for Each Feature Activity with Dark as ‘Pass’ and Light as ‘Fail’ Class</td>
<td>90</td>
</tr>
<tr>
<td>5.3</td>
<td>Quarterly Accuracy for each Category</td>
<td>93</td>
</tr>
<tr>
<td>5.4</td>
<td>Quartiles Learning Accuracy Curves for all Categories</td>
<td>95</td>
</tr>
<tr>
<td>5.6</td>
<td>ANN’s evaluation with the baseline methods for Q1-Q4</td>
<td>96</td>
</tr>
<tr>
<td>5.5</td>
<td>Quartiles Learning Loss Curves for all Categories</td>
<td>97</td>
</tr>
<tr>
<td>5.7</td>
<td>LSTM Validation Accuracy, Loss, Precision and Recall Curves for 5th, 10th, 20th, 30th &amp; 38th weeks at 50th epoch</td>
<td>101</td>
</tr>
<tr>
<td>5.8</td>
<td>Using SHAP to Identify Important Features impacting Performances of Students</td>
<td>103</td>
</tr>
<tr>
<td>6.1</td>
<td>Up-Sampling Time-Series Data</td>
<td>111</td>
</tr>
<tr>
<td>6.2</td>
<td>Data Pipeline for Temporal GAN-based Up-Sampling Methods</td>
<td>113</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Scopus-Compatible Query String for Bibliometric Study</td>
<td>34</td>
</tr>
<tr>
<td>3.2</td>
<td>Top 10 Countries in Learning Analytics w.r.t. Publication Counts</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>from 2000 to 2017</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>Top 15 Institutions in Learning Analytics with Publication Counts</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>from 2000 to 2017</td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td>Top 10 highly Cited Publications in Learning Analytics</td>
<td>55</td>
</tr>
<tr>
<td>3.5</td>
<td>Top 10 Recent Publications in Learning Analytics</td>
<td>57</td>
</tr>
<tr>
<td>4.1</td>
<td>Class Label Categories</td>
<td>64</td>
</tr>
<tr>
<td>4.2</td>
<td>Features Representing Before and After Course Activities of Students</td>
<td>64</td>
</tr>
<tr>
<td>4.3</td>
<td>Evaluation of the Results of Students Demographics &amp; VLE Portal Data</td>
<td>72</td>
</tr>
<tr>
<td>5.1</td>
<td>Description of each Activity Feature in OULA Dataset, Total Clicks</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>and Average Clicks per Class</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>Classification Results on the Balanced Dataset for Pass/Fail at the</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>End of a Course Duration</td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>Quarterly Accuracy and Loss for each Category</td>
<td>92</td>
</tr>
<tr>
<td>5.4</td>
<td>ANN’s Evaluation with the Baseline Methods for Q1-Q4</td>
<td>96</td>
</tr>
<tr>
<td>5.5</td>
<td>Comparison of ML Classifiers and LSTM Week-wise Results</td>
<td>99</td>
</tr>
<tr>
<td>6.1</td>
<td>Mean L1-Norm for all GAN-Variants</td>
<td>144</td>
</tr>
<tr>
<td>6.2</td>
<td>Comparison of the Proposed GAN-based Models with Various Over-sampling</td>
<td>147</td>
</tr>
<tr>
<td></td>
<td>Techniques on Neural Network</td>
<td></td>
</tr>
<tr>
<td>6.3</td>
<td>Quarter-wise Machine Learning Efficacy for EDD-GAN</td>
<td>148</td>
</tr>
<tr>
<td>6.4</td>
<td>Quarter-wise Machine Learning Efficacy for EDC-GAN</td>
<td>148</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

“Everything is theoretically impossible, until it is done.”
— Robert Heinlein

The proliferation in the online learning has enabled the emergence of several online educational platforms to assist students in their learning. The changing landscape in the global economy has resulted in many governments setting priorities for higher education institutions that aim to enhance personalized support for learners and increase the number of graduates [1, 2]. These priorities are often not matched by an increase in funding for higher education. On the contrary, higher education institutions are often confronted by cuts in government funding. All these put the pressure on such institutions to develop programs for education innovation [3], transform their existing learning and teaching practices [4], introduce flexible learning opportunities (e.g., self-paced learning), and embrace novel technological innovations [5]. However, all this comes with potential risks associated with student success. Therefore, higher education institutions are also in need of looking for solutions that can address some of those risks and support learners in achieving their goals.

1.1 Background

Rapid advancements in online educational platforms have yielded sufficient educational repositories, demonstrating significant impact on higher education by providing opportunities to address learner behavior and optimizing these behaviors to enhance the learning environment [6]. The progression of the accumulated educational data has stimulated the emergence of several research communities, such as learning analytics to predict learners behavior, providing indicators for optimized policy formulations. Educational data, a by-product of the interaction between learners and instructors, has been substantiated as a multidisciplinary field of study, involving researchers from various research disciplines. Therefore, many terms are associated with the analysis of educational data, for instance learning analytics, academic analytics, educational data mining, predictive analytics and learners’ analytics. The commonality in all of these terms is the use of different or similar sets of educational
data for many purposes. Recently, a new term has been introduced: ‘educational data science’, which clarifies how disciplines and researchers with different research interests and backgrounds can work in this field [7].

1.2 Learning Analytics

Learning analytics constitutes of collecting, examining, inspecting, analysing and reporting on learners’ data to improve the learning experience and environment in order to optimize learners’ and instructors’ performance [8]. Another interpretation of learning analytics lays emphasis on measuring and understanding learners’ performance on an individual basis and how it impacts on the institute’s overall conduct [9, 10]. Furthermore, learning analytics is used in the context of predicting new insights not only to enhance learning and teaching experiences but develop strategies for optimizing the efficiency of institutes at an academician level, promoting effective decision-making [11, 12]. Khalil and Ebner [13] described it as an analysis technique applied to the educational data stream to infer patterns for improving and elevating students’ performance and assisting in teaching mechanisms.

The integration of learning analytics to advance decision-making in higher education, requires a deep understanding of the multiple facets of the phenomenon. In fact, the investigation of the demand and supply sides of the learning analytics flow within higher education institutes presents many unforeseen challenges relating to education administration, policy making, innovation and teaching and learning excellence. For the development of the learning analytics scientific community, it is mandatory to explore interesting and distinct paths by developing a clear understanding of the research activity in this field, visualizing the various dimensions of learning analytics [14] and its impact on the learning community and environment.

1.2.1 Related Terms and Semantic Mapping of Learning Analytics

Learning analytics has been receiving enormous attention from policy makers and the research community ever since the term was coined in 2011. It resonates strongly with educational data mining, which is a multidisciplinary term applying various methods, including machine learning, data-mining
techniques, information extraction, statistical analysis and pedagogical methods, to investigate, ex-
plore and resolve educational-related issues [15]. By employing data-mining techniques on an educa-
tional dataset, an insight is generated at the level of interaction among learners/students and instruc-
tors, students’ attitude to learning and how their behaviour and interaction impacts on learning and
grades. It assists in understanding the pre-requisites of an educational environment, improving the
efficiency of an institute [16].

The trend of using the internet for learning and educational purposes has yielded a collection of large
datasets of students and learners information. Further, with everything being computerized, e-learning
has been immensely developed and established, providing sufficient repositories for employing data-
mining techniques [17, 18]. With the advent of internet e-learning, collaborative online learning
and the newly emerged concept of Massive Open Online Courses (MOOCs) have become immensely
popular [19], generating datasets that can be analysed for predicting patterns in an educational setting,
understanding the relationships between the actors involved in such settings and how these can be
used to improve and optimize learning and learning environments [20, 21].

Learning analytics, being a multidisciplinary field, is akin to educational data mining. Ever since its
emergence in 2011, it has been evolving and various dimensions have become a part of it, including
academic analytics and action analytics. Academic analytics relates to a holistic view of educational
data analytics from an institutional perspective, and relates to improving policy-level strategies [8].
It is associated with business intelligence, improving an institute’s decision-making process and opti-
mizing operations such as its recruitment and administrative processes [22].

In higher education, learning analytics is used cumulatively with concepts of academic analytics,
helping institutions in their business, economics and finance sectors by optimizing learning outcomes,
understanding student behaviour and suggesting corrective strategies for instructors that ultimately
make an institute stable, maintaining an appropriate resource allocation. Lately, with increased com-
petition between universities, it has become imperative for an institute to focus on student retention,
reducing attrition and subsequently increasing graduation rates [23].

Aljohani [21] demonstrated the overlap between learning and academic analytics as ‘educational ana-
lytics’, as demonstrated in Fig.1.1. This may be attributed to the fact that both learning and academic
analytics deal with learners and instructors, aiming to optimize their learning processes. However,
they do differentiate on the terms; Academic analytics provides a holistic view of the overall insti-

Figure 1.1: Semantic Mapping of Educational Analytics with Learning Analytics
tution optimization, whereas learning analytics works on improving learners’ experiences, indirectly impacting on the institute’s overall performance. Hence the former takes the institute as a direct stakeholder, involving decision-making at an operational and administration level, while the latter impacts on it in an indirect manner. Cumulatively, both provide a better understanding of the educational domain, assisting the learning environment, producing positive outcomes for both learners and instructors and optimizing an institution’s decision-making [20].

1.3 Evolution of Learning Analytics

The learning analytics discipline assists universities to maintain their reputation by predicting future outcomes. Many studies conducted from this perspective produce significant and successful results, constructing analytical frameworks and highlighting the importance of learning analytics in higher education [24]. However, the collection of student records for gaining insights has prompted debate on the ethical and legal issues, tightly interconnected to privacy, trust and accountability [25, 26, 27]. It aids in improving and optimizing inter-institutional cooperation by setting corrective strategies for the instructor and learner community, encouraging them to work in a cohesive manner, producing positive outcomes [28].

Learning analytics has also been defined as a tool for analysing the quantitative data gathered from online learning systems such as Learning Management Systems (LMS), MOOCS, Virtual Learning Environments (VLE) and other online tutoring systems [29]. In recent years, with the increase in online education systems, extensive repositories are maintained by the institutions as a by-product [11]. For instance, when students use LMS [21], Intelligent Tutoring systems (ITS) [30], or other online platforms [31], they leave behind trails of data that can be analysed for predicting future outcomes, such as predicting on-risk students, determining the behavior of intellectual students, improving teamwork assessments [29] and suggesting various courses of actions depending on individual student’s performance [32]. Learning analytics makes use of statistical analysis to categorize students on the basis of their current performances, revealing patterns of on-risk and successful students, suggesting possible outcomes and predicting potential problems before time [11, 20]. It has also proved to be a measure of assistance for teachers/tutors and institutions, intervening in the learning environment and providing a better means of teaching by tailoring teaching methods and producing successful results.
Consequently, this assists an institute to retain students, improve overall business by reducing student attrition, focus on at-risk learners, improve learning outcomes and provide instructors with substantial, optimized guidelines to enhance teaching criteria [35, 36, 37].

Khalil and Ebner [13] proposed a learning analytics lifecycle constituting of four phases, referring to learners, instructors and the educational institute as ‘stakeholders’ contributing to the construction of large repositories of educational data known as ‘Big Data’. The collected data is processed by applying statistical and other analytic techniques to make predictions and recommendations for the future, optimizing key stakeholders’ outcomes. While learning analytics emphasizes improving the learning process, all stakeholders benefit from it. Aljohani and Davis [38] emphasized developing learning analytic strategies to assess the performance and learning behavior of mobile learners. This area is sparsely investigated in learning analytics, consequently it is emerging as an upcoming area for future research [38, 39]. As an extension to this work, they developed an app for collecting immediate feedback from learners after each class, aiding instructors to evaluate the cognitive understanding of learners [40].

### 1.3.1 Deep Learning and Learning Analytics

Learning analytics comprises of techniques to improve the learning process. Artificial Neural Networks (ANNs) are the most prominent practice utilized in the Educational Data Mining (EDM) domain [41]. Although there have been issues associated with ANNs, especially when extracting human-interpretable patterns from the predicted results, most of these concerns were resolved in the last decade, with the emergence of Deep ANNs [41, 42]. Deep Learning, evolved from machine learning, and characterized by numerous computational layers, enables the model to learn from examples, superseding the traditional techniques of hand-engineering the features [43].

Contrary to the increasing rise in Deep Learning techniques, sufficient evidence of Deep ANNs in EDM and learning analytics literature is not available. Coelho et al. [41] conducted a systematic literature review to investigate the evidence found in learning analytics and deep learning studies. They identified student performance [44, 45, 46], student assessment [47] and handwriting recognition [48], as being some of the areas where deep learning was deployed, proving better than the baseline models. Numerous models have been explored in the learning analytics research paradigm, however
identifying the significance of deep learning in the learning analytics domain is still in its infancy, with studies on the adoption of this technique emerging in the last few years.

In higher education, learning analytics is defined consistently with notions of academic analytics, aiding institutions in their financial strength by reducing attrition rates, improving learning outcomes by considering learners behavior, recommending corrective policies for instructors that eventually yield in establishing a stable institute and exercising a suitable resource allocation method. Student retention has become a standard strategic imperative for institutions, and the learning analytics phenomena cumulatively aids in retaining students, consequently resulting in accumulated graduation rates [23]. Academic and learning analytics consistently overlap in formulating the ‘Educational Analytics’ paradigm - where learning analytics is associated with the learner’s experience and academic analytics implicitly incorporates the overall institute and its performance [49]. Moreover, a semantic mapping of deep learning with educational data science is presented in Fig.1.2, constituting of the objectives related to educational data science and the tools/techniques associated with machine learning that are deployed in the existing literature. The paradigm of educational data science encompasses the various overlapping educational analytics terms such as Learning analytics, Educational analytics and Academic analytics, as depicted in Fig.1.2. Deploying deep learning techniques enables to capture the students’ learning perception in a more rigorous manner, facilitating longitudinal interventions by the academia, and tailoring teaching methods to improve learner’s experience.

1.4 Thesis Focus

To address the demands of modern education and increase flexibility, many higher education institutions are considering self-paced education programs. However, student retention is yet a widely recognized challenge faced in self-paced education [50, 51]. While many studies have examined the potential of the use of data about student interaction with learning technologies to predict student success, studies that focus on self-paced education are scarce [11, 37]. The identification of students that are at-risk of a failure in higher education is connected to numerous educational policies, to enhance their competencies and skills through timely interventions by the academia. We intend to employ deep learning techniques to early predict the students at-risk to enable the development of an early warning system and to inform timely interventions. Moreover, we intend to identify specific factors
Deep Learning for Educational Data Science

**Objectives**
- Academic Engagements
- Learner’s Intuition
- Longitudinal Interventions

**Tools & Techniques**
- Statistical Analysis
- Machine Learning
  - Classification
  - Regression
- Deep Learning
  - Deep ANN
  - LSTM

---

**Learning Analytics**

**Objectives**
- Student Progress
- Improving Learner’s Experience
- Improve Collaboration

**Tools & Techniques**
- Statistical Analysis
- Machine Learning
  - Classification (Pass-Fail)
  - Regression (Scores)

---

**Educational Analytics**

**Objectives**
- Student Success
- Academic Difficulty
- Customizing Courses
- Tailoring Teaching Methods
- Student Retention

**Tools & Techniques**
- Machine Learning
  - Classification
  - Regression

---

**Academic Analytics**

**Objectives**
- Administration Level
  - Policies
  - Guidelines
- Student Affordability
- Student Access
- Resource Allocation

**Tools & Techniques**
- Machine Learning
  - Classification (SVM, Naïve Bayes, Decision Tree)
  - Regression

---

*Figure 1.2*: Deep Learning for Educational Data Science
that are the highest predictors of students at-risk in order to assist the development of pedagogical interventions.

To predict the student performances during an on-going course, the engagement information of the students is transformed into a sequential format. These sequences of interactions enable the machine learning modes to early identify the students at-risk of a failure. In the literature, several studies exist that focus on early predicting the students at-risk [52, 53, 54]. To cater the imbalance in the early prediction of academic performances of students, the task becomes more complex due to the sequential nature of the data, where each time sequence represents a particular student and several sequences formulate the overall engagement of that student over the duration of a course. A major limitation in the learning analytics community is the hindrance of balancing the sequential data of students. The conventional methods for generating synthetic data do not account for the association of sequences, where each synthetically generated student is associated with its previous sequence. In this dissertation, we intend to fill this research gap by employing adversarial techniques.

1.5 Statement of the Problem

This research study intends to exploit the temporal sequential classification problem of predicting students performances in a self-paced education environment. VLEs are one such infrastructure that complement the self-paced education environment and the learning analytics paradigm by effectively providing datasets for analyzing and reporting students’ learning process, assisting in disclosing their learning approaches, and its reflection and contribution in their respective performances. With the increase in the self-paced learning phenomenon, class imbalance in the learning analytics community has become more prevalent with lesser students dropping out in comparison to students that intend to pursue. Similarly, an imbalance in the behavioral performances of students is also observed. This imbalance problem becomes more complex for temporal data, where each time sequence represents a particular student and several sequences formulate the overall engagement of that student over the duration of a course. The conventional methods for generating synthetic data do not account for this association of sequences, where each synthetically generated student is associated with its previous sequence. This study proposes a novel method based on adversarial networks, that encapsulates the behavior of the up-sampled students throughout the sequences and retains their previous behavior to
predict the behavior for next sequences. More specifically, the problem statement of this research is as follows:

*This dissertation intends to exploit the temporal sequential classification problem of predicting the academic performances of students in a self-paced education environment and eliminating the class imbalance, of the award-gap performances, in a temporal setting.*

### 1.6 Research Objectives

This section formalizes the objectives of this dissertation, described as below:

- The first objective of this dissertation is to understand the multi-facet dimensions of the learning analytics community, by conducting a bibliometric research to examine the world’s research landscape in this discipline. Through qualitative and quantitative bibliometric techniques, the study intends to present a comprehensive view of the new era of learning analytics which, in fact, is multidisciplinary, flexible, dynamic and powered by a new, sophisticated series of computational tools, enabled by cognitive computing and Big Data techniques. State-of-the-art bibliometric indices are deployed, such as co-authorship network, citation networks and terms co-occurrence networks to quantitatively study the multidisciplinary field of learning analytics over time, specifically in terms of publication and citation counts. Moreover, studying collaboration network patterns with respect to institutions and authors, analysing the temporal thematic evolution of the field and exploring the tools and techniques being deployed for data-driven decision-making in the learning analytics research space.

- The second objective of this dissertation leverages behavioral characteristics of students in terms of their performance predictability and capability through an array of machine learning algorithms, to predict the award-gap in performances among students. The effectiveness in the prediction of students’ academic performances is examined for the following categories: ‘withdrawn-pass’, ‘pass-fail’, ‘distinction-pass’ and ‘distinction-fail’ using VLE datasets. Exploiting clickstream data generated through students interaction with the virtual learning management system to assess their performance.

- The third objective is to explore the effectiveness of deep learning models in the early predic-
tion of students at-risk of a failure and determine the effectiveness of deep learning models in predicting students at-risk using VLE dataset. It comprehensively compares how deep learning algorithms perform vs. other conventionally used machine learning techniques to early predict students at risk of failing a course. The data is transformed using smart data processing into a sequential week-wise format for a more granular specific and in-depth analysis of student behavior. It also identifies specific factors that are the highest predictors of students at risk in order to inform the development of pedagogical interventions. Moreover, predicting students at risk as early in a course as possible to enable the development of an early warning system and to inform timely interventions.

- Lastly, the class imbalance issue in the learning analytics discipline is examined, where a significant award-gap in academic performances of students is observed. Generally the number of students passing in a course will be multiple times more than the number of students demonstrating poor performances. This class imbalance issue becomes more complex for sequential time series problem, where the existing up-sampling techniques are unable to handle the complexity of the temporal data such that the behavior of the newly generated instances is encapsulated throughout the entire sequence. We intend to extend the existing Conditional-Generative Adversarial Networks (CGANs) and propose two variations of educational GANs that encompass the complexity of temporal data. This proposed method may also assist researchers in other disciplines dealing with temporal characteristics of the data.

1.7 Contributions

The proliferation in the online learning has enabled the emergence of several online educational platforms to assist students in their learning, revolutionizing the learning processes with the trend of increased analytical computations, inferring unique and significant characteristics impacting learners’ performances.
1.7.1 Bibliometric Analysis on Learning Analytics

Learning analytics is an emerging field of research, motivated by the wide spectrum of the available educational information that can be analysed to provide a data-driven decision about various learning problems. This contribution intends to examine the research landscape of learning analytics, to deliver a comprehensive understanding of the research activities in this multidisciplinary field, using scientific literature from the Scopus database. An array of state-of-the-art bibliometric indices is deployed on 2,811 procured publication datasets: publication counts; citation counts; co-authorship patterns; citation networks; and term co-occurrence. The results indicate that the field of learning analytics appears to have been instantiated around 2011, thus before this time period no significant research activity can be observed. The temporal evolution indicates that the terms ‘students’, ‘teachers’, ‘higher education institutions’ and ‘learning process’ appear to be the major components of the field. More recent trends in the field are the tools that tap into Big Data analytics and data-mining techniques for more rational data-driven decision-making services. A future direction research depicts a need to integrate learning analytics research with multidisciplinary smart education and smart library services. The vision towards smart city research requires a meta-level of smart learning analytics value integration and policy making.

1.7.2 Leveraging Students’ Behavioral Characteristics from VLEs

The abundance of accessible educational data provides opportunities to tap students’ learning behavior, addressing their issues, optimizing the educational environment and enabling data-driven decision making. VLE are one such infrastructure that complement the Learning Analytics paradigm by effectively providing datasets for analyzing and reporting students’ learning process, assisting in disclosing their learning approaches, and its reflection and contribution in their respective performances. The study intends to investigate the efficacy of the VLEs log dataset in predicting students’ performance by examining students’ behavior towards it. This includes their degree of involvement, extent of use, clickstream data and accessing these activities to measure their performance, in order to detect students at-risk of failure and predict withdrawals. This contribution deploys a deep ANN on the features, extracted from the VLEs clickstream data, to predict students on risk of a failure and providing measures for intervening the students on-time. More specifically, this study focuses on the
following four classes: Distinction, Pass, Fail and Withdrawal. Results show the proposed model to achieve an accuracy of 84%-93%. We show that deep ANN outperforms Logistic Regression (LR) and Support Vector Machine (SVM) models. While LR achieves an accuracy of 79.82% - 85.60%, the SVM achieves 79.95% - 89.14%. The inclusion of significant features tends to have an impact on the model’s accuracy, such as student’s legacy data including assessment submission time and number of assessments submitted. A contribution of this research is the informed novel approach to the advanced higher education decision-making towards sustainable education. It is a bold effort for student-centric policies, promoting the trust and the loyalty of students in courses and programs. In the near future there is a plan to introduce a framework for an ecosystem of educational policies geared towards this learning analytics layer.

1.7.3 Influential Attributes in Predicting Academic Performances during On-going Courses

To adapt to the demands of modern education and increase flexibility, many higher education institutions are considering self-paced education programs. However, student retention is yet a widely recognized challenge faced in online self-paced education. While many studies have examined the potential of the use of data about student interaction with learning technologies to predict student success, studies that focus on self-paced education are scarce. To address this gap in the literature, this contribution reports on the findings of a study that has investigated the performance of a well-known deep learning technique i.e., Long Short-term Memory (LSTM), in the prediction of students at-risk of failing a course offered in a self-paced mode of online education. This contribution has utilized a freely accessible Open University Learning Analytics (OULA) Dataset comprising 22,437 students with 69% pass, and 31% failed instances. The deep LSTM shows the highest predictive power to classify between pass and fail students, compared to all other alternatives by achieving an accuracy of 84.57%, precision of 82.24% and recall of 79.43%. Interestingly, with only first five weeks of course activity log data used for training, the receiver operating characteristic based diagnostic accuracy of the LSTM algorithm is achieved up to 71%, that outperforms almost all other conventional algorithms - despite trained on the complete dataset collected for the entire duration of the course i.e. up to 38 weeks. Furthermore, this study has also employed a shapely additive explanation model to
identify the most important predictors of student retention, e.g., assessment submission and attempted quizzes. This approach is essential in order to increase the interpretability of deep learning techniques and, thus, increase their potential to generate actionable insights.

1.7.4 Balancing the Award-gap Imbalance in Students’ Learning

With the increase in the self-paced learning phenomenon, class imbalance in the learning analytics community has become more prevalent with lesser students dropping out in comparison to students that intend to pursue. Similarly, an imbalance in the behavioral performances of students is also observed. This imbalance problem becomes more complex for temporal data, where each time sequence represents a particular student and several sequences formulate the overall engagement of students over the duration of a course. The conventional methods for generating synthetic data do not account for this association of sequences, where each synthetically generated student is associated with its previous sequence. For this contribution, we propose two GANs, specifically designed to encapsulate each student’s behavior over the entire course duration and thus generating instances that capture the data distribution through previous sequences, referred to as Educational (ED)-GANs. Furthermore, we utilize the Open University’s UK dataset, to demonstrate the effectiveness of our proposed ED-GANs. Through the inclusion of engagement information of students, the proposed GANs are trained to generate instances of minority class and augmenting the training set to eliminate class imbalance. The proposed networks are evaluated by deploying a series of conventional non-GAN based and some GAN-based up-sampling techniques. The results are compared with neural networks to demonstrate the effectiveness of the proposed GAN-based models. The ED-GAN models show an increase in accuracy of up to 5.33% compared to conventional GANs. The proposed adversarial networks can be served as a baseline for other communities with temporal longitudinal studies.

1.8 Main Findings of this Thesis

From the above deployed contributions, following are the main findings of this research.

- For each of the academic performance categories the engagement information collected af-
ter the module initiation was found to be significantly associated with student’s performance, implying that active participation in a module is positively associated with academic performance.

- In early prediction of academic performances, a sequential time-series format of students’ engagement produced more robust results, identifying at-risk students in-need of support and interventions, as early as the 5th week during an on-going course.

- Attempted quizzes and assessment submission participation behavior were found to be significantly impactful on the academic performances of students.

- In a sequential time-series setting, adversarial based techniques assist in up-sampling of the data of students such that each time-sequence of an up-sampled student is associated with its previous time-sequence.

- Introducing the students’ engagement homogeneous vector, as an additional conditional layer in GANs, assists the network in improved learning distribution of the students interaction patterns.

1.9 Thesis Outline

The rest of the dissertation is organized as follows in the chapters.

- In Chapter 2 the related work pertaining to existing studies on VLE datatsets are discussed in detail, encompassing the performance prediction of students, at the end of the courses and during an on-going course, and the associated attributes signifying their behavior. Moreover, an extensive review is also presented on the up-sampling techniques in learning analytics, to eliminate the award-gap in the performances of students. The existing studies deploying adversarial networks for class imbalance are also discussed in detail. The literature highlights the significance of the proposed approach, in a temporal class imbalance setting.

- In Chapter 3, the data methodology for conducting the bibliometric study is discussed, and the related terms relevant to this bibliometric research are defined along with a discussion on the results. The results of this study are also published in the following journal:
• In chapter 4, academic performances of students are predicted using the interactions from a VLE dataset using demographics along with engagement information. Some influential characteristics, impacting these academic performances, are identified and discussed. The chapter further defines the methodology of this contribution, the deployed dataset, its transformation and feature engineering. The architectures of the deployed machine learning algorithms are discussed along with the results of the study, that highlight the effectiveness of the proposed methods in the prediction of academic performances at the end of the courses. This study is published in the following journal:


• In Chapter 5 the significance of a sequential approach is explored, encompassing the transformation of student engagements in a quarterly and week-wise setting, through smart data processing. Deep-learning based models are deployed for predicting the academic performances of students during an on-going course, and the results are compared with conventional machine learning algorithms. The results reveal the effectiveness of the deployed deep-learning based approaches to predict the award-gap in the performances during an on-going course. The results of this study are under-review in the following journal:


• In Chapter 6 the issue of award-gap class imbalance in a temporal setting is discussed. The significance of the adversarial networks, to handle the award-gap class imbalance in a temporal setting are discussed and an adversarial based approach is proposed to tackle this issue in the learning analytics community. Subsequently, the results of this contribution are compared with other gan-based and non-gan based approaches. This work is under-review in the following journal:

• Finally, in Chapter 7, the dissertation is concluded by discussing the significance of the obtained results and deployed approaches, its implications on the learning analytics community, especially in building early warning systems, are discussed. Further, concluding remarks and possible future directions are presented to highlight this crucial aspect of the learning analytics community.
Chapter 2

Literature Review

"An organization’s ability to learn, and translate that learning into action rapidly, is the ultimate competitive advantage.”
- Jack Welch

With the growth of the internet, online education has become one of the rising key phenomena providing sufficient repositories of student and learner information. This allows analysis for predicting patterns in an educational setting, defining association between the stakeholders involved and optimizing the learning environment [17, 18]. Online systems include, but not limited to, Learning Management Systems (LMS), Course Management Systems (CMS), Massive Open Online Courses (MOOCS), Virtual Learning Environments (VLE), Intelligent Tutoring systems (ITS) and other web-based educational systems; contribute to generating digital footprints that can be examined to assess the prospective behavior of learners, analysing activities of successful and at-risk students, providing corrective strategies based on learner’s performances, consequently assisting instructors in improving the pedagogical methods [29, 31].

2.1 Self-paced Online Learning

Self-paced online learning is a commonly used approach in many open universities around the world (e.g., the Open University of UK and Athabasca University) [55]. The approach allows students to get an extended period to study (e.g., 6 or 9 months) to afford an extra level of flexibility. Although the courses offer a recommended schedule of a range of tasks defined in the coursework, students in such courses typically studying part-time and, thus, can have a greater degree of flexibility to organize their learning. While being very flexible, self-paced learning can increase potential risks for the success of students, given that it may impose greater expectations for students in terms of time management and self-regulated learning [56]. There has been a considerable amount of research to understand student retention in distance and online learning and to evaluate the efficacy of different interventions
Existing research in learning analytics have even produced some important insights into the ways how students manage their time [60] and how the efficacy of learning design can be evaluated in self-paced learning afforded through formal or credit courses offered fully online [61]. However, there has been much less research on predictive modeling in this context. Although some related research on predictive modeling has been done in the context of MOOCs offered on-demand is relevant [62], the non-formal nature of education in MOOCs sets a different context from online learning in formal education.

2.1.1 Predictive Modeling in Learning Analytics

Student retention is a widely recognized challenge faced by higher education institutions [63]. Although flexibility in learning is highly appreciated to address needs for the contemporary student population (e.g., work at least part-time and have families), it equally increases the risks of a student failing to complete their education programs. For example, it is well-established that fully online and self-pace education programs require strong skills for self-regulated learning to assure success [56]. Moreover, students in fully online learning can often experience isolation, weak social integration, and simply lose the pace of their studies [64].

Higher education institutions have recognized a large potential of data collected by digital technologies used in learning and teaching practices to provide better support for their students and increase student success. Specifically, many higher education institutions started investing in institutional programs for the adoption and implementation of learning analytics [65]. Early prediction of students at risk of failing a course or dropping out of a degree program has been among the most popular uses of learning analytics in higher education [66]. These applications of learning analytics are seen by senior management in higher education institutions as highly promising to increase student retention [23]. The keen interest in early prediction by senior management in higher education institutions is driven by their aim to address government priorities and, thus, maximize access to government funding for higher education. More importantly, this provides a great service to society by producing a higher trained workforce. Implementation of learning analytics for early prediction is connected with changes in educational support structures in institutions such as revision of educational designs, cost-effective allocation of resources, and formation of student counseling teams [9, 49, 67, 68].
Existing research in learning analytics has dedicated much attention to the prediction of students at-risk in online learning, massive open online courses, and blended learning in situations when students are following a predefined schedule of activities [69]. However, much less attention has been dedicated to self-paced learning [55], which has traditionally been the main delivery mode of education in main open universities (e.g., the UK Open University and Athabasca University). This mode is also recognized and available in massive open online courses and often referred to as on-demand [70]. As instructional conditions matter, it is essential to investigate how predictive models can be developed to support this delivery model. This model is also attracting the attention of conventional higher education institutions in their aim to cater to the changing needs of students who often have at least part-time jobs and dependents.

The existing literature in learning analytics follows a two-fold analysis for predicting students’ performance where the task is defined as either prediction of final grades/scores in a course [71] or prediction of passing or failing in a course [72]. Typical input predictors used in such models are based on a combination of socio-economic demographic data as commonly available in student information systems with the log-data that are collected as a by-product of learners’ interaction with learning technology. The next subsections highlight the conventional modeling approaches, in chronological order, to predict the performances of students at the end of a course. Further, the next subsection presents the modelling approaches in the early prediction process during an on-going course, along-with a discussion of the significant predictors of student success and presenting the more recently deep learning and other methods deployed to early identify the students at-risk of failure. The last subsection, discusses the class-imbalance issue in the learning analytics community and presents the existing techniques to eradicate this issue.

### 2.2 At-risk Student Prediction at the End of a Course

The research area related to students’ performance prediction is multidimensional and can be explored and analysed via multiple perspectives, including early prediction of dropouts and withdrawal students. Various data mining techniques are deployed on educational datasets to predict students’ performance, assessing slow learners and dropouts [50, 51, 73, 74]. The techniques employed on these learning analytics datasets aid in data-driven decision making [49]. Early prediction is a new
phenomenon in this domain, encompassing methods to timely assess the students in order to retain them, by suggesting suitable corrective strategies and policies, subsequently managing and reducing attrition rates. In the literature, there has been substantial debate on the subject of student retention in MOOCs. Numerous studies have emphasised the identification of the factors contributing to students’ dropout rates [75, 76, 77, 78]. Various factors, such as the attribute of time, motivational factors, non-existent interactivity between students-instructors, lack of knowledge of the course pre-requisites and effectiveness of the course content, are associated with the early withdrawal of students in MOOCs [79, 80]. Jaggers & Xu [81] conducted a study identifying the major factors influencing student performance in an online environment and concluded student-instructor interaction to be the foremost attribute positively impacting one’s performance.

Learning analytics assists in providing automated, real time opinions and recommendations for constructive pedagogical strategies through multiple learning analytics platforms (such as learning analytics dashboard, LMS visualization systems). These feedback systems aid in constructing a more robust platform to measure, examine and predict student involvement, contributing self-regulated learning and driving motivational factors to regulate successful goal achievement [82, 83, 84, 8]. Corrin & de Barba [82] initiated a study to interpret the influence that feedback provided via learning analytics dashboards had on students’ performance, motivational drive and strategy development for course work. These applications induce a positive drive in students learning, consequently impacting performance [20].

Different attributes are associated with the prediction of students’ performance. Shahiri & Husain [85] conducted a systematic literature review to observe the characteristics substantially contributing to the prediction of class-room performances. Cumulative grade point (CGPA) and assessments (such as assignments and quiz marks) were considered to be the two primary attributes to assess students’ performance [86, 87, 88, 89]. Another perspective encompasses factors of students’ legacy data (such as past performances in previous assessments/entry test etc.) and demographics in being significant contributors to assessing performances [11]. Furthermore, another study employed family characteristics, such as family expenditure, income and students personal information, to assess the impact on their performance [37]. They concluded that excessive rental expenditures and health expenses impacted the overall environment ultimately influencing student’s performance. Emotional stability, a student’s inherent attribute, was considered to be an important predictor of his/her performance. Married students were found to be more emotionally stable as compared to unmarried students. Similarly,
self-employment was found to have a substantial impact on their performance. Employed students were found to be more responsible and industrious as compared to un-employed.

Marbouti & Diefes-Dux [90] utilised machine learning techniques to predict at-risk students. They predicted students at risk of failure in the 2nd, 4th and 9th week of their first year of engineering. Their dataset comprised of attendances, quizzes and assignments, with the inclusion at 9th week, an additional attribute of mid-term exams. Deploying a logistic regression model, they achieved an accuracy of 98% by week 9. Additionally, in a study, at-risk students were predicted by deploying various data mining techniques, including Support Vector Machines (SVM), Naive Bayes Classifier, Decision Tree, K-Nearest Neighbor and Multi-Layer Perceptron, to identify the best prediction modeling method. Logistic Regression was employed as the baseline model [91, 92].

Deploying deep learning techniques on learning analytics to predict successful and at-risk students, is rather a new area of research. Deep learning employs techniques that encompass constructing a model comprising of multiple layers to learn representations from raw data. This representation learning consists of multiple layers, where each layer transforms the representation to a more abstract form, for the next layer [42]. Corrigan & Smeaton [93] predicted student success via a VLE by including the number of times a student interacts with the environment. They deployed Recurrent Neural Network’s (RNN) variation Long Short Term Memory (LSTM), to predict the success of students based on their interaction with the Moodle based learning environment. The results were evaluated using Random Forests, and LSTM outperformed it by 13.3% of the variance of the model, as opposed to 8.1%.

Okubo et al. [45] predicted student success by utilizing various features extracted from Kyushu University’s learning system named M2B. They collected 108 student learning logs for an ‘Information Science’ course. The learning logs comprised of a feature set of weekly quiz, attendance, notes on the lectures delivered, sides view and book markers. They deployed the RNN model with hidden layers of LSTM to predict students grades based on their interactions and defined features. The results were compared with multiple regression analysis and the proposed model surpassed the regression models through early prediction of the grades. Fei & Yeung [76] employed different machine learning techniques to predict student dropout rates in two courses from Massive Open Online Courses (MOOCs). Since the data is time dependent, they viewed it as a sequence classification problem. MOOCs are subjected to high attrition rates, hence apart from baseline models such as SVM and Logistic regres-
sion, they implemented Input Output Hidden Markov Model (IOHMM), RNN and LSTM to identify the best technique among these. The feature set incorporated lecture view, lectures downloaded, quiz attempts, numbers of access, forum activities, forum views and number of times students commented on forums. After deploying the above mentioned techniques, they found that a combination of RNN and LSTM had the highest accuracy.

2.3 Early Performance Prediction of Students At-risk

The literature in learning analytics and educational data mining offers a plethora of approaches to predictive modeling for early detection of students at risk [52, 53]. Predictive modeling in learning analytics is heavily rooted in the use of machine learning with the primary focus on classic machine learning techniques (e.g., decision tree, random forest, and support vector machines). Only recently has the research community started paying attention to the use of deep (neural network-based) learning techniques. A possible reason for the relatively slow adoption of deep neural networks in learning analytics could be attributed to the need to use ‘white-box’ algorithms to promote actionable insights through the use of interpretable models.

Early prediction of students’ academic performance is an interesting area of research with existing research [66, 54, 94, 95] emphasizing on the value of the Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) algorithms that analyze the activities of students on a fine-grained level such as analyzing their daily or weekly log data. Such fine-grained practices provide a more in-depth analysis of the behavior of students.

Brooks et al. [96] analyzed the dataset of Coursera courses to build a predictive model using a decision tree to assess the academic performance of students in an on-going course. Their dataset comprised of the vectors of students’ resources accessed, such as discussion forums, quizzes, and lectures viewed. They analyzed each student’s performance on a daily, weekly, and monthly basis where each student vector was composed of the respective aggregated log data. With data of the first five weeks, they correctly classified 68.59% of students and achieved Cohen’s kappa of 0.374. Furthermore, they extended their study to analyze the behavior of distinction students and predicting the performance of active students [97]. They split the data into two cohorts of high and low achievers, with 85% or above marks demarcated for distinction students. In order to eliminate class imbalance, they used
random sub-sampling to balance the two cohorts. Decision trees with 10-fold cross-validation were deployed to build descriptive models for the two cohorts achieving a kappa score of 0.8199. In order to predict distinction students, the same model was applied to other courses achieving a kappa of 0.307 and 0.347, respectively.

Corrigan & Smeaton [93] deployed a variation of Recurrent Neural Networks to predict the success of students through a regression analysis by incorporating data about students’ interactions with the learning platform. Their deployed deep learning approach, with an r-squared value of 0.13547, outperformed the traditional random forest regressors, which had an r-squared value of 0.080. Similarly, another study predicted the success of students through the inclusion of their attendance rate and tapped their behavior through log data information [45]. They deployed a RNN model with hidden layers of LSTM to early predict students’ grades based on their interaction with the learning platform. The results were compared with multiple regression analysis, and the proposed model surpassed the regression models by 90% of prediction accuracy at almost half of the course completion. Fei & Yeung [76] employed a feature set consisting of the lectures watched and downloaded, assessments attempts and scores, participation in the forum activities, and number of times commented to predict students’ academic performance by assessing students at-risk. They implemented an array of techniques, such as SVM, Logistic regression, Input-Output Hidden Markov Model, RNN, and LSTM, and found LSTMs to outperform the other techniques.

Wang et al. [46] also implemented LSTMs to analyze student behavior in a MOOCs course and their performance in assessment related to course exercises, where based on their previous submission patterns, each student’s academic performance was assessed. The results were observed to have a 5% improvement compared to logistic regression, which was used as a baseline. Furthermore, Kim et al. [98] developed a model named GritNet that predicted performance by transforming each student’s activities into a sequence of uniform length, such that each vector represented unique actions. The sequences for each student were merged to a single vector, encompassing their previous engagement pattern. The model embedded the event vectors to reduce them into a lower dimension, and these were then fed to the Bi-directional LSTM. A GMP layer was introduced before the output layer that enabled the model to focus on the most relevant part of the event. The model was observed to have attained superior performance in the upcoming weeks, with an AUC of 74% in the first week to 80% in the second week, reaching above 90% in the final weeks, as compared to a logistic regression model that was used as the baseline and achieved an AUC of 68.6% in the first week and 76% in the
second week.

At first, we discuss recent relevant studies pertaining to the self-paced online learning. Furthermore, we systematically discuss a detailed review of the predictive modeling approaches in learning analytics vertical in chronological order.

2.3.1 Predictive Modeling Approaches in Self-paced Online Learning

Existing research on predictive modeling in learning analytics has used a range of approaches from standard statistical approaches to conventional machine learning and, more recently, deep learning and neural networks-based approaches. Among the many practices, regression models and Markov models have been used in a number of studies to predict potential course-specific performances for the next term/semester, where the course is characterized in a latent space [99, 100]. Morsy & Karypis [99] deployed a knowledge-based regression model to predict academic performance for the next term by accumulating the performance in various courses undertaken by students, using the data of 12 majors from the College of Science and Engineering at University of Minnesota. The results from this study were compared with the course-specific regression model [101] that undertakes the course-level performance to predict course level grades for the next term, and their deployed method provided better results with less error rate. They reported an RMSE of 0.661 vs. the 0.632 RMSE achieved with the previously best-performing schemes.

Lacave & Lemus [102] employed Bayesian models to predict at-risk students by leveraging their academic (grades, enrollment year, subject undertook, scores in subjects) and social information (demographics) from the institute database by including students from the degree of computer engineering. They found K2 Bayesian model performed better than others with a log-likelihood score of -12.99 on a 10-fold cross-validation. Burgos et al. [71] made use of logistic regression to analyze student performance and predict potential dropouts by leveraging student’s legacy data pertaining to course activities and its submission in an online learning environment. The deployment of this predictive model reportedly reduced the dropout rate by 14% with an accuracy of 97.13%.

Chen et al. [103] deployed a survival-based statistical approach to identify undergraduate dropouts by including their demographical information, high school GPA, entry test SAT scores, and semester
information constituting of credit hours, semester grades, and major undertaken. They conducted a semester-wise analysis to identify students at-risk of a drop-out. Their deployed statistical approach, achieving an F1 score of 73.3%, performed better in the initial weeks compared to the deployed machine learning algorithms: logistic regression, random forest, decision trees, and AdaBoost.

The traditional approach to predicting student performance relies on deploying linear models such as Logistic regression [76], linear Support Vector Machine (SVMs) [62], and statistical methods like survival analysis [103]. Moreover, an array of machine learning algorithms, such as Decision Trees (DT) and Naïve Bayes, are also constituted as significantly influential predictors [104]. Marbouti et al. [92] used an array of machine learning algorithms to identify students at-risk of a failure in an online course. Their dataset constituted of students’ attendance, assessments and homework grades, in-class quizzes, project participation, and exam results. From the deployed six algorithms, multiple layered perceptron was found to give better results with an overall accuracy of 93%; however, the model only predicted 48.3% of the students who failed the courses.

Another emerging approach relies on exploring deep machine learning techniques with reference to prediction of performances at the end of a course and during an on-going course. With the adoption of deep learning algorithms for analysis of educational data, a recent trend in the community has been towards the early prediction of students’ performance [45, 105]. Along with Deep ANN, few prominent works include Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN) with its variants. This area of research, with the inclusion of deep learning techniques to predict the performance of students, is still in a preliminary stage with limited evidence reported. The core principle of deep learning techniques relies on the existence of multiple non-linear layers, with each layer supporting the concept of hierarchical representation learning supporting non-human engineered representations and transforming these learned representations to the above layers [42].

An array of several data analytic approaches, constituting of various machine learning methods, are conducted for the prediction of potential learner performances, with Deep ANNs being a noticeable practice due to their hierarchical representation and learning abilities [41]. The deep learning techniques consist of several non-linear computational layers, assisting an algorithm to learn from existing fundamental patterns and superseding the conventional feature engineering practices [43, 44]. Some evidence, in terms of the number of publications, can be observed in the literature that analyzed the effectiveness of deep learning techniques in learning analytics, specifically for understanding learners’
behavior and measuring the improvement in student performances through interventions. A systematic literature review conducted by Coelho & Silveira [41] analyzed this paradigm in relation to deep learning techniques. It demarcated the superior performance of these techniques over the traditional statistical approaches for the prediction of learners’ performance [23, 67, 93]. While the literature supports the use of deep learning algorithms for predictive modeling in education, there has been limited research that systematically compares the performance of such algorithms with a range of conventional algorithms and on a publicly shared dataset about self-paced education to enable replicability of findings.

### 2.3.2 Important Predictors of Student Success in Self-paced Online Learning

Bonafini et al. [106], through quantitative analysis, investigated the participation in the discussion forums while watching video lectures and found a high involvement in the discussion forums to be positively correlated with student success in that course. Similarly, Sunar et al. [107] laid emphasis on participation in discussion forums and their impact on academic success. They applied social network techniques to identify communities of students and their followers and suggested active participation in forums to demonstrate a lower retention rate for a particular course. Another study [102], through an application of Bayesian algorithms, found students’ basic information such as age and awarded the scholarship as important predictors of academic performance. Similarly, the study also found that if course information, such as grades and scores achieved, is provided, then the students’ entry test scores hold less significance. Mutahi et al. [108], through statistical techniques, showed that high interaction with the quizzes demonstrate positive association on the academic performance of students. Moreover, another study analyzed the access to course contents and concluded increased access to the content to be positively associated with the improved academic performance [109].

Legacy data of students such as their previous grades, assessment performances, quiz performance, participation in course activities, and engagement patterns are also used to deliver an understanding of their learning behavior [11, 92, 110]. Hlosta identified the significance of the first assessment and its impact on the academic performance of students, using various machine learning classifiers. In line with the previous researches, they found the first assessment to be a good predictor of student success for a particular course [110]. O’Connell et al. [111] conducted a statistical study to identify the
indicators of student success, through a multiple linear regression and identified students’ past grades and earned credit hours to have an influential impact on their current course performance. Moreover, they also found assessments to be positively correlated with improved academic performance. The existing literature follows statistical and conventional approaches to identify important predictors of success, such as finding highly correlated attributes with the performance of students [112]. However, the literature lacks substantial studies in this aspect where model agnostic techniques are deployed to identify significant predictors of success. This study contributes in this direction to find influential predictors impacting the performance of students.

2.4 Class Balancing of Students Logs in Self-paced Online Learning

This section discusses the existing studies on the class balancing problem in the learning analytics discipline. The section is further organized in two parts. At first, we review the existing studies and the up-sampling techniques deployed in these studies pertaining to the educational data science field. Secondly, we discuss some generic studies on GANs for tabular data generation.

2.4.1 Review on Up-Sampling Techniques in Educational Data Science

Class imbalance is one of the crucial challenges incurred by the learning analytics community with the disparity in the performances of students. In a conventional setting, usually majority of the students tend to pass a course and a minority of students tend to fail. Similarly, students dropping out of a degree will be much less than those who tend to pursue it. This disparity in the performances of the students can be mapped to the classic issue of class imbalance which is catered through various data-level and algorithmic-level methods [113]. Synthetic Minority Over-sampling Technique (SMOTE) is one of the most widely applicable methods in the learning analytics community with several studies deploying it to eliminate the imbalance by over-sampling the minority classes [114, 115, 116]. Ghorbani Ghousi [117] deployed various versions of SMOTE in comparison with Random Over-sampling (ROS) on two educational datasets from Iran and Portugal universities. Both the datasets consisted of various features to predict the performance of students in terms of their GPA (binary-classification) and categories of the GPA such as poor, medium, good, excellent (multi-class classification). An ar-
ray of over-sampling methods; SMOTE, SMOTE-ENN, SMOTE-SVM, SMOTE-Tomak, were used in comparison to ROS and the results were validated using various machine learning algorithms. Comparison between the results of balanced and imbalanced data revealed SVM-SMOTE to outperform other sampling methods with 73% and 81.27% reported accuracies.

Farrow et al. [118] predicted the cognitive presence of students by modeling their interactions in the discussion forums. By undertaking the data of discussion forums they eliminated the class imbalance by using a variant of SMOTE, named as SMOTE exact, where the minority classes were oversampled in a controlled environment such that all the classes had equal number of instances. The study emphasized on the significance of the data contamination that occurs due to the over-sampling methods and avoiding it by re-balancing the classes inside the cross validation, and thus eliminating biasedness and producing better results. Krawczyk [119] conducted a survey and discussed the challenges presented by the imbalanced data in the machine learning domain. Specifically, it mentioned the learning process from the imbalanced data distributions where the imbalance ratio between classes may change as time progresses. Furthermore, following this concept of varying ratios of imbalanced classes over the time, Wang et al. [120] examined the speed of the changing ratios through various bagging sampling methods of over-sampling and under-sampling and found the former one to perform better than the later one. Students assessment submission, where a cut-off defines the deadline, can be mapped to this concept of varying imbalance ratios, since more submissions are expected near the submission deadline [121]. This student behavior, termed as procrastination, has also been examined by other studies [110, 122]. Since such problems will yield towards a low number of students submitting in the beginning of the assessment compared to those submitting near the deadline, thus the data will present an imbalance ratio of submissions. Hlosta et al. [121] applied some conventional up-sampling techniques; ROS, SMOTE, SMOTE-ENN, SMOTE-Tomek, Tomek Links and some domain-specific sampling techniques; EQ,CLS equal number of instances from both classes, EST_RT estimating the submission ratio of students, and eliminating the overlap between classes, to mitigate the complexities of the inherent problem of class imbalance in the assessment submissions. Their results demonstrated the superiority of the domain-specific techniques over the conventional sampling methods.

Overall, in the existing literature several studies practice the conventional sampling techniques to handle the inherent problem of class imbalance in the educational datasets.
2.4.2 GANs for Temporal Data Generation

With the recent advancements in the adversarial networks [123, 124], an array of disciplines have adopted the use of these networks for synthetic data generation, from generating images and videos to polyphonic music generation [125, 126]. A GAN consists of two components, a generator network, that imitates the actual data and attempts to generate synthetic but realistic instances, and a discriminator network, responsible for distinguishing between the synthetically generated instances and the actual ones. In the existing literature limited evidences exist for generating the temporal data with few studies observed in the healthcare community, notable among them being MedGan and Recurrent Conditional Gan (RCGan) [127]. MedGan [128] is based on a combination of auto-encoders and adversarial network and learns the distribution of high dimensional discrete variables. In contrast to other generative methods in the healthcare systems, MedGan is not domain specific and thus does not cater a diverse range of applications. The generator and discriminator for RCGan are based on LSTM to capture the distribution of the data and are subjected to a condition at each time step [127]. Another noticeable work in temporal GANs is the generation of continuous sequences of polyphonic music using LSTM based GANs [129].

Some prominent works have also been conducted in modeling the financial time-series data using adversarial networks [130, 131, 132]. However, these studies are domain-driven and specifically target the domain-level specificities of stock exchange. Another study, applies a conventional GAN on the fraud detection data to alleviate the class imbalance, however this study does not cater the temporal data [133]. They augmented the training data with the generated instances to improve the performance of the classifiers and compared the results with SMOTE to demonstrate the effectiveness of GANs.

In the learning analytics community, one study combines CGAN, InfoGAN [134] and auxiliary classifier GAN (ACGAN) [135] to implement an improved conditional GAN (ICGAN) [136]. They deploy their proposed ICGAN on an educational dataset [137] consisting of several attributes related to family and school tutoring, such as parents’ education, their quality of life, relationships, legacy information of students, time given to studies, extra classes taken, internet access available at home etc. The study did not mention inclusion of any clickstream data. A deep SVM model is implemented as a classifier to predict the performance of students and results are compared with base line SVM to show the effectiveness of their proposed GAN-based model. Although the study uses GANs to
eliminate the class imbalance, it does not cater temporal functionality of a time-series data.

In the learning analytics community, sufficient evidences can be observed on the prediction of temporal data [66, 93, 94, 95, 96, 97], however they lack an implementation on temporal data generation. Temporal data consists of multiple sequences which are time dependent and each new sequence is associated with its previous sequence. For instance, if the students’ engagement patterns are presented longitudinally (such as day-wise or week-wise etc.) then each student’s daily interaction will be associated with its previous day’s interaction. In terms of generating such temporal data, this association of sequences should be adhered to. The existing studies on GANs lack this implementation of temporal data for capturing the behavior of students.

2.5 Summary

This chapter discussed the significance of the self-paced online learning in the discipline of learning analytics, existing studies pertaining to the prediction of at-risk learners and the prevalent modeling approaches in this context. The prediction of academic performances is categorized into two aspects; 1) predicting at the end of the courses along with the legacy data and 2) early prediction of the academic performances where the students’ interactions are modeled on a temporal basis. Further, the issue of class imbalance in the learning analytics community is discussed and the relevant existing studies eliminating this challenging problem are presented. However, the up-sampling of such educational data in a temporal setting presents some new challenges that have not been addressed in the learning analytics community. In this dissertation, we intend to fill this research gap such that the behavior of the up-sampled students is retained throughout the course duration.
Chapter 3

Bibliometric Analysis of Learning Analytics

“Numbers have an important story to tell. They rely on you to give them a voice.”

– Stephen Few

This chapter describes the bibliometric analysis on the discipline under-study and the related indices associated with it. This chapter intends to examine the world’s research landscape in learning analytics techniques by deploying qualitative and quantitative bibliometric analyses. Bibliometric studies tend to statistically examine the quantitative aspects of scientific publications within a field [138, 139]. The growth in scientific literature can be explored through its quantitative analysis, with publication outputs signifying the research productivity and citation counts indicating the scientific impact [140, 141]. The number of citations for a publication signifies its higher influence [142]. Citation analysis reflects the international scientific influence and scholarly impact of a publication [138, 143]. The number of international co-publications shared between two regions indicates the extent of collaboration between those regions, implying knowledge flows [144, 145]. Bibliometric measures are deployed in various academic research areas to explore the quality and impact of the publications produced in that area, enabling researchers to analyse scientific products [146]. Hassan and Haddawy [147] deployed bibliometric techniques to measure quantitatively the flow of knowledge among countries and intellectual impact. In this paper, we deploy state-of-the-art bibliometric indices such as co-authorship network, citation networks and terms co-occurrence networks, the details of which are provided in subsequent sections. Specifically, the following are the objectives of our first contribution:

(i) To procure the relevant set of scientific publications that belongs to the learning analytics research space

(ii) To study quantitatively the multidisciplinary field of learning analytics over time, specifically in terms of publication and citation counts

(iii) To identify the institutions and countries dominant in the field

(iv) To study collaboration network patterns with respect to institutions and authors
(v) To identify the citation exchange among the source titles (conferences/journals)

(vi) To analyse the temporal thematic evolution of the field

(vii) To explore the tools and techniques being deployed for data-driven decision-making in the learning analytics research space.

The intended contribution as derived from the above objectives is multidimensional. We present a comprehensive view of the new era of learning analytics which, in fact, is multidisciplinary, flexible, dynamic and powered by a new, sophisticated series of computational tools, enabled by cognitive computing and Big Data techniques. Finally, the significance of the human factor is highlighted. The evolution of learning analytics research in higher education institutions both requires and promotes academic and scientific collaboration towards the accomplishment of various soft factors.

Bibliographic analysis assists in identifying influential communities, institutions and countries participating actively in an area, and helps in exploring the cutting-edge trends for that area, revealing interconnection patterns among different communities [148, 149]. In such analysis, various key terms are of significant importance, namely co-authorship network, citation network and the term co-occurrence network. A brief description of each is given below. These terms were applied to the procured dataset to explore the learning analytics research space.

3.1 Bibliography Dataset for Learning Analytics

Bibliographic database resources organize scientific publications by mapping them with predefined categories in the database. For instance, Scopus employs All Science Journal Classification (ASJC) to classify sources (journals/conference proceedings) in an organized hierarchy of disciplines and sub-disciplines. However, for emerging and multidisciplinary areas, such a structured ordering scheme does not cater for the entire publication dataset [150, 151]. Therefore, in order to procure the entire dataset in interdisciplinary and multidisciplinary areas, either simple terms, such as ‘learning analytics’, are searched for in the titles, abstracts and keywords or a collection of key-word terms related to that area is constructed [152]. Our present work employs the latter approach by listing a combination of different key words relating to learning analytics, using the seed key words method. These key
words are selected with the help of domain experts. Finally, a Scopus-compatible query is constructed for all the terms relating to learning analytics to extract relevant scientific publications.

**Table 3.1: Scopus-Compatible Query String for Bibliometric Study**

| Scopus-compatible query string | TITLE-ABS-KEY ("Learning Analytics") OR TITLE-ABS-KEY ("Academic Analytics") OR TITLE-ABS-KEY ("Educational Analytics") OR TITLE-ABS-KEY ("Educational Data Science") OR TITLE-ABS-KEY ("Educational Data Mining") OR (TITLE-ABS-KEY ("Learning Process") AND TITLE-ABS-KEY ("Data Mining") AND TITLE-ABS-KEY ("Education") ) |

As shown in Table 3.1, a Scopus-compatible query is matched against key words in the title, author-defined key words and abstracts. Since learning analytics resonates so strongly with educational data mining, this term was included in the query as well. Learning analytics constitutes various sub domains, including academic analytics and educational analytics, therefore both terms are merged in the query. In order to procure all the relevant publications relating to learning analytics, a term ‘learning process’ is also incorporated into the query. However, it is quite a general term and may fetch publications that are unrelated to the field. Therefore, in the final query, this term is ‘ANDed’ with the terms ‘Education’ and ‘Data Mining’, limiting publications in that field and procuring results relevant to learning analytics. The term ‘publication’ is mentioned to obtain scientific literature published in recognized technical journals or conference proceedings, including reviews and scientific articles.

### 3.1.1 Data Pre-processing for Bibliometric Analysis

The data is acquired from Scopus in .csv (Comma Separated Values) format. Scopus files constitute bibliographic information, such as a publication’s name and year, citation count, affiliations and source title information, which assists in performing bibliometric analysis. A dataset of 2,925 publications was procured. The data was pre-processed and cleaned by excluding irrelevant attributes, such as editor, sponsor and so on, and incomplete records with no author names. After pre-processing, a dataset of 2,811 publications was finalized for bibliometric analysis. Fig. 3.1 presents the document types of the publications in the dataset.

The attributes used for bibliometric analysis of learning analytics consist of authors, publication title, year of publication, source title, cited by, affiliations, authors with affiliations, author key word, in-
Figure 3.1: Document Types

dex key word, references, electronic identification of document (EID) and document language. The country names are extracted from the affiliation attribute in order to procure and highlight areas participating in learning analytics research worldwide.

3.2 Bibliometric Indices

This section discusses the bibliometric indices that have been deployed on the procured dataset.

**Co-authorship Network:** In scientific literature, ‘co-authorship’ is a measure of collaboration between different authors, aiding scientific development and progression growth [153]. Such collaborations assist in developing social networks, improving knowledge and promoting growth [154]. It is most prevalent in interdisciplinary domains, where authors from multiple areas collaborate to produce an intellectual study [155]. Such networks assist in interpreting the behavioral characteristics of scientists in various domains, highlighting the phenomenon of knowledge flows. According to a study by [156], scientific articles, co-authored by various scholars, represent a link between different institutes, depicting the structure of knowledge flows. In this study, we constructed collaboration
networks with respect to authors. The details are provided in the next section.

**Citation Network:** In bibliometric citation, networks are constructed using co-citation or bibliographic coupling networks using scientific literature. These networks help to quantify interdependencies and scholarly influence among the entities at various levels of detail, including between scientists, journals, subject categories, institutions and countries. More specifically, co-citation occurs when two articles in the dataset are independently cited by one or more articles, indicating their relativity of working areas [157]. Similarly, bibliographic coupling is a phenomenon where two or more publications reference a third common publication in their bibliographies, implying that the two publications are working on relatively similar areas and are interrelated with each other. Fig.3.2 demonstrates the difference between these two key methods. For instance, Document A and B are two articles in our dataset. If both cite a common third article, this relation will be termed Bibliography Coupling, whereas if an independent third article simultaneously cites Document A and B, then it is termed co-citation. In this study, we have constructed citation networks w.r.t. source titles. The details are provided in the next section.

**Terms Co-occurrence Network:** To study the thematic structure of Learning Analytics, we created term maps of the text corpus for the field. We used VOSviewer version 1.6.6 to analyse the temporal evolution of learning analytics. Specifically, we sought to explore the tools and techniques being deployed in data-driven decision-making in the learning analytics research landscape. In order to give input to VOSviewer utility to create concept maps, terms from title, abstract and author-defined keywords of the articles were used as candidate items to represent each article. VOSviewer tends to extract the nouns, as candidate terms, from the title, abstracts and author-defined keywords. Using the selected candidate terms, VOSviewer returned concept maps on the recommended normalization method 1, with 0 convergence, 1 resolution and 100 random starts. Each node in the map represents a semantic concept, which ultimately are the terms extracted from articles’ titles, abstract or author-defined keywords, size of a node shows the frequency of each concept and the distance between two nodes illustrates the associated strength between the concepts [158]. This association strength indicates the similarity between the two concepts and is computed as shown in equation 3.1.
Figure 3.2: Bibliographic Coupling and Co-citation Method

Bibliographic coupling

Doc A citing
Doc C
Doc D

Co-citation network

Doc B citing
Doc D
Doc C
Doc A Cited
Doc B Cited

Cites
Cites
Cites
where $s_{ij}$ indicates the similarity, $c_{ij}$ represents the number of co-occurrences between two concepts $i$ and $j$, $w_i$ is the total number of occurrences of the concept $i$ and $w_j$ is the total number of occurrences of the concept $j$. These concept maps aid in illustrating and understanding the thematic structure of the concepts/terms deployed in the research space. VOSviewer analyses these concepts among different documents.

3.3 Results & Analysis

Firstly, the results of the conducted bibliometric analysis are presented to highlight the research landscape of the learning analytics discipline. We present an analysis of the top institutes, sources (journal or conferences) and countries that are the most research active and prevalent in learning analytics, in terms of the number of publications, citations and collaboration, during 2000 to 2017. This time window was chosen because publications on learning analytics do not exist before 2000. Therefore, our Scopus-compatible query only extracts papers from 2000 onwards. At country level, the research activities of those countries that have the highest number of publications among them were examined. At institute level, the research strengths of various institutes around the world were explored to determine the topmost in terms of research activity. Google API was used to find the locations of the universities producing learning analytics research in order to map them visually on Tableau (free student version). A comparison was performed on the sources to evaluate their research strengths and participation in terms of number of publications and magnitude of collaboration. Finally, for qualitative analysis, term co-occurrence maps were created using VOSviewer.
3.3.1 Publication Output of Countries and Institutions in Learning Analytics

This section presents the publication output of countries in learning analytics. Fig. 3.3a demonstrates the top countries in terms of publication count. The United States (US) appears at the top by producing 677 publications, followed by Spain, the second country worldwide active in learning analytics research, producing 336 publications. The US is clearly leading the research in this area, and a gap can be seen between it and Spain. Similarly, a clear demarcation in publication counts can be visualized between Spain and other countries at lower ranks. This depicts the strength of the top countries in this area w.r.t. the number of publications and demonstrates the research landscape of learning analytics. Interestingly, most of the top countries in terms of publication output also emerge as the most cited (refer to Fig.3.3b), implying the quality of their work in this field. The US emerges on top with 10,136 citations, followed by Spain with 9,044 citations. The publication and citation counts of the top-10 countries are shown in Table 3.2. While India, China and Japan appear in the top-10 countries w.r.t. publication counts, they rank low in terms of citation counts and are supplanted by countries such as Belgium, Greece and Taiwan, with citation counts of 1,584, 823 and 552, respectively.

Table 3.2: Top 10 Countries in Learning Analytics w.r.t. Publication Counts from 2000 to 2017

<table>
<thead>
<tr>
<th>Sr.#</th>
<th>Country</th>
<th>Publication Count</th>
<th>Citation Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>United States (US)</td>
<td>677</td>
<td>10,136</td>
</tr>
<tr>
<td>2.</td>
<td>Spain</td>
<td>336</td>
<td>9044</td>
</tr>
<tr>
<td>3.</td>
<td>United Kingdom (UK)</td>
<td>281</td>
<td>2,478</td>
</tr>
<tr>
<td>4.</td>
<td>Australia</td>
<td>263</td>
<td>3,497</td>
</tr>
<tr>
<td>5.</td>
<td>Germany</td>
<td>198</td>
<td>2,667</td>
</tr>
<tr>
<td>6.</td>
<td>Canada</td>
<td>160</td>
<td>2,478</td>
</tr>
<tr>
<td>7.</td>
<td>India</td>
<td>130</td>
<td>390</td>
</tr>
<tr>
<td>8.</td>
<td>Netherlands</td>
<td>125</td>
<td>2,008</td>
</tr>
<tr>
<td>9.</td>
<td>Japan</td>
<td>101</td>
<td>399</td>
</tr>
<tr>
<td>10.</td>
<td>China</td>
<td>94</td>
<td>274</td>
</tr>
</tbody>
</table>

Fig. 3.4 illustrates the research landscape of learning analytics over time. It can be seen from Fig. 3.4a that learning analytics is a relatively new research area. Though publications started in 2001, it only became an active research area from 2011 onwards. This can be attributed to the fact that the term ‘learning analytics’ was only coined in 2011 by Siemens after the first Learning Analytics
Figure 3.3: Publication and Citation Counts of Countries in Learning Analytics
Conference in 2011. The figures depict a decline in 2017, attributed to the fact that we conducted this analysis in October of that year, so not all the publications for 2017 were not yet indexed in Scopus.

Although the publications in this area are from the year 2000, Fig. 3.4b illustrates that the core research on ‘learning analytics’ started in 2011. Previously, key terms such as educational data mining, data mining, e-learning, higher education and so on were prominent. Academic analytics was also considered an individual term. However, with the emergence of learning analytics, this is used cohesively with that term. Despite it being a relatively new area, a higher number of publications, up to 1,315, were found compared to other key terms in the query, implying that it is an emerging significant area for the research community.
(a) Evolution of Scientific Literature in the Field

(b) Key Terms Evolution

Figure 3.4: Evolution of Learning Analytics Research
Figure 3.5: Institute-wise Publications in Learning Analytics from 2000 to 2017
Table 3.3: Top 15 Institutions in Learning Analytics with Publication Counts from 2000 to 2017

<table>
<thead>
<tr>
<th>Sr.#</th>
<th>Institution</th>
<th>Publication Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Spanish University for Distance Education UNED, Spain</td>
<td>254</td>
</tr>
<tr>
<td>2.</td>
<td>University of Melbourne, Australia</td>
<td>237</td>
</tr>
<tr>
<td>3.</td>
<td>Deusto Institute of Technology, Spain</td>
<td>225</td>
</tr>
<tr>
<td>4.</td>
<td>Old Dominion University, US</td>
<td>218</td>
</tr>
<tr>
<td>5.</td>
<td>Tsinghua University, China</td>
<td>218</td>
</tr>
<tr>
<td>6.</td>
<td>University of Heldesheim, Germany</td>
<td>213</td>
</tr>
<tr>
<td>7.</td>
<td>University of Geneva, Switzerland</td>
<td>211</td>
</tr>
<tr>
<td>8.</td>
<td>University of Aizu, Japan</td>
<td>209</td>
</tr>
<tr>
<td>9.</td>
<td>Athabasca University, Canada</td>
<td>143</td>
</tr>
<tr>
<td>10.</td>
<td>Worcester Polytechnic Institute, US</td>
<td>77</td>
</tr>
<tr>
<td>11.</td>
<td>University of Hong Kong</td>
<td>45</td>
</tr>
<tr>
<td>12.</td>
<td>Tampere University of Technology, Finland</td>
<td>35</td>
</tr>
<tr>
<td>13.</td>
<td>Open University, UK</td>
<td>28</td>
</tr>
<tr>
<td>14.</td>
<td>Ho Chi Minh City University of Technology, VietNam</td>
<td>16</td>
</tr>
<tr>
<td>15.</td>
<td>Complutense University of Madrid, Spain</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 3.3 shows the publication output of top 15 institutions in this area. The Spanish University for Distance Education (UNED) supersedes other institutions by producing 254 publications, followed by the University of Melbourne, Australia, with a count of 237. Though the US appears on top in the countries, its institutes come second to Spain. China produces relatively fewer publications than other top countries; yet Tsinghua University, China, appears in the top five producing institutes. This implies that, while China does have fewer publications, some of its institutes are on par with other top worldwide institutes. Nonetheless, it appears to rank low in terms of citation counts. Fig 3.5 illustrates worldwide research activity in learning analytics w.r.t. institutions. The red dots denote the publication count of each institute; evidently the institutions of Spain, France, Germany, Australia, US, China and Japan are the most active in this research area.

The Fig. 3.6 shows the evolution in this research area, regions are mapped to examine the key areas active in producing publications during 2008 to 2017. Since there were no significant publications before 2008, a time window of 10 years was selected to analyse the learning analytics research landscape, further splitting into two windows, each of five years. A five-year time window was chosen to
identify the early producers and pioneers in this research area. Fig. 3.6 (left side), depicts the regions contributing to learning analytics from the beginning. The regions of Europe and US emerge as the top contributors, followed by China, Japan and Australia. The color scheme of the dots changes from light to dark, denoting the earliest and more recent work during the first five-year time window.

The second time window, illustrated in Fig. 3.6 (right side), ranges from 2013 to 2017, with darker circles representing areas with recent research activity. Evidently, this recent time range demonstrates greater activity, with newer regions contributing to the research community. As depicted, the regions of Europe and the US emerge as the top contributors, followed by newer regions including but not limited to Brazil, South Africa, India, New Zealand, United Arab Emirates and Malaysia. Overall, very sparse research activity is observed in the regions of South Africa, Egypt, United Arab Emirates and Russia, implying their reduced contribution to the Learning Analytics research community.

3.3.2 Discussion on Collaboration Network among Authors and Countries

In order to visualize the collaboration patterns among authors and countries, co-authorship networks are created for both authors’ affiliation and countries’ collaboration. Interestingly, the authors Pardon and Dawson are found to be the major contributors in this research area, in terms of the highest number of publications, with a count of 40 and 36 publications respectively. However, despite being major contributors, they are not highly cited by the research community, as shown in Fig. 3.7a. In the figure, the bars depict the publication counts of each author and the labels of each bar indicate the citation counts. Romero and Ventura, with a publication count of less than 25, are found to be the most cited authors, with a citation count of 1,785 and 1,777 respectively.

Further, an authors’ co-authorship network was constructed to analyse the authors and their collaborations, in order to determine the pioneers of this research community. Different cluster networks can be visualized between various authors, as depicted in Fig. 3.7b. Pardon and Dawson, the major contributors in the field, can be seen forming two different clusters, interconnecting with other author networks, implying their significance to the research community. However, some disconnected smaller clusters can also be observed that are not interconnected with others, implying the existence of smaller research communities and a lack of collaboration among such clusters. Furthermore, another phenomenon observed is the non-existence of the highly cited authors, Romero and Ventura, from the
Figure 3.6: Publication Output of Countries in Learning Analytics from 2008 to 2017
Figure 3.7: Publication/Citation Counts of Authors and Co-authorship Network in Learning Analytics
co-authored network. Though Romero can be observed in the figure, however it forms a very small cluster, implying its reduced participation in producing collaborated work in the research community. Although Romero and Ventura do not actively collaborate with other authors in the community, some of their articles are still highly cited by the research community, as depicted in Table 3.4.

In addition, a co-authorship network among countries from 2008 to 2017 is displayed in Fig. 3.8. A five-year time window was used to analyse the complete research activity of this research area. Fig. 3.8a shows the affiliation patterns of countries between 2008 and 2012. It shows 41 countries out of 58, including countries that have produced a minimum of one document in collaboration with another country. The size of the clusters represents the citations received by each source. The co-authorship network depicts the clusters where affiliation between different countries is most significantly observed. Significant clusters can be observed for regions of the US, UK, Spain, Germany, Australia and Canada, followed by the regions of China, Japan, Greece, Portugal and New Zealand, which form minute clusters. A distinct affiliation pattern cannot be observed among countries, as each one collaborates with others worldwide to produce innovative research. However, the US appears as the top collaborator, with a significant cluster size indicating its substantial contribution to the research community and justifying its increased publication count. The co-authorship network also depicts worldwide research activity in this research area, highlighting key productive countries.

The recent time window 2013 to 2017 shows 79 connected countries out 136, as depicted in Fig. 3.8b, generating a lattice of interconnecting networks between different nations worldwide, implying the contribution of several countries in learning analytics. This indicates that, in recent years, authors across the world have become more aware of the existence of this field and their acceptance towards it has evolved, generating a higher number of internationally co-authored publications. Countries such as Japan, Austria, Greece and India, which had formed negligible clusters in the early time window, form significant clusters in the recent time window. Additionally, the lattice illustrates the emergence of several countries that did not appear in the earlier time window, including but not limited to China, Brazil, France, Egypt, Norway and Sweden.
(a) Time window: 2008 to 2012

(b) Time window: 2013 to 2017

**Figure 3.8:** Co-authorship Network of Learning Analytics Among Nations from 2008 to 2017
3.3.3 Discussion on Citation Network among Sources

In order to analyse the Learning Analytics research landscape, a network of sources citing one another was constructed to identify the dominant and influential sources in the community, depicted in Fig. 3.9. The map shows 95 sources out of 822 using a minimum of five documents of a source as a threshold to highlight key source titles in the field. The size of the clusters represents the citations received by each source.

It was observed that conferences are more prevalent than journals, attributed to the fact that learning analytics is a relatively new and emerging research area. Conferences and journals interconnect by citing one another, sharing ideas and producing innovative research. The cluster size of most of the entities is the same, aside from a few, indicating that the participation level at these conferences and journals is equivalent. Some smaller circles indicate the emergence of new conferences and journals participating in the research community, indicating the evolution and growth of this research area. Some significant journals such as, Computer and Education, Computers in Human Behavior, Expert Systems with applications and International Journal of Technology Enhanced Learning can be noted. However, these are not specifically on learning analytics, hence cover a broader research range. This implies a lack of journals for this research field.

3.3.4 Discussion in Terms of Co-occurrence Network

Finally, to examine the key words associated within the field, a text-based map on ‘titles’ was constructed, as shown in Fig. 3.10, displaying frequently occurring key words using VOSviewer. The map shows 48 terms out of 5,284, using a minimum 20 number of occurrences of a concept as a threshold to highlight major research streams in the field. It represents a holistic view of the Learning Analytics research landscape, showcasing various terminologies associated with this domain. Here the related terminologies are clustered into one, indicating their correlation in terms of published articles relating to them.

Drilling down to the ‘learning analytic’ cluster, it can be observed that it deals with educational data including student assessment and evaluation, predicting their performance, tailoring teaching methods and practices, visualizing MOOC activities and developing applications to evaluate them,
**Figure 3.9**: Citation Network Among Source Titles
Figure 3.10: Title-based Thematic Analysis on Learning Analytics Keywords
modifying courses, designing systems and frameworks to help higher education and optimizing the
learning environment. Higher education is directly connected to learning analytics, implying the
influence it has on the education system. The cluster size of MOOCs and higher education is quite
small, implying their recent emergence in the Learning Analytics research space. A time window is
also displayed, representing the temporal evolution of the research area, indicating that the learning
analytics cluster formed from 2015 onwards.

Drilling down further, in order to explore the relationship between learning analytics and data-driven
decision-making, a network of author-defined key words was constructed, as illustrated in Fig. 3.11.
The map shows 157 terms out of 4285 using minimum 11 number of occurrences of a concept as
threshold to highlight major research terms in the field. It can be observed that decision-making is
directly connected to analytics, data and data mining. This implies that data-mining techniques are
used to perform analysis on the data, later to be used for decision-making. Since we are dealing
with an educational dataset, the insights inferred from analytical techniques are deployed in making
data-driven decisions. A temporal line is represented in this figure, representing decision-making as
a relatively new area from 2015.

3.3.5 Discussion on Recent and Most Significant Papers in Learning Analytics

Table 3.4 represents the top-10 highly cited publications in the research community. It can be ob-
served that publications by Romero and Ventura emerge as the top three most highly cited articles. A
full text analysis was performed on these publications to understand their context. These publications
give a holistic view of the current trends in educational data mining, discussing the state-of-the-art
and describing the content involved in this domain. It specifically emphasizes mining techniques to
extract meaningful information from student data [16, 17, 159].

Ferguson [160] discussed the significance of learning analytics, its distinction from academic analyt-
ics, the challenges of procuring datasets and the ethical issues associated with it, thus providing an
overview of this domain and forming the baseline for the literature on learning analytics. Kizilcec
[161] studied the behavior of users in MOOCs and identified various categories of users. The dis-
tinctions and similarities between educational data mining and learning analytics are also discussed,
indicating the overlap between the two communities; the former is inclined to automated discovery,
Figure 3.11: Author-defined Keyword Evolution in Learning Analytics
Table 3.4: Top 10 highly Cited Publications in Learning Analytics

<table>
<thead>
<tr>
<th>Sr#</th>
<th>Paper Names</th>
<th>Authors</th>
<th>Year</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>Educational data mining: a review of the state of the art</td>
<td>Romero and Ventura [16]</td>
<td>2010</td>
<td>465</td>
</tr>
<tr>
<td>3.</td>
<td>Data mining in course management systems: Moodle case study and tutorial</td>
<td>Romero et al., [17]</td>
<td>2008</td>
<td>376</td>
</tr>
<tr>
<td>4.</td>
<td>Learning analytics: drivers, development and challenges</td>
<td>Ferguson [160]</td>
<td>2012</td>
<td>236</td>
</tr>
<tr>
<td>6.</td>
<td>Social learning analytics</td>
<td>Shum and Ferguson [162]</td>
<td>2012</td>
<td>159</td>
</tr>
<tr>
<td>10.</td>
<td>A reference model for learning analytics</td>
<td>Chatti et al., [165]</td>
<td>2012</td>
<td>134</td>
</tr>
</tbody>
</table>
whereas the latter uses the inferred information for human judgement [163]. Chatti [165] provides a learning analytics reference model, consisting of data, stakeholders and analytical techniques, to yield detailed knowledge of the learning analytics cycle.

While the most highly cited papers are inclined to form ground rules for learning analytics, recent publications focus more on the practical advantages of learning analytics, such as examining student learning behavior to identify deep learning students who study with intent to gain conceptual knowledge and surface learners who are inclined towards bookish knowledge [166]. Pardo [167] investigated patterns of students who are more inclined to academic success and compared them with those of poor performers, using approaches of self-regulating learning and learning analytics cohesively. Casey & Azcona [168] identified weak students by classifying them on their performance when they used the taught concepts for the first time. A list of the recent publications for 2017, is given in Table 3.5.

### 3.3.6 Overall Discussion on the Bibliometric Study

In this section, a bibliometric study of the research productivity in learning analytics was presented using Scopus database over 2000 to 2017. Its research landscape was examined and explored to analyse it at various levels, including investigating the prominent countries, institutions and sources to visualize current trends in this field, summarized as follows:

- In the last five years, research in this area has evolved, and countries worldwide are participating and contributing in the research community.

- The US, Spain, Australia, UK and Germany emerge as the top countries in learning analytics in terms of publication output and citation count. This implies that these countries lay emphasis on their research activities in terms of both quantity and quality.

- The combination of country-level and institution-level analysis envisages the research landscape with US and Spain emerging as top countries, with their institutions contributing actively in this domain and surfacing in the top 10 institutes around the globe.

- China ranks low in terms of publication count and citation count, but a major portion of its publications is produced by a single institute. This may justify its lower rank in citation.
Table 3.5: Top 10 Recent Publications in Learning Analytics

<table>
<thead>
<tr>
<th>Sr.#</th>
<th>Paper Names</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Lostrego: a distributed stream-based infrastructure for the real-time gathering and analysis of heterogeneous educational data</td>
<td>Estevez-Ayres et al.,[169]</td>
</tr>
<tr>
<td>2.</td>
<td>Generating descriptive model for student dropout: a review of clustering approach</td>
<td>Iam-On &amp; Boongoen [170]</td>
</tr>
<tr>
<td>3.</td>
<td>Utilizing student activity patterns to predict performance</td>
<td>Casey &amp; Azcona [168]</td>
</tr>
<tr>
<td>5.</td>
<td>An analysis of collaborative problem-solving activities mediated by individual-based and collaborative computer simulations</td>
<td>Chang et al.,[172]</td>
</tr>
<tr>
<td>6.</td>
<td>On expressiveness and uncertainty awareness in rule-based classification for data streams</td>
<td>Le et al.,[173]</td>
</tr>
<tr>
<td>7.</td>
<td>Learning analytics: challenges and limitations</td>
<td>Wilson, et al.,[174]</td>
</tr>
<tr>
<td>8.</td>
<td>Using learning analytics to evaluate a video-based lecture series</td>
<td>Lau et al.,[175]</td>
</tr>
<tr>
<td>9.</td>
<td>Detecting learning strategies with analytics: links with self-reported measures and academic performance</td>
<td>Gasevic et al.,[166]</td>
</tr>
<tr>
<td>10.</td>
<td>Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance</td>
<td>Pardo et al.,[167]</td>
</tr>
</tbody>
</table>
Since it is a new emerging area, a discrete pattern of affiliations between countries was not encountered. Countries, worldwide, collaborate with one another in order to produce publications. However, regions of Europe and US emerge as the early producers, contributing the most to this research domain.

The Learning Analytics research area is still evolving, with the conference count being significantly greater than the journal count. In the coming years, more journals will surface to counter work in this domain.

The authors Pardo and Dawson are found to be the two most influential scholars with highest publication count, forming a collaborated-network with other authors in the research community. Overall, learning analytics resonates strongly with educational datasets, performing data-mining techniques on it, developing strategies to assess and evaluate learners’ performance, and assisting higher education decision-making. Learning analytics’ influence on higher education and its decision-making is a relatively new area, therefore no significant clusters for it were observed. In the coming years, with new and innovative research in this domain, more rational strategies will be developed to aid higher education decision-making.

3.4 Summary

This chapter presented the research landscape of learning analytics through a bibliometric study conducted through a Scopus-compatible query. It perused prominent countries, institutions and sources to visualize the current trends in this field. Moreover, it presented the evolution of eminent keywords applicable in this discipline, highlighting the evolving nature and growth of learning analytics in the research community.
Chapter 4

Academic Performance Prediction of Students and their Behavior

“Once we know something, we find it hard to imagine what it was like not to know it.” -

Chip & Dan Heath, Made to Stick

In this chapter and for our second contribution of this dissertation, we aim to leverage machine and deep learning techniques to predict the academic performances of students at the end of the course. The overall research agenda for this contribution is to measure the effectiveness of VLEs in predicting students’ performance, for timely intervention by the instructors, providing suitable pedagogical support. Moreover, determining learners’ behavior and various patterns associated with them, the performance of the students can be assessed by the extent of their interaction with the virtual environment. More specifically, students’ interactions with the deployed learning management tools generate an abundance of clickstream data, consequently reflecting their participation with the learning environment. The current students’ performance is indicative of their legacy data, such as their past performance in assessments and quizzes. The research questions to be addressed for this section are as follows:


(ii) Exploiting clickstream data generated through students interaction with the virtual learning management system to assess their performance.

The proposed contribution of this study, as derived from the above research objectives, is essential in identifying the effectiveness of the learner’s digital footprints in the VLE, including their legacy data and ascertaining interaction patterns of students’ at-risk of failure or withdrawal.
4.1 Dataset for Academic Performance Prediction of Students

The dataset is retrieved from the Open University Learning Analytics which provides demographic, clickstream behavior and assessment performance of 32,593 students over a course period of 9 months, from 2014-2015 [176]. It is comprised of 7 courses, referred to as modules, each presented at least twice and at different times in a year. The resultant marks of students are classified into four classes; distinction (3,024), pass (12,361), fail (7,052) and withdrawn (10,156).

4.1.1 Descriptive Analysis of the Dataset

The data descriptives assist in formulating an initial understanding and present a generic behavior of the data. Age, gender, highest education achieved by the students and their imd-band are some of the attributes that have been visualized and presented in Fig. 4.1, with each bar graph presenting the percentage of students with respect to their performance. From bottom to top, for each graph, the performance division is sorted as pass, fail and distinction, respectively. Note, that for this data exploration the original dataset with class imbalance was utilized with 7,052 fail instances and 15,385 pass instances, however further on for this contribution all other experiments, explorations and analysis are performed on the down-sampled balanced dataset. Fig. 4.1a shows performance distribution with respect to the three age bands provided in the dataset. It can be observed that students in the age band of 35-55 show better performance variation with lesser withdrawals and more pass, distinction instances as compared to students within the age-band of 0-35. The third age-band, of students greater than 55 years, constitutes of a very few number of students.

Furthermore, Fig. 4.1c shows performance variation with respect to gender, with M and F signifying males and females, respectively. Little performance variation is observed in terms of gender where males outnumber females; however, a very slight performance difference can be visualized with females performing better. According to the highest education achieved by a student, five measures exist in the dataset as provided in Fig.4.1d with majority of the students with an A-level or equivalent qualification. Between the first two education measures, which are also the highest in terms of the number of students, A-level or equivalent to it demonstrate better performance variation as compared to the other measure (having qualification lower than A-level). A generic trend can be observed where
Figure 4.1: OULA Dataset Descriptives for Age, Gender, Highest Education & IMD-Band
the post graduate students demonstrate better performances with fewer withdrawals and more distinction instances, however these are very few in number. Similar to post graduate instances, a very few number of instances are available with no formal qualifications; however they demonstrate poor performance behavior with a higher number of withdrawals and failed instances unlike post graduated instances.

Imd-band represents the index of multiple deprivation, signifying the economic deprivation of the area that the student belonged to, where 0-10% represent the most deprived areas and 90-100% represent the least deprived areas. This economic deprivation is categorized in terms of the employment opportunities, health facilities etc. A visual performance improvement is depicted with students demonstrating better performance distribution as the graph move towards lesser deprived areas (see Fig. 4.1b). Moreover, in order to understand some influential feature activities and their correlation with each other and on the basis of performance, a correlation heatmap is constructed, provided in Fig. 4.2. The corresponding feature codes and their description are provided in Table 1. This correlation matrix is constructed from the balanced dataset that has been primarily deployed in this study. Fig. 3 presents the feature activities as code and student performance (pass/fail) as the class label. The first left column of the matrix depicts the features correlation with the class label, with lighter colors denoting negligible impact to darker colors denoting an influential impact on the performance. A more in-depth analysis of the presented heatmap highlights some important features with respect to the class label. As illustrated in Fig. 4.2 graded as well as non-graded activities are observed to have an influential effect on performance, such as quiz and assignment related activities are graded and discussion forums, pdf resources, link to other sites and additional materials are non-graded. However a combination of both such activities is observed to be having a positive impact on the performance. Some more in-depth analysis are required to identify the individual impact of both types of activities.

The class labels pass and distinction are merged into one label ‘pass’, owing to the class imbalance problem. To predict students’ at-risk of failure, class labels ‘pass’, merged with ‘distinction’, and ‘fail’ are deployed on the model. The problem, hence, becomes a binary classification problem with 22437 unique students and ‘pass’ or ‘fail’ as class labels, where class 0 depicts pass and 1 represents fail. The withdrawal cases tend to have a distinct attitude where a student might withdraw in the first few weeks. Therefore, it is considered another category and withdrawals are predicted with the ‘pass merged with distinction’ cases. Again the problem is converted to a binary classification problem with 25,541 unique students and ‘withdrawn’, ‘pass’ as class labels (see Table 4.1). Similarly, to
Figure 4.2: Heat-map of all Feature Activities with their Feature Codes
predict students with distinctions, they are employed on the model with ‘pass’ and ‘fail’. Overall four categories of the dataset are computed with respect to the class labels. Each category represents a binary classification system, to address the objectives of this contribution. Further, for other objectives as mentioned in Section 1.6, OULA data was employed, however with different processing steps.

Table 4.1: Class Label Categories

<table>
<thead>
<tr>
<th>Categories</th>
<th># of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass-Fail</td>
<td>22,437</td>
</tr>
<tr>
<td>Withdrawn-Pass</td>
<td>25,541</td>
</tr>
<tr>
<td>Distinction-Fail</td>
<td>10,076</td>
</tr>
<tr>
<td>Distinction-Pass</td>
<td>15,385</td>
</tr>
</tbody>
</table>

The dataset consists of several files, consisting of students’ demographics, VLE data, interaction with the VLE represented by clickstream data, assessment marks and modules information. The VLE data is comprised of 20 activities, where each activity type represents an important element of the VLE, and students interact with the VLE to access course material, through these activities. Further, in this dissertation, these activities with their names and intended meanings have been provided in Chapter 5.

Table 4.2: Features Representing Before and After Course Activities of Students

<table>
<thead>
<tr>
<th>Features</th>
<th>Feature Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>highest_education</td>
<td>highest education of a student when registering for a module</td>
</tr>
<tr>
<td>F2</td>
<td>imd_band</td>
<td>Index of Multiple Deprivation band of the area where student resided during the module work</td>
</tr>
<tr>
<td>F3</td>
<td>age_band</td>
<td>students age band</td>
</tr>
<tr>
<td>F4</td>
<td>num_of_prev_attempts</td>
<td>number of times a student has attempted a particular module</td>
</tr>
<tr>
<td>F5</td>
<td>studied_credits</td>
<td>number of credits being currently studied by the student</td>
</tr>
<tr>
<td>F6</td>
<td>disability</td>
<td>if student has stated a disability</td>
</tr>
<tr>
<td>F7</td>
<td>AC T. Clicks</td>
<td>after course total clicks of a student for a particular module</td>
</tr>
</tbody>
</table>

(To be continued)
<table>
<thead>
<tr>
<th>Features</th>
<th>Feature Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F8</td>
<td>BC T. Clicks</td>
<td>before course total clicks of a student for a particular module</td>
</tr>
<tr>
<td>F9</td>
<td>AC DataPlus</td>
<td>after course total clicks on the additional information such as videos, audios, sites etc.</td>
</tr>
<tr>
<td>F10</td>
<td>AC DualPane</td>
<td>after course total clicks on the information on site and activity related to that information</td>
</tr>
<tr>
<td>F11</td>
<td>AC Folder</td>
<td>after course total clicks on the files relevant to course</td>
</tr>
<tr>
<td>F12</td>
<td>AC Forumng</td>
<td>after course total clicks on the discussion forum</td>
</tr>
<tr>
<td>F13</td>
<td>AC HomePage</td>
<td>After course total clicks on the course homepage</td>
</tr>
<tr>
<td>F14</td>
<td>AC OuCollaborate</td>
<td>after course total clicks on the online video discussions</td>
</tr>
<tr>
<td>F15</td>
<td>AC Oucontent</td>
<td>after course total clicks on the contents of the assignment</td>
</tr>
<tr>
<td>F16</td>
<td>AC OuWiki</td>
<td>after course total clicks on the Wikipedia content</td>
</tr>
<tr>
<td>F17</td>
<td>AC Page</td>
<td>after course total clicks on the information related to course</td>
</tr>
<tr>
<td>F18</td>
<td>AC Questionnaire</td>
<td>after course total clicks on the questionnaires related to course</td>
</tr>
<tr>
<td>F19</td>
<td>AC Quiz</td>
<td>after course total clicks on the course quiz</td>
</tr>
<tr>
<td>F20</td>
<td>AC RepeatActivity</td>
<td>after course total clicks on the course contents from previous weeks</td>
</tr>
<tr>
<td>F21</td>
<td>AC Resource</td>
<td>after course total clicks on the pdf resources such as books</td>
</tr>
<tr>
<td>F22</td>
<td>AC SubPage</td>
<td>after course total clicks on the other sites enabled in the course</td>
</tr>
<tr>
<td>F23</td>
<td>AC Url</td>
<td>after course total clicks on the links to audio/video contents</td>
</tr>
<tr>
<td>F24</td>
<td>Tc_Activity</td>
<td>total course activity clicks for each student for a particular module</td>
</tr>
<tr>
<td>F25</td>
<td>PreA-1</td>
<td>number of clicks for a student, one day before the assessment deadline submission</td>
</tr>
<tr>
<td>F26</td>
<td>BC HomePage</td>
<td>before course total clicks on the course homepage</td>
</tr>
</tbody>
</table>

(To be continued)
<table>
<thead>
<tr>
<th>Features</th>
<th>Feature Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F27</td>
<td>OnAsClicks</td>
<td>number of clicks for a student, on the assessment day deadline submission</td>
</tr>
<tr>
<td>F28</td>
<td>ModuleAsigns</td>
<td>number of assessments in a module</td>
</tr>
<tr>
<td>F29</td>
<td>LateAsignsSub</td>
<td>number of assignments submitted late by a student</td>
</tr>
<tr>
<td>F30</td>
<td>PostA-1</td>
<td>number of clicks for a student, one day after the assessment deadline submission</td>
</tr>
<tr>
<td>F31</td>
<td>BC Oucontent</td>
<td>before course clicks on the contents of the assignment</td>
</tr>
<tr>
<td>F32</td>
<td>BC SubPage</td>
<td>before course clicks on the subpage activity</td>
</tr>
<tr>
<td>F33</td>
<td>AC External Quiz</td>
<td>after course clicks on the external quiz activity</td>
</tr>
<tr>
<td>F34</td>
<td>BC Forumng</td>
<td>before course clicks on the discussion forum</td>
</tr>
<tr>
<td>F35</td>
<td>BC Url</td>
<td>before course clicks on the links to audio/video contents</td>
</tr>
<tr>
<td>F36</td>
<td>AC Ouelluminate</td>
<td>after course clicks on the online tutorial sessions</td>
</tr>
<tr>
<td>F37</td>
<td>BC Page</td>
<td>before course clicks on the information related to course</td>
</tr>
<tr>
<td>F38</td>
<td>BC OUwiki</td>
<td>before course clicks on the Wikipedia content</td>
</tr>
<tr>
<td>F39</td>
<td>BC Quiz</td>
<td>before course clicks on the course quiz</td>
</tr>
<tr>
<td>F40</td>
<td>AC Glossary</td>
<td>after course clicks on the basic glossary related to contents of course</td>
</tr>
<tr>
<td>F41</td>
<td>BC HtmlActivity</td>
<td>before course clicks on the interactive html page</td>
</tr>
<tr>
<td>F42</td>
<td>AC HtmlActivity</td>
<td>after course clicks on the interactive html page</td>
</tr>
<tr>
<td>F43</td>
<td>BC Ouelluminate</td>
<td>before course clicks on the ouelluminate activity</td>
</tr>
<tr>
<td>F44</td>
<td>BC Glossary</td>
<td>before course clicks on the basic glossary related to contents of course</td>
</tr>
<tr>
<td>F45</td>
<td>BC Oucollaborate</td>
<td>before course clicks on the online video discussions</td>
</tr>
</tbody>
</table>

(To be continued)
A range of analytical techniques are employed on this dataset to predict students’ performance by identifying students likely to fail their course, early prediction of at-risk and withdrawal students and identifying patterns of students passing with distinction. To encompass the objectives of our study demographics, assignments and total clicks of each student are computed to analyse the behavior of each student. A detailed description of all the features employed is described below in the pre-processing section.

### 4.2 Data Pre-processing

For this contribution OULA data is processed such that it includes student activity with the VLE portal and static demographics data. In order to evaluate students’ performance, the OULA dataset is
pre-processed to obtain meaningful attributes and features reflecting their performance.

4.2.1 Demographics and VLE Portal Data

The constructed dataset includes the students’ demographics and VLE portal information, such as the number of clicks for each activity type provided in the dataset. Moreover, students are able to access the VLE a few weeks before the start of the module; hence the number of clicks for each activity type, before the module started, is considered another range of features. Similarly, the data relevant to the assessments and number of late assignments submitted by a student, are also included in the dataset. A list of these 54 features is provided in Table 4.2. The dataset, comprising students’ demographics and VLE portal information, is sparse with 54 features and one binary class label for each category. Sparse feature reduction technique is employed to find the optimal features impacting student’s performance. To cater the data sparsity on this feature set, sparse feature reduction technique using truncated Singular Value Decomposition (SVD) is deployed, to identify the top 30 significant attributes. This technique performs linear dimensionality reduction by means of truncated SVD. SVD corresponds to dimensionality reduction by means of a low rank approximation, assuming there is a matrix ‘X’ and it is to be estimated to a rank ‘r’ matrix ‘X̂’, where r = rank(X). The estimated inaccuracy is computed by the Frobenius norm, which corresponds to the square root of the squared error. It reduces the ‘squared distance’ between each data point. The problem then becomes: \(\min_{X^\wedge} |X - X^\wedge|_F\) such that, rank \((X^\wedge) = r\). The rank ‘r’ denotes the dimension of the linear sub-space of the data points. \((X^\wedge)\) is the computed low rank estimation of the data matrix implying that the data points in ‘X’ matrix are compressed into a r-dimensional sub-space. In terms of centered data SVD corresponds to Principle Component Analysis (PCA), which can be defined as reducing the estimation error [177]. However, in terms of non-centered data, such as in this case where the data is sparse, SVD instead of the covariance matrix, is expressed as the Eigen-decomposition of the matrix XTX. A list of the 30 features (F1-F30) selected by the sparse reduction technique (SVD) is provided in Table 4.2, where each feature is listed with respect to its significance on the student’s performance. The demographics data is available in the OULA dataset. However, the VLE portal information is computed through the available features in the dataset.
4.3 Artificial Neural Networks (ANN)

The Deep Artificial Neural Network or ANN classification model is employed to learn the predictive function of predicting students’ at-risk of failure, those likely to withdraw from their courses, early prediction of withdrawal students, and determining the students who outperform others with distinction.

Deep learning methods are referred to as representation-learning methods constituting several layers of non-linear modules. This enables the system to be proficient enough to learn complex functions, making it robust enough to be sensitive to intricate and minute specificities. In contrast to the statistical methods, ANNs facilitate generalization, which enables them to correctly infer hidden patterns from the data, assisting in making data driven assumptions [178]. The network learns from the examples in the training data, thus increasing the training split leads to a more robust accuracy [179]. The stack of non-linear layers between the input and output layers are referred to as hidden layers, weights are adjusted in the layers through stochastic gradient, to calculate the error computed in classifying and predicting correct answers. An ANN with multiple non-linear layers, also referred to as a Multiple Layered Perceptron, is capable of implementing complicated input functions [42].

An ANN is composed of inter-connected objects referred to as processing units. Each unit takes the weighted sum of inputs and produces an output. An ANN has a layered architecture where neurons are assembled in consecutive layers and output of each layer is fed to its successive layer. A combination of non-linear functions is deployed on the input, hidden and output layers. In binary classification the output layer is activated with ‘sigmoid’ function, because sigmoid squashes the values into 0 or 1. For a supervised learning problem, the weight vector ‘w’ for the ‘n\textsuperscript{th}’ unit at time instant ‘t’ is defined as shown in equation 4.1.

\[
w_i(t + 1) = w_i(t) + \Delta w_i t \tag{Equation 4.1}
\]

where \(\delta w_i t\) denotes the change in the weight vector. For each instance the weight vector is changed with respect to the difference in the actual and desired output,
\[ \Delta w_i = \eta \times (D \times Y)I_i \quad \text{(Equation 4.2)} \]

where \( \eta \) is the learning rate of the model, \( D \) and \( Y \) are the desired and actual outputs, respectively and \( I_i \) denotes the \( i^{th} \) input instance. The model is trained to find appropriate weight vector values that correspond to a robust prediction [180].

In this section, ANN is employed to predict the students’ performance, in terms of identifying students’ at-risk of failure in their modules. After feature selection, min / max scaling is employed to normalize the data. The data is split into train test and is fed to the neural network in the form of a feature vector. Fig. 4.3 illustrates the proposed architecture level system of the ANN. For each of the four categories defined in Table 4.1, ANN with three hidden layers is implemented, where each hidden layer is activated with ‘relu’ function and the output layer with ‘sigmoid’ function.

### 4.4 Results: Experimental Setup

This chapter presents the experimental setup for the over-all data including student activity with the VLE portal and demographics data. For each of the defined award-gap category different experiments were performed with different parameters, however optimized results were obtained by deploying the deep ANN model with three hidden layers of 50, 20 and 10 neurons, a batch size ranging from 32 to 64 and either ‘adam’ or ‘rmsprop’ as optimizers.

#### 4.4.1 Results using Demographics and VLE Portal Data

To predict students’ performance, considering their demographics and overall portal VLE information, 30 features, through sparse reduction technique were selected from a total of 54 features, provided in Appendix A Table 4.2. From these 54 features F1-F30 were selected through sparse reduction technique.
Figure 4.3: Proposed Architecture using Artificial Neural Network for Predicting Student Performance at the End of a Course
### Table 4.3: Evaluation of the Results of Students Demographics & VLE Portal Data

<table>
<thead>
<tr>
<th>Categories</th>
<th>Techniques</th>
<th>Accuracy%</th>
<th>Loss%</th>
<th>Precision%</th>
<th>Recall%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass/Fail</td>
<td>ANN</td>
<td>84.48</td>
<td>0.385</td>
<td>0.86</td>
<td>0.617</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>67.94</td>
<td>0.320</td>
<td>0.46</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>82.35</td>
<td>0.161</td>
<td>0.84</td>
<td>0.83</td>
</tr>
<tr>
<td>Distinction/Fail</td>
<td>ANN</td>
<td>86.40</td>
<td>0.295</td>
<td>0.78</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>72.03</td>
<td>0.279</td>
<td>0.78</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>80.08</td>
<td>0.191</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>Distinction/Pass</td>
<td>ANN</td>
<td>80.54</td>
<td>0.495</td>
<td>0.69</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>79.82</td>
<td>0.202</td>
<td>0.64</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>80.01</td>
<td>0.199</td>
<td>0.77</td>
<td>0.80</td>
</tr>
<tr>
<td>Withdrawn/Pass</td>
<td>ANN</td>
<td>94.70</td>
<td>0.136</td>
<td>0.94</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>88.31</td>
<td>0.116</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>89.44</td>
<td>0.105</td>
<td>0.89</td>
<td>0.89</td>
</tr>
</tbody>
</table>

### 4.4.1.1 Predicting On-risk Students

The problem of predicting at-risk students was converted to a binary classification problem by defining two classes ‘pass’ and ‘fail’. Table 4.3 represents the evaluation results of our Deep ANN with the baseline models Support Vector Machine (SVM) and Logistic Regression (LR). It can be observed that Deep ANN yields a better accuracy in predicting students’ at-risk of failure. Therefore, according to the features provided in Table 4.2 in Appendix A, Deep ANN produces better predictions by yielding an accuracy of 84%. The significant features associated with the desired student performance are illustrated in the heat-map (part (b) of Fig. 4.4).
Figure 4.4: Correlation Heat-maps among Features and Class Categories
Figure 4.4: Correlation Heat-maps among Features and Class Categories
It can be observed that demographic features, previous education history, legacy data of students including assessments submission and the overall activity of a student in a module, are significantly impacting student’s performance.

### 4.4.1.2 Predicting Students with Distinction

To predict students with distinction, the problem was considered a binary classification problem by defining two sets of categories, a) ‘distinction’ and ‘pass’ b) ‘distinction’ and ‘fail’. As depicted in Table 4.3, ANNs accuracy in ‘distinction-pass’ category does not have a significant difference from other baseline models. This may be attributed to the class imbalance problem; also another justification can be that both classes do not have significant differences in their patterns. In the ‘distinction-fail’ category ANN yields more accurate results than other baseline models, as shown in Table 4.3. The significant features associated with the desired student performance are illustrated in the heat-map (parts (c) and (d) of Fig 4.4. To predict ‘distinction’ instances from failures, students’ portal information is observed to be positively associated with the performance, whereas demographic features including geographical region and education history are seen to be negatively associated with performance.

It can be observed that demographic features, previous education history, legacy data of students including assessments submission and the overall activity of a student in a module, are significantly impacting the student’s performance.

### 4.4.1.3 Predicting Withdrawals Cases

To predict ‘withdrawal’ instances, ‘withdrawals’ and ‘pass’ are the two classes that were defined. The pattern of withdrawals will distinguish from ‘pass’ instances, since withdrawals tend to left/drop out of their modules. The evaluation results show an accuracy of 94% in predicting the two aforementioned class labels, provided in Table 4.3. Comparing the ANN results with SVM and LR, ANN tends to give a higher accuracy in predicting withdrawals. The significant features associated with the desired student performance are depicted in the heat-map (part (a) of Fig. 4.4). In case of withdrawals,
the activities before the initiation of the module are observed to be significantly impacting the performance. Students interested in registering in a particular module are more likely to be active on the portal before the module starts. It can be observed that demographic features, previous education history, legacy data of students including assessments submission and the overall activity of a student in a module, are significantly impacting the student’s performance.

4.5 Summary

This chapter presented some analysis to predict the academic performances of students by leveraging machine learning techniques. Its constituted of the demographics and the computed clickstream information of the students, encompassing the clicks before and after the module initiation. Moreover, the analysis also presented some influential attributes significantly impacting the academic performances such as assessment submission, their previous education history and demographics.
Chapter 5

Early Performance Prediction of Students in Self-paced Online Learning

“Learning is never done without errors and defeat.”

– Vladimir Lenin

In this chapter, we intend to early predict the academic performances of students in an on-going course. This chapter presents a two-fold analysis in the early prediction of performances, pertaining to quarterly and week-wise granularity of student engagement patterns. The research questions to be addressed for this contribution are as follows:

(i) Compare comprehensively how deep learning algorithms perform vs. other conventionally used machine learning techniques to predict students at risk of failing a course accurately. Moreover, many studies lack the evaluation of effects of hyper-parameter tuning on prediction accuracy;

(ii) Identify specific factors that are the highest predictors of students at risk in order to inform the development of pedagogical interventions [46, 47, 112, 181];

(iii) Predict students at-risk as early in a course as possible to enable the development of an early warning system and to inform timely interventions.

5.1 Dataset for Students’ Early Performance Prediction

This contribution reports on the findings of the analysis performed on the accessible online dataset from OULA [176], comprising the log data representing students’ engagement with an online learning platform, student demographic data, and assessment-related scores and information in a raw format. Engagement of the students with the online environment is presented in the form of clicks and hence is also referred to as the clickstream data. A detailed description of the OULA data has been presented in Section 4.1.
The dataset contains student demographic data, clickstream data including students’ engagement with the Moodle platform, and assessment related information such as assignment submissions and scores. The dataset comprises seven courses (modules), with each course being assigned a code and taught through the Moodle learning platform at different intervals in a year. The duration of each course varies, with a maximum duration ranging from 259 to 264 days. The combination of course name and course offerings at different times throughout the year resulted in 20 unique course offerings. The data set distinguishes between four performance groups: distinction, pass, fail, and withdrawal. The students’ engagement with the VLE is characterized into 20 different activity types with each activity referring to a specific action, such as downloading lectures, viewing course contents, attempting quizzes, and logging in to the homepage. The names of each of these activity types, as provided in the dataset are as follows: dataplus, dualpane, externalquiz, folder, forumng, glossary, homepage, htm-lactivity, oucollaborate, oucontent, ouelluminate, ouwiki, page, questionnaire, quiz, repeatactivity, resource, sharedsubpage, subpage and url; with the description of each feature activity, total clicks recorded for each activity and average clicks per student for both pass and fail students, provided in Table 5.1. The activity feature codes are also presented in this table and are referred to in the remainder of this section.

5.1.1 Data Processing for Quarterly VLE Clickstream Data

The foremost approach for this contribution is the early intervention for all the award-gap categories (as provided in Table 4.1). To address early intervention, modules are divided into quarters. The duration of a module is nine months, so each module is divided into 4 quartiles and for each quartile the clickstream data of each activity is computed. For each quarter temporal features for each activity are computed, that change with respect to each quarter. This forms another array of derived features, in order to assess the most significant time in a module affecting a student’s performance, which can further be utilized for early intervention by the academia. Each module quarter contains the computed clickstream data of the temporal longitudinal activities, which vary with respect to each quarter. Each new quarter is computed by adding clickstream data of its previous quarters, such that Q2 will contain the clickstream activities of Q1 and Q2 cumulatively. Similarly, Q3 and Q4 are computed, with each of them including the clickstream information of their previous quarters. The activities in OULA VLE have been discussed previously in the dataset section. The number of features in each quarter
Table 5.1: Description of each Activity Feature in OULA Dataset, Total Clicks and Average Clicks per Class

<table>
<thead>
<tr>
<th>F. Code</th>
<th>Activity</th>
<th>Description</th>
<th>Total clicks</th>
<th>Avg clicks/pass</th>
<th>Avg clicks/fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>DataPlus</td>
<td>supplementary information and clicks on videos, audios, sites.</td>
<td>46,817</td>
<td>3.0167</td>
<td>0.238</td>
</tr>
<tr>
<td>F2</td>
<td>DualPane</td>
<td>engagement the site and activity related to that information</td>
<td>18,931</td>
<td>1.1116</td>
<td>0.3224</td>
</tr>
<tr>
<td>F3</td>
<td>External Quiz</td>
<td>engagement with the external quiz activity</td>
<td>54,540</td>
<td>3.2533</td>
<td>1.2717</td>
</tr>
<tr>
<td>F4</td>
<td>Folder</td>
<td>engagement the files relevant to course</td>
<td>5,397</td>
<td>0.3533</td>
<td>0.0712</td>
</tr>
<tr>
<td>F5</td>
<td>Forumming</td>
<td>engagement with the discussion forum</td>
<td>6,913,303</td>
<td>411.9130</td>
<td>104.9300</td>
</tr>
<tr>
<td>F6</td>
<td>Glossary</td>
<td>engagement with the course contents</td>
<td>75,411</td>
<td>4.5008</td>
<td>1.1284</td>
</tr>
<tr>
<td>F7</td>
<td>HomePage</td>
<td>engagement with the course homepage</td>
<td>5,887,338</td>
<td>342.6710</td>
<td>106.7500</td>
</tr>
<tr>
<td>F8</td>
<td>HtmlActivity</td>
<td>engagement the interactive html page</td>
<td>3,676</td>
<td>0.1974</td>
<td>0.1021</td>
</tr>
<tr>
<td>F9</td>
<td>Oucollaborate</td>
<td>engagement the online video discussions</td>
<td>98,046</td>
<td>5.7865</td>
<td>1.6068</td>
</tr>
<tr>
<td>F10</td>
<td>Oucontent</td>
<td>engagement the contents of the assignment</td>
<td>10,046,916</td>
<td>592.0995</td>
<td>166.4800</td>
</tr>
<tr>
<td>F11</td>
<td>Ouelluminate</td>
<td>engagement the online tutorial sessions</td>
<td>33,144</td>
<td>1.7808</td>
<td>0.9188</td>
</tr>
<tr>
<td>F12</td>
<td>Ouwiki</td>
<td>engagement the Wikipedia content</td>
<td>835,103</td>
<td>50.4855</td>
<td>11.1170</td>
</tr>
<tr>
<td>F13</td>
<td>Page</td>
<td>engagement the information related to course</td>
<td>39,498</td>
<td>2.20323</td>
<td>0.9212</td>
</tr>
<tr>
<td>F14</td>
<td>Questionnaire</td>
<td>engagement the questionnaires related to course</td>
<td>61,297</td>
<td>3.7463</td>
<td>0.7288</td>
</tr>
<tr>
<td>F15</td>
<td>Quiz</td>
<td>engagement with the course quiz</td>
<td>6,383,071</td>
<td>367.9307</td>
<td>123.4400</td>
</tr>
<tr>
<td>F16</td>
<td>RepeatActivity</td>
<td>engagement with the course contents from previous weeks</td>
<td>5</td>
<td>0.0001</td>
<td>0.0000</td>
</tr>
<tr>
<td>F17</td>
<td>Resource</td>
<td>engagement with pdf resources such as books</td>
<td>876,295</td>
<td>48.6829</td>
<td>20.8640</td>
</tr>
<tr>
<td>F18</td>
<td>SharedSubPage</td>
<td>shared information between courses and faculty</td>
<td>121</td>
<td>0.0065</td>
<td>0.0034</td>
</tr>
<tr>
<td>F19</td>
<td>SubPage</td>
<td>clicks on the other sites enabled in the course</td>
<td>2,814,113</td>
<td>158.5762</td>
<td>62.2100</td>
</tr>
<tr>
<td>F20</td>
<td>Url</td>
<td>engagement with the links to audio/video contents</td>
<td>458,514</td>
<td>26.4078</td>
<td>8.9138</td>
</tr>
</tbody>
</table>

corresponds to the number of activities in the VLE, however the clickstream information of each quarter varies.

5.1.2 Data Processing for Week-wise Student Engagements

The OULA dataset, acquired in a raw form consisting of several log files, was transformed into a structured format in order to understand student behavior and obtain conclusive results [182]. The entire pipeline used for data processing and analysis is shown in Fig. 5.1. Feature engineering steps
were performed to obtain relevant clickstream data for each student according to their engagement with the VLE. The activity-related features, depicting student engagement, were formulated in oracle 11g where several inter-related log files were joined to obtain the required clickstream data for each student, depicting their interaction with the learning platform. The data is then transformed into a homogenous vector comprising the overall cumulative engagement data for each student, fed to the classifiers. Furthermore, this clickstream data was aggregated in a sequential week-wise manner, with each week forming a vector comprising of the engagement pattern of the students’ interactions and passed to the sequential LSTM classifier. Such deep learning classifiers with their learning ability of hierarchical representation enable an effective performance prediction.

5.2 Handling Class Imbalance

A class imbalance was observed in the dataset with 15,385 ‘pass’ instances exceeding the 7,052 ‘fail’ instances. This inhibited the performance of the prediction; therefore, different class balancing techniques were employed to find the optimal one for this dataset. In the case of up-sampling of the ‘fail’ instances, each instance would have been repeated since there were twice as many instances in the ‘paas than in the ‘fail’ class. Such an up-sample method may overfit the training process and produce biased results, inhibiting the learning process of the training set [118, 183, 184]. Similarly, down-sampling of the ‘pass’ instances may eliminate some important interaction patterns of students.

In order to avoid the issues associated with up- and down-sampling, a cluster-based sampling was performed on the ‘pass’ class. A k-means clustering method was applied, which works by identifying random centroids and assigning each instance to its nearest centroid. Such a clustering enables the instances to be grouped with respect to their similarity [161]. The ‘pass’ class was clustered into two clusters, and equal proportions of students were extracted from each cluster. A clustering approach assisted in capturing the major themes in the ‘pass’ class without losing any significant information of pass or distinction instances [185]. Since the ‘pass’ class constitutes of students that passed the courses and those that passed with distinction, therefore this class encompasses students with varying behavior. To downsample the ‘pass’ class, random downsampling would have been a threat to the validity of the results. Therefore, students with similar interaction patterns were clustered into groups and then down sampled with equal proportions of students from both clusters. This approach for
Figure 5.1: Data Processing Steps for the Early Prediction Model using LSTM
down sampling enabled the inclusion of all the significant interaction patterns of students and also
decreased the risk of losing important information from the majority class. A consistent set of student
ids were kept in the aggregated and week-wise performance prediction analysis.

5.3 Modeling Approaches for Students Performance Prediction

This contribution followed a two-fold analysis approach. First, it included several classic machine
learning algorithms (base-lines) that are commonly used for the prediction of student performance in
the literature. Deep learning models were trained to compare their performance over the baselines.
A second analysis was performed to predict the early performance of students through a sequential
week-wise analysis of their interactions with the VLE.

5.3.1 Conventional Machine Learning Models

A series of conventional machine learning algorithms, commonly used by the educational research
community to predict at-risk students, were deployed to predict students’ performance.

**Logistic Regression (LR):** LR is one of the most commonly used methods in the educational commu-
nity. LR predicts the probability of a categorical dichotomous variable (class label such as pass/fail)
from a number of independent predicting attributes [186, 187]. Mathematically, the function is rep-
resented as given in equation 5.1:

\[
\ln \frac{p}{1 - p} = b_0 + b_1 x
\]  

(Equation 5.1)

Where \(b_0\) is the constant moving the curve towards left or right, \(b_1\) is the slope defining steepness,
and $p$ is the logistic model, represented in equation 5.2:

$$p = \frac{1}{1 + e^{-b_0 + b_1x}} \quad \text{(Equation 5.2)}$$

**Support Vector Machine (SVM):** It works by separating different categories of data through a hyperplane [188]. The task is to find an optimized hyperplane such that it separates different categories under study (class labels). In terms of binary classification, SVM will find the maximum margin between the two classes (such as pass/fail). For a case where the classes do not separate linearly, a non-linear SVM kernel finds an optimal surface. Since the OULA dataset is not linearly separable, multiple experiments with different kernels were performed to find optimal results. Kernels were used for hyperplane learning, where the linear kernel is a dot product between the input $y$ and the support vector $y_i$, calculated as represented in equation 5.3:

$$f(y) = B_0 + \text{sum}(a_i \ast (y, y_i)) \quad \text{(Equation 5.3)}$$

where the inner product of each input $y$ is calculated with all the support vectors in the training data, and $B_0$ and $a_i$ are the coefficients computed for each input from the training data.

**Decision Tree (DT):** A DT represented through a tree-like structure showcases internal nodes as rectangles and leaves as ovals. It is primarily based on the concept of partitioning, where the data is split based on one variable such that either all the attributes have been used one by one or the data in each node belongs to one of the class labels (e.g., pass/fail). These nodes are referred to as leaves and each leaf contains a class related to the dataset. The data is split on the basis of a score function, responsible for computing the purity of each node and selecting the variable that generates the purest nodes [189]. For this study, the Gini index and entropy information gain were used to calculate the purity of each node for each of the available variables.

Entropy performed better than the other measure, although generically significant performance differences were not observed between the two concepts. DT was implemented with entropy and Gini index, with the depth ranging from 3 to 6 and minimum leaves 4 to 7. The increasing depth of a DT leads to over-fitting, therefore, a range has to be set. From these experiments, entropy with a depth of
4 and minimum leaves of 6 was found to be the best fit amongst other DT experiments.

**Gradient Boosting Trees (GBT):** GBT algorithms are used for supervised regression and classification problems. Like other supervised algorithms, GBT also utilizes training data with multiple features to predict the target class. These trees are constructed greedily from the root node and split on the basis of mean square error. GBT is also widely applicable to the education community. It is an ensemble approach, where a new model is generated by predicting the residual errors of the previous models. It makes use of the gradient descent algorithm to minimize the residual errors and improve prediction.

**K-nearest Neighbor (KNN):** KNN is a non-parametric classifier and, unlike other classifiers, does not train on the available data attributes. It calculates the distance of each sample with its neighbors to designate a class label to it through majority voting [190]. In this study, the Euclidean distance metric was computed between samples, and KNN with values of 3 and 5 were used, that is, 3 and 5 students with the most similar grades were used to find optimal results.

**Artificial Neural Network (ANN):** The structure of an ANN is composed of many non-linear levels, where each level plays a significant role in transcending the learned representation in a more abstract manner to its higher layers, consequently assisting in learning complex functionalities [179, 191]. As opposed to the conventional statistical approaches, such self-adaptive techniques effectively determine the underlying associations between the data, by generalizing the input sequence and learning from it through back-propagation phenomena [178]. ANNs, imitating the human brain, consist of several neurons and layers, and with their added generalizability attribute assists inconclusively inferring hidden patterns from the data. Between the non-linear layers, the adjustment of weights assists in calculating the prediction error. The weight adjustment at a particular time t for a unit ‘n’ is defined in equation 5.4:

$$w_i(t + 1) = w_i(t) + \Delta w_i t$$  \hspace{1cm} (Equation 5.4)

where $\delta w_i t$ represents the change in the weights. For each input sample, the weight is adjusted with
respect to the difference between the predicted and actual results, as given in equation 5.5:

\[ \Delta w_i = \eta \times (D \times Y)I_i \]  

(Equation 5.5)

where \( \eta \) denotes the learning rate of the classifier, D and Y are the desired and actual outputs, respectively, and \( I_i \) denotes the \( i^{th} \) input instance. Several experiments were performed to tune the hyper-parameters for ANN and get optimal results, and the details have been provided in Section

5.3.2 Time-series Data Modeling

RNNs explicitly intend to learn the long-term dependencies in the sequential data through a recursive loop at each cell that supports it to keep a check on the previous input data as par with the existing input. The neurons’ weight in RNN is updated by backpropagation through time that enables the transmission of the error and the gradient over the entire vector sequence [192]. Though RNNs were designed for long-term dependencies, existing evidence provided by empirical research demonstrates otherwise [182, 193]. The RNN cells inhibit the transfer of error to distant past due to the issues emerging because of vanishing and exploding gradients [183, 194]. To resolve these issues, LSTM was introduced, where a memory cell is augmented in the network, enabling it to retain longer sequences [195]. The LSTM block is comprised of an additional memory unit and three gates responsible for regulating the flow of information, which is input i, output, o, and forget f gates. For any specific time instant \( t \), the input gate it is responsible for administering the information in the cell, forget gate \( f_t \) controls the information that should be overlooked, memory cell \( C_t \) is responsible for retaining the past information and output gate \( o_t \) manages the representation of the delivered output. The computation of \( f_t \) is defined in equation 5.6:

\[ f_t = \sigma W_f[h_{(t-1)}, x_t] + b_f \]  

(Equation 5.6)

where \( \sigma \) is the sigmoid function, \( x_t \) is the input at a time instant \( t \), \( h_{t-1} \) is the previous hidden state, \( W_f \) and \( b_f \) are the weights and biases of the forget gate, respectively. The augmented memory cell in
LSTM blocks facilitate the model with a dynamic window split flexible enough to adjust and regulate data, enabling the model to predict the outputs with respect to the existing and earlier inputs [105].

This contribution deploys a deep LSTM model for the early identification of students at-risk based on their engagement with the online platform in the form of clickstream data. To attain this purpose, the OULA dataset was processed to retrieve weekly clickstream data for each activity, where each week \( w_i \) consisted of the same students \( s_i \) and feature activities \( a_i \). The computed week-wise engagements formulated the clickstream behavior for each \( a_i \) such that for each \( w_i \) a vector \( v_i \) is formed consisting of the clickstream behavior for each \( s_i \). These week-wise vectors, consisting of the interaction activities, were stacked on each other and passed to the LSTM architecture, such that for every \( s_i \), \( v_i \) was created for a particular time \( t_i \). Before feeding the vectors to the layers, the vectors were padded to transform the sequence into an equal length, since for each additional week the length of the vector would increase. The padded vectors were masked to ignore the padding values, enabling the model to learn and update based on the engagement information of each vector. Through this procedure, a flexible and adaptable window length was implemented, based on the number of weeks facilitating the prediction of performance in the early weeks. The architecture of the LSTM model consisted of multiple layers, with each next layer accepting the output of the above layers as input, adhering the model to learn intricate details and complex representations more comprehensively that ultimately yields better predictions [196]. Such a layered architecture also captured the structure of the data sequence more precisely and efficiently. Other conventional classifiers, as discussed in the previous section, were also deployed for the early prediction by aggregating week-wise interaction patterns of students in order to compare the results with LSTM. Thus, the OULA data was explored in a two-fold manner where firstly, the overall clickstream information for each activity was aggregated, and secondly, the clickstream data was appended in a week-wise manner to convert it into a time-series classification problem to facilitate early prediction during an on-going course. Classifiers were then trained separately on the two types of computed clickstream information.

5.3.3 Evaluation of Feature Significance for Performance Prediction

Machine learning methods, with their added benefit of generating predictions, face a challenge of interpretability. One of the characteristics of interpretability is finding key features that significantly
contribute to a prediction task. Due to the complex nature of this challenge, a model agnostic representation for feature importance was employed for this contribution which uses Shapley values [197, 198] to analyze the significance and contribution of each feature in the output prediction. Shapely Additive Explanation (SHAP) unifies previous techniques for model interpretations; LIME, DEEP LIFT, Tree Interpreter, Shapely Sampling Values, and Shapely regression values to present a model agnostic method that explains the predictions for all machine learning models [199]. SHAP [200] assigns a shapely value to each feature based on its impact on the model. The impact of each feature is calculated through equation 5.7.

\[
\Phi_i = \sum_{F \subseteq S \setminus i} \frac{|S|! |F|! |S| - |F| - 1)!}{|F|!} [f_{S \cup i} (x_{S \cup i}) - f_S (x_S)]
\]

(Equation 5.7)

Where \( \Phi \) is the feature importance computed for a feature \( i \) and \( S \) is a subset of features \( F \). A model \( f_{S \cup i} \) is trained with the feature \( i \) and another model \( f_S \) is trained for all the possible combinations without that feature \( i \). The output of the two models is compared on the input \( f_{S \cup i} (x_{S \cup i}) - f_S (x_S) \), where \( x_S \) denotes the input values of the features. These shapely values thus assist in assessing the significance of that particular feature on the model prediction and are the weighted sum averaged over all possible combinations. The feature significance is defined as the change in the model output observed with that particular feature versus the feature withheld.

In practice, many terms contribute in evaluating the sum for each feature. Therefore, the contribution of each feature can be approximated through nested sampling [201, 202] where shapely values are estimated first, and sampling is done again to compute \( f_S \) and \( f_{S \cup i} \). This nested sampling process requires multiple steps that can be reduced by taking k-medians of the entire dataset, enabling the dataset to be divided into \( k \) summary portions and performing a weighted evaluation for these summaries only, depicting an overall approximation of the dataset. In our study, various \( k \) values were selected, ranging from 10 to 30, and depending on the time it took for evaluating shapely values, a \( k \) of 30 was selected. Therefore, our dataset was reduced to 30 median summary portions as an approximation of the entire dataset, and then a weighted evaluation was performed for only these 30 summary portions [202]. Shapely values are the only such method that obeys the desirable properties of local accuracy and consistency, having their connections with game theory and are the newly de-
vised model agnostic methods [200]. The concept of local accuracy works on the principle that the sum of the feature attributions will be equal to the output of the function, whereas consistency states that the attribution assigned to a feature will not change irrespective of the model dynamics [203].

5.3.4 Evaluation

In this section, we explain the output metrics used to evaluate the classifiers along with the training and validation measures.

**Outcome Metrics:** Accuracy, Precision, and Recall are common metrics used to evaluate the performance of a classifier. For an imbalanced dataset, accuracy is not a good measure to signify the performance of a classifier; instead recall and precision are used to evaluate such imbalanced classification. Accuracy refers to the proportion of the correct predictions by the classifier. Precision is defined as the ratio of correctly classifying the students at-risk of failure from the overall students predicted as on-risk by the classifier, whereas recall measures the effectiveness of the classifier incorrectly predicting at-risk students from the total students that were actually on-risk.

Similarly, AUC (area under the curve) is also an important metric for classifiers evaluation. It is a trade-off between recall and the false-positive rates and gives the probability of ranking a randomly chosen positive instance above a random negative instance. The values of AUC range from 0.5 to 1, with 1 indicating a perfect classifier that discriminates positive instances from negative ones [204].

**Training and Validation:** The balanced data was split into a ratio of 75:25 with the test set consisting of new course sessions that were not included in the training set. This was done in order to avoid overfitting issues and evaluate the classifier in a more rigorous manner since a random split is observed to produce biased results and inhibit robust predictions [118, 205]. The training set included 4,937 students belonging to each class (i.e., pass and fail). The equal proportion of instances from both classes enabled the classifier to make predictions without bias. Similarly, the test set consisted of 2,117 instances from both classes and included unseen data from new course sessions that had not been included in the training set.
5.4 Results for Students’ Early Performance Prediction

This section reports the results observed in the study for the early prediction of students academic performance, using the OULA dataset that was transformed into a sequential format to support early interventions.

5.4.1 Academic Performance Prediction at the End of a Course

Moreover, in order to observe the behavioral patterns of students with respect to their performances, dataset distribution on the basis of each activity were visualized, as depicted in Fig. 5.2 where each of the activities are plotted in a kernel density form with y-axis representing the activity density and x-axis insinuating the activity values. It represents the univariate distribution of each activity on the basis of the target class (pass/fail), with blue (lighter color) representing ‘pass’ and pink (darker color) illustrating ‘fail’ instances. The density distribution for some of the features appears to be greater than 1, however considering the x-axis and calculating the total area under the curve the density approximately reaches 1. A skewed distribution of the data can be observed with a massive amount of data skewed towards the initial values. This illustrates that a large number of non-significant clicks were observed for some students. The dataset is sparse with a large number of zeroes, representing no clicks for that particular activity. Therefore the data for almost all of the activities is skewed towards the left with some instances having a large number of clicks, as can be observed in Fig. 5.2.

Moreover, correlations of the activities on the basis of student performances were evaluated and Fig. 5.2, in a descending order, depicts highly correlated features and their distribution. So the first row of this figure illustrates the top 5 highly correlated activities, which coincidently also appeared to be the activities with the highest number of clicks. Highest correlated activities were found on the basis of Spearman correlation coefficient. HomePage, Forumng, Subpage, Oucontent, Resource and Quiz were found to be the top six highly correlated activities. Excluding resource activity, the other five activities had the highest number of clicks amongst others. A pairplot with respect to their performances was constructed in matplotlib, to represent the overall data distribution and analyzing it with respect to performances of the students. Such exploratory analysis assists in visualizing the data and interpreting patterns that exists between the data. We can conclusively interpret that the
Figure 5.2: Dataset Distribution for Each Feature Activity with Dark as 'Pass' and Light as 'Fail' Class
behavioral patterns for both pass and fail instances are inter-mingled and not easily distinguishable. This blended nature of the two classes makes the student performance prediction more challengeable and the key question is identifying significant characteristics impacting performance that make an instance inclined towards a certain class.

To predict student performance at the end of the course, the conventional ML algorithms were deployed to the processed aggregated dataset. Aggregated clickstream data for each student, comprising of their overall interaction with VLE activities, were included in these experiments. Table 5.2 provides the results of the models. The results revealed that the linear SVM performed best amongst the SVM kernels, although the data were not linearly separable because the interaction patterns of the pass and fail do not differ with a prominent margin. One of the reasons that non-linear kernels performed poorly is due to the fact that the data was quite sparse, and since SVM maps the features on a high dimensional space, therefore non-linear kernels perform poorly. KNN also performed poorly in predicting the performances with an AUC of 0.69, which indicates that it did not distinguish failed students from the ones that passed the course. GBT and decision trees were second best in predicting the performance of students. Further, GBT with learning rates ranging from 0.1 to 0.0021 and 0.0001 were executed, from which 0.001 learning rate produced optimal results, as provided in Table 5.2. GBT tends to overfit the data, and hence a low learning rate produced better results. DT and GBT gave comparable performances on the dataset; GBT tends to reduce the model bias in contrast to DT that reduces the variance. Tree-based models tend to perform fairly well because they split the data to make possible combinations and decide the best split based on the purity of the leaf nodes.

**Table 5.2:** Classification Results on the Balanced Dataset for Pass/Fail at the End of a Course Duration

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM - RBF</td>
<td>68.43</td>
<td>72.52</td>
<td>68.52</td>
<td>0.67</td>
</tr>
<tr>
<td>SVM - sigmoid</td>
<td>66.04</td>
<td>72.51</td>
<td>66.26</td>
<td>0.65</td>
</tr>
<tr>
<td>SVM-linear</td>
<td>73.96</td>
<td>75.54</td>
<td>66.73</td>
<td>0.64</td>
</tr>
<tr>
<td>LR</td>
<td>73.67</td>
<td>75.63</td>
<td>72.23</td>
<td>0.68</td>
</tr>
<tr>
<td>DT</td>
<td>79.82</td>
<td>78.30</td>
<td>76.20</td>
<td>0.73</td>
</tr>
<tr>
<td>GBT</td>
<td>78.70</td>
<td>79.18</td>
<td>75.52</td>
<td>0.72</td>
</tr>
<tr>
<td>KNN</td>
<td>78.35</td>
<td>78.27</td>
<td>68.12</td>
<td>0.69</td>
</tr>
<tr>
<td>ANN</td>
<td>83.36</td>
<td>80.39</td>
<td>78.51</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Therefore, these tree-based models performed well in distinguishing the behavior of students. ANN performed the best in the prediction of at-risk students. To tune ANN architecture, various experiments were performed with different optimizers and their default values. Further, the default learning rate values were then changed between the range of 0.1 to 0.0001, in various batches ranging from 64 to 264, activations relu, tanh, and sigmoid, with a combination of three hidden layers of 100, 200, 50 and 100, 200, 100. A learning rate of 0.01 with relu in the hidden layers (100, 200, 50 units) and sigmoid in the dense layer on a batch size of 64 produced optimal results. The ANN also produced optimal results in comparison to other ML classifiers, suggesting the significance of the hierarchical deep learning approach.

5.4.2 Early Performance Prediction in a Quarterly Setting

For each quarter, the four categories corresponding to at-risk students, distinction students and withdrawals were predicted through the deep ANN classifier. Each quarter constituted of the features presented in Table 5.1. The detailed analysis for each of these quarters is discussed below and provided in the Table 5.3.

Table 5.3: Quarterly Accuracy and Loss for each Category

<table>
<thead>
<tr>
<th>Quartiles</th>
<th>Categories</th>
<th>Accuracy %</th>
<th>Loss %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Pass-Fail</td>
<td>77.22</td>
<td>0.485</td>
</tr>
<tr>
<td></td>
<td>Distinction-Pass</td>
<td>80.25</td>
<td>0.479</td>
</tr>
<tr>
<td></td>
<td>Distinction-Fail</td>
<td>80.63</td>
<td>0.360</td>
</tr>
<tr>
<td></td>
<td>Withdrawn-Pass</td>
<td>78.68</td>
<td>0.444</td>
</tr>
<tr>
<td>Q1-Q2</td>
<td>Pass-Fail</td>
<td>81.63</td>
<td>0.413</td>
</tr>
<tr>
<td></td>
<td>Distinction-Pass</td>
<td>80.48</td>
<td>0.469</td>
</tr>
<tr>
<td></td>
<td>Distinction-Fail</td>
<td>81.58</td>
<td>0.392</td>
</tr>
<tr>
<td></td>
<td>Withdrawn-Pass</td>
<td>86.03</td>
<td>0.349</td>
</tr>
<tr>
<td>Q1-Q3</td>
<td>Pass-Fail</td>
<td>86.15</td>
<td>0.345</td>
</tr>
<tr>
<td></td>
<td>Distinction-Pass</td>
<td>80.44</td>
<td>0.484</td>
</tr>
<tr>
<td></td>
<td>Distinction-Fail</td>
<td>85.81</td>
<td>0.301</td>
</tr>
<tr>
<td></td>
<td>Withdrawn-Pass</td>
<td>90.42</td>
<td>0.275</td>
</tr>
<tr>
<td>Q1-Q4</td>
<td>Pass-Fail</td>
<td>88.62</td>
<td>0.318</td>
</tr>
<tr>
<td></td>
<td>Distinction-Pass</td>
<td>81.01</td>
<td>0.468</td>
</tr>
<tr>
<td></td>
<td>Distinction-Fail</td>
<td>85.81</td>
<td>0.308</td>
</tr>
<tr>
<td></td>
<td>Withdrawn-Pass</td>
<td>93.23</td>
<td>0.223</td>
</tr>
</tbody>
</table>
5.4.2.1 Early Prediction of On-risk Students

The problem of predicting at-risk students was converted to a binary classification problem by defining two classes ‘pass’ and ‘fail’. For each quarter, at-risk students were predicted, in order to analyse the improvement in our model, insinuating early prediction of students’ at risk of failure. Fig 5.3 illustrates the accuracy for each of the quarters. It can be observed that the best accuracy is achieved in the last quarter that cumulatively integrates the overall clicks for each activity. However, 2nd and 3rd quarters also do not perform poorly; they too reach an accuracy of above 80%. Therefore, our classifier gives an accuracy of 81-86% for early prediction of at-risk students.

![Figure 5.3: Quarterly Accuracy for each Category](image)

5.4.2.2 Early Prediction of Students with Distinction

To predict the students outperforming others with distinction, the problem was again converted to a binary classification problem by defining two sets of this problem; a) ‘distinction-pass’ b) ‘distinction-
fail’. We show that for the category ‘distinction-pass’, a major change or rise in the accuracy for all the quarters is not observed (see Fig. 5.3). The accuracy throughout the quarters rather remains stagnant, with a little difference in the last quarter (Q1-4), implying that a distinct decision boundary in the case of ‘distinction’ and ‘pass’ is not available. This may be attributed to the class imbalance problem, because the ‘distinction’ instances are quite scarce as compared to ‘pass’ instances. If ‘distinction’ instances are increased then a distinct pattern between such students may be visualized.

Similarly, to evaluate the ‘distinction-fail’ instances, these two were defined as class labels. Deep ANN was implemented on this set of instances, for all the quarters. Due to the class imbalance problem, a distinguishing pattern is not observed, as depicted in Fig. 5.3. It can be visualized that the last two quarters have a stagnant accuracy, however better than the previous quarters, nevertheless due to the low number of ‘distinction’ instances, an unusual accuracy rise between quarters, cannot be observed for this category.

### 5.4.2.3 Early Prediction of Withdrawal Cases

For early intervention in predicting withdrawals, the problem was considered a binary classification by deploying ‘withdrawals’ and ‘pass’ as two defined class labels. A distinct rise in accuracy, with respect to each of the quarters, is observed, with the last quarter reaching an overall accuracy of 93%. Fig. 5.3 demarcates a discrete pattern in the prediction of ‘withdrawal’ instances from ‘pass’ instances. It can be inferred that the model constantly learns the patterns of withdrawals for each new quarter, with accuracy ranging from 78% to 93%.

Fig. 5.4 depicts the quarterly learning accuracy curves for all the four categories specified in this study. We show that for the three categories; ‘withdrawn-pass’, ‘pass-fail’ and ‘distinction-fail’, a clear demarcation can be observed with respect to the quarterly data. Learning accuracy improves with the increased clickstream quarterly data, implying the possibility of early prediction of students’ at-risk of failure, detecting early withdrawals and distinguishing students with distinction from failure instances. Owing to the class imbalance problem, the ‘distinction-pass’ category does not yield substantial results and a clear demarcation line, in terms of improved accuracy between quartiles, is thus missing for this case.
Figure 5.4: Quartiles Learning Accuracy Curves for all Categories
Table 5.4: ANN’s Evaluation with the Baseline Methods for Q1-Q4

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Categories</th>
<th>Accuracy%</th>
<th>Loss%</th>
<th>Precision%</th>
<th>Recall%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>Pass-Fail</td>
<td>88.62</td>
<td>0.318</td>
<td>0.93</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Distinction-Pass</td>
<td>81.01</td>
<td>0.468</td>
<td>0.37</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Distinction-Fail</td>
<td>85.81</td>
<td>0.308</td>
<td>0.74</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>Withdrawn-Pass</td>
<td>93.23</td>
<td>0.223</td>
<td>0.96</td>
<td>0.86</td>
</tr>
<tr>
<td>SVM</td>
<td>Pass-Fail</td>
<td>85.65</td>
<td>0.157</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Distinction-Pass</td>
<td>79.95</td>
<td>0.202</td>
<td>0.64</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Distinction-Fail</td>
<td>78.08</td>
<td>0.219</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Withdrawn-Pass</td>
<td>89.14</td>
<td>0.108</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>LR</td>
<td>Pass-Fail</td>
<td>84.23</td>
<td>0.157</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Distinction-Pass</td>
<td>79.82</td>
<td>0.202</td>
<td>0.75</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Distinction-Fail</td>
<td>81.69</td>
<td>0.183</td>
<td>0.81</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>Withdrawn-Pass</td>
<td>85.60</td>
<td>0.144</td>
<td>0.86</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Fig. 5.5 illustrates the quarterly learning loss curves for all the four categories mentioned. Similar to

Fig. 5.6: ANN’s evaluation with the baseline methods for Q1-Q4
Figure 5.5: Quarters Learning Loss Curves for all Categories
the learning accuracy, the loss too depicts a clear demarcation for each quartile. As accuracy increases the corresponding loss for each quartile decreases. Similar to the accuracy of the ‘withdrawn-pass’ category, the learning loss also does not yield a distinct decrease. In order to evaluate the quarterly results, the last quartile encompassing data from Q1-4 was compared with SVM and ANN, details provided in Table 5.4. As illustrated in Fig. 5.6, the proposed model yields better accuracy compared to baseline models.

This section presents a contribution to knowledge in early prediction of students’ at-risk of failure, determining students likely to withdraw from modules and ascertaining significant features that enable a student to outperform others. Results reveal demographic characteristics and student’s clickstream activity, after the module initiation, as having a significant impact on student performance. This study also determines the effectiveness of the deep learning model in the early prediction of student performance, enabling timely intervention by the university to implement corrective strategies for students support and counseling. Due to the class imbalance problem in ‘distinction’ instances, a discrete pattern for such students was not observed, a limitation of our study. However, demographic and geographic characteristics tend to significantly impact performance. The performance evaluation model shows a sensitivity of 69%, a precision of 93% and overall accuracy of 88% in predicting at-risk students; a sensitivity of 86%, a precision of 96% and overall accuracy 93% in predicting early withdrawals. Similarly, ascertaining ‘distinction’ students from ‘fails’, a sensitivity of 74%, and precision of 81% and overall accuracy of 85% is achieved.

### 5.4.3 Early Performance Prediction in a Week-wise Setting

In order to analyze early performance prediction, the activities pertaining to students’ engagement, as provided in Table 5.1, were computed in a week-wise manner. According to the duration of the course, which consisted of a maximum of 264 days (as provided in the dataset), each student’s activities for 38 weeks were formulated. Due to the increased number of weeks, experimentation was performed on some selective weeks in order to analyzes the gradual performance improvement in the prediction process. Weeks 5th, 10th, 20th, 30th, and 38th were selected to observe students’ performance and classifiers prediction. The results for early prediction were compared with the time-series classifier LSTM and other classifiers, as discussed in Section 5.1.
Table 5.5: Comparison of ML Classifiers and LSTM Week-wise Results

<table>
<thead>
<tr>
<th>Week</th>
<th>Classifiers</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-5</td>
<td>SVM–RBF</td>
<td>58.27</td>
<td>65.23</td>
<td>53.63</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>SVM-sigmoid</td>
<td>55.86</td>
<td>65.09</td>
<td>56.74</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>SVM-linear</td>
<td>62.73</td>
<td>65.32</td>
<td>58.70</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>64.04</td>
<td>66.20</td>
<td>65.34</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>66.10</td>
<td>61.10</td>
<td>60.12</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>GBT</td>
<td>63.80</td>
<td>63.39</td>
<td>59.01</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>63.30</td>
<td>60.02</td>
<td>58.21</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>65.85</td>
<td>63.01</td>
<td>62.12</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>68.92</td>
<td>70.23</td>
<td>69.74</td>
<td>0.71</td>
</tr>
<tr>
<td>1-10</td>
<td>SVM-RBF</td>
<td>60.20</td>
<td>62.01</td>
<td>55.32</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>SVM-sigmoid</td>
<td>58.08</td>
<td>59.50</td>
<td>57.36</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>SVM-linear</td>
<td>66.30</td>
<td>62.50</td>
<td>59.09</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>67.29</td>
<td>68.52</td>
<td>67.45</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>67.60</td>
<td>68.00</td>
<td>68.04</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>GBT</td>
<td>67.24</td>
<td>69.53</td>
<td>69.51</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>66.61</td>
<td>66.52</td>
<td>66.53</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>70.01</td>
<td>70.23</td>
<td>69.32</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>72.65</td>
<td>72.71</td>
<td>70.92</td>
<td>0.74</td>
</tr>
<tr>
<td>1-20</td>
<td>SVM-RBF</td>
<td>64.15</td>
<td>70.23</td>
<td>60.02</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>SVM-sigmoid</td>
<td>61.79</td>
<td>70.50</td>
<td>63.50</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>SVM-linear</td>
<td>69.94</td>
<td>72.32</td>
<td>60.32</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>70.32</td>
<td>71.52</td>
<td>68.05</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>71.08</td>
<td>70.04</td>
<td>69.53</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>GBT</td>
<td>72.44</td>
<td>69.52</td>
<td>68.54</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>71.12</td>
<td>66.53</td>
<td>62.56</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>73.68</td>
<td>72.65</td>
<td>70.27</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>75.65</td>
<td>74.57</td>
<td>72.54</td>
<td>0.78</td>
</tr>
<tr>
<td>1-30</td>
<td>SVM-RBF</td>
<td>67.10</td>
<td>72.22</td>
<td>65.43</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>SVM-sigmoid</td>
<td>64.46</td>
<td>71.52</td>
<td>67.50</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>SVM-linear</td>
<td>72.37</td>
<td>72.26</td>
<td>64.51</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>72.06</td>
<td>73.50</td>
<td>70.40</td>
<td>0.67</td>
</tr>
</tbody>
</table>

(To be continued)
The results obtained from each of the deployed classifiers are provided in Table 5.5. The LSTM architecture consisting of four-layers was observed to produce the best results for all the mentioned weeks, with each LSTM layer, including a different range of units, from 68 to 264 units. Extensive experiments were performed to tune the classifier and improve the prediction process by using a combination of different optimizers and activation functions with a varying range of learning rates and batch sizes.

Deep learning models are prone to overfitting and require different architectures to reduce overfitting. This can be attained by introducing dropout and regularization between layers of the classifier, that randomly drop nodes in each layer, to avoid inter-dependency and ultimately yielding in the reduction of overfitting [206]. For each week’s training, the computed week-wise information $w_i$ until the $i^{th}$ week was passed to the layered architecture of the LSTM network. For the hyperparameter tuning, different optimizers (adam, adamax, rmsprop, adagrad) were tuned with default values and later with various learning rates ranging from 0.1 to 0.00001, and batch sizes varying from 64 to 1364.

The ADAM optimizer with 0.001 learning rate and ‘relu’ as activation in the dense layer was observed to provide optimal results for all the mentioned weeks but with different units in the architecture and varying batch sizes. For the $5^{th}$, $10^{th}$, $20^{th}$, $30^{th}$, and $38^{th}$ week, batch sizes of 864, 964, 1064, 1264,
Figure 5.7: LSTM Validation Accuracy, Loss, Precision and Recall Curves for 5\textsuperscript{th}, 10\textsuperscript{th}, 20\textsuperscript{th}, 30\textsuperscript{th} & 38\textsuperscript{th} weeks at 50\textsuperscript{th} epoch
and 1364, respectively, were found to get optimal results. As batch size increased from the value of 64, the performance metrics for all the weeks improved, however, when the batch size was increased further from 1364, the performance metrics degraded with AUC decreasing by a value of 0.04. An increased batch size enables the model to learn different pattern behavior of students and to lead to good predictions. The architecture for each week ranged from four to five layers with each layer consisting of 50 to 200 units, and a dropout of 0.2 applied on each layer, and a recurrent dropout of 0.3 applied on the first layer only.

To find the optimal loss function, several built-in and customized Keras loss functions were applied, including binary cross-entropy, hinge loss, stochastic gradient descent, square-hinge loss, and focal loss. Focal loss [207] deals with the class imbalance by hard mining the difficult problems; however, it performed poorly on the original imbalanced dataset, due to data sparsity. After several experiments, binary cross-entropy, which inherently emphasizes on the binary nature of the problem, provided optimal predictions with loss values degrading with each additional week. Since LSTM requires high computational resources, therefore, these experiments were performed on a 24GB GPU machine with 124GB RAM, implemented in Python using different sklearn, Keras, and matplotlib libraries.

Fig. 5.7 illustrates the results of the deployed time-series classification model (LSTM) by depicting the weekly trend in the performance measures. Gradual progress is observed in the week-wise accuracy curves, with 68.92% observed in the 5th week to 84.57% obtained in the last week (refer to Fig. 5.7 (up-per left)). Each ith week consisted of a vector appending the engagement pattern of students, for each of the activities, till that ith week. In the starting weeks of course offerings, conclusive engagement patterns of students could not be observed because of the varying nature of courses where students could just join-in the course for auditing and could withdraw without any fee. Also, in the first few weeks, the past interaction history of a student is not sufficient enough. Therefore, the engagement behavior is analyzed from as early as the 5th week. Fig. 5.7 depicts the effectiveness of the LSTM model in predicting the performance of students through the gradual refinement in the accuracy scale range with additional weeks.
Figure 5.8: Using SHAP to Identify Important Features impacting Performances of Students
Figure 5.8: Using SHAP to Identify Important Features Impacting Performances of Students
Furthermore, Fig. 5.7 (upper right) presents the degradation in the loss values with additional week-wise interactions. The loss measure demonstrates the difference between the actual performance and the predicted output. The degrading loss curves illustrate the improvement in the prediction throughout the course presentation. The performance measures shown in the figure are the validation curves, with 75% of the dataset included in training and 25% included for validation, achieved at the 50th epoch of the model. Moreover, precision and recall curves are also provided in Fig. 5.7 (bottom left and right). The precision and recall curves illustrate the improvement in the model with each additional week, signifying the learning behavior of the model. With additional weekly information, the model improved the recall and precision measures. A notable improvement was observed in the precision values after the 10th week. Similarly, for recall, a notable increase can be observed after the 20th week. The values of precision ranged from 70.23% in the 5th week to 82.25% in the last week, whereas recall ranged from 69.74% in the 5th week to 79.43% in the last week.

A comparison between the conventional classifiers and LSTM is provided in Table 5.5. In the initial weeks, LSTM outperformed other classifiers with a margin. With the accumulated data in additional weeks, other classifiers also tended to produce good predictions. However, for early predicting the students in the initial weeks of the course offerings, LSTM produced optimal results. This pattern of improvement highlights the robust predictions by the model. This time-series classifier was fed with week-wise interactions of each student and thus could predict student performance on a more granular level. Other classifiers could be tailored to accept such vectors, and thus through the aggregated data were unable to perform optimally in the initial weeks.

### 5.4.4 Identifying Features impacting Students’ Academic Performance

The Shapely method of identifying important features was deployed on the OULA dataset consisting of the overall click-stream data of students throughout the course modules. SHAP requires a kernel on the basis of which the model is trained. From the extensive experiments conducted on our data, detail provided in the previous section, LSTM was found to produce the best results in the classification of students on the basis of their clickstream data. Therefore, the same LSTM settings (for 38th week) were used for SHAP (refer to Section IV.A), and the results are visualized using summary plots, which list the features according to their significance (refer to Fig. 5.8) where darker color red denotes higher
activity value, and a lighter color blue denotes lower activity values.

The SHAP values for each activity are along the x-axis ranging from -1 to 1. A negative SHAP value increases the risk of failure, and a positive SHAP value decreases the risk of being a failure. The results reveal the influence of individual features, as illustrated in Fig. 5.8. Levels of participation in quizzes, student interaction with the homepages of their courses, and assessment submission were found to be significant predictors. The analysis showed that the Quiz activity was the most significant feature where high values in this feature were associated with a low risk of failing a course and where low values were associated with a high risk of failure.

A similar pattern of a positive association between the student risk level and the use of the Homepages of the course was found. However, active student participation in the discussion forums (Forumng) does not necessarily represent a positive impact on their performance. A mixed pattern was observed for some features, where both high and low values either cumulatively increased or decreased the risk of failure. For example, for the OuContent feature, which deals with the students’ assignments and assessments, an amalgamated pattern is observed where students who submitted their assessments or appeared to be active in their assessments still face the challenge of being at risk of failure. Similarly, being an active member in discussion forms (Forumng) did not decrease the risk of failure. The top features shown in Fig. 5.8, such as Quiz, Homepage, OuContent were observed as more impactful as compared to the bottom ones, which showed a negligible impact on the model such as RepeatActivity and SharedSubpage. These bottom features, Fig. 5.8, also tend to show minimum to none activity in terms of the number of clicks by students. From such results, it can be inferred that activities demonstrating active student participation tended to have an impact on the academic performance of students as compared to activities with negligible to none student participation.

The results are shown in Fig. 5.8b, 5.8c and 5.8d are dependence plots and visualize how the model depends on and behaves with respect to a certain activity. Dependence plots for three activities (from Fig. 5.8) are provided, two from the top: Quiz and OuContent (refer to Fig. 5.8b and 5.8c) and the third one from the middle (refer to Fig. 5.8d): Forumng. This selection of activities caters to the variation of the features and their influence on the predictions. In the dependence plots, activity values are provided on the x-axis and SHAP values on the y-axis. These plots depict a discrete pattern of behavior where OuContent seems to reflect a more discerning pattern as compared to Forumng. With higher activity in the OuContent, SHAP values also increase, as represented by darker color
in the figure and demonstrating a positive impact on the model. However, for a discrete pattern cannot be observed. For increased activities in this feature, the model shows both a positive and a negative impact, suggesting that predictable behaviors cannot be inferred from this feature. Assignments and assessments (OuContent) activities appear to be more impactful where students with higher values appear to have a lower risk of failure. The significance of quizzes (refer to Fig. 5.8b) can be illustrated by the fact that as the number of clicks on a quiz increases, so does its shapely values, suggesting a positive impact on the model.

This contribution examined the successful prediction of students through an application of conventional classifiers by monitoring the engagement of students with a learning platform used by learners enrolled in fully online courses that were offered in self-paced learning mode. The first research objective of the study was to provide a comprehensive comparison of deep learning algorithms with conventional machine learning algorithms to predict students at risk to fail a course based on log data about the students' interaction with the learning platform. The study results indicate that a deep learning algorithm outperformed the conventional and commonly used machine learning options. The deep LSTM technique outperformed all other alternatives and achieved an accuracy of 84.57%, precision of 82.24%, recall of 79.43%, and AUC of 82%. This demonstrates a strong potential of deep algorithms to provide a foundation for higher education institutions to address challenges associated with self-paced education in fully online learning settings.

The chapter presented the effectiveness of the deep LSTM, outperforming the conventional machine learning algorithms in the early prediction of students at risk. In fact, the accuracy of the LSTM algorithm measured with AUC with the data accumulated after the first 5 (0.71) and 10 (0.74) weeks of the course outperformed almost all other conventional algorithms even when they are used on the complete dataset collected for the entire duration of the course. This comes from the property of LSTM to use computed log data on a weekly basis to convert performance prediction into a time-series classification problem for early prediction. In practical terms, a relatively high level of prediction accuracy, only after five weeks of course duration, offers promising opportunities for early interventions in self-paced education where courses are offered for an extended period (6 or 9 months).
5.5 Summary

This chapter presented the methodology, experimental set-up and the corresponding results for the early prediction of academic performances of students in a quarterly and week-wise setting. For the prediction of students at the end of the course, ANNs and other machine learning approaches were deployed. This prediction was then further processed on a quarterly and week-wise granularity to early predict the academic performances for timely interventions and for formulating early alarm systems. The deployed Deep LSTM model outperforms other conventional machine learning algorithms with a significant margin in the week-wise performances prediction.
Chapter 6

Balancing the Award-Gap Imbalance in the Academic Performances of Students

“We live on an island surrounded by a sea of ignorance. As our island of knowledge grows, so does the shore of our ignorance.”
- John Archibald Wheeler

In the Learning Analytics community, considerable significance can be observed on the early prediction of students’ performances which inherently constitute it as a time-series problem. However, evidences on the up-sampling of such sequential data cannot be observed in the learning analytics community. This study intends to eliminate the award-gap class imbalance and up-sample the openly accessible data by transforming it into a sequential quarterly format and analysing the improvement in the prediction of the performances in the transformed dataset using classical machine learning algorithms.

6.1 Up-Sampling Time-Series Data

Conventionally, in a classroom setting the rate of students passing a course will surpass those that are inclined towards failure, thereby creating an imbalance in the performance ratios and thus producing biased results in terms of the prediction of this performance. In order to realistically predict the performances of students a mechanism is required to balance the ratio of the students’ performances. Several techniques have been introduced in the literature to synthetically generate data for up-sampling the minority class and thus balancing the ratios. Generative Adversarial Networks (GANs) have been rapidly adopted to eliminate the class imbalance by learning the probability distribution and generating synthetic data [123]. The Learning Analytics research community lacks an implementation of GANs to overcome the class imbalance with respect to the students’ performances. Moreover, for an early prediction of a student performance the challenge is to sequentially up-sample the minority class students such that the synthetic instances generated have an association in each sequence. A
detailed illustration is depicted in Fig 6.1 where the data is presented in a quarterly manner and ‘pass’ instances in green \( P(a_1) \) exceed the ‘fail’ instances in red , \( F(a_1) \), thus representing a class imbalance in terms of the resultant performance (pass/fail). Each student’s log information is represented homogeneously in a quarterly manner, where ‘1’ in \( a_1 \) represents the quarter and ‘a’ represents a specific student. We aim to devise a mechanism such that for each synthetic generated student \( F_a \) its behavior is encapsulated in the next quarters, represented by dotted lines in Fig 6.1. Since in this study ‘fail’ is the minority class hence the figure only shows the synthetic instances of the minority class representing the up-sampling of those instances.

We aim to address the following research objectives for this contribution:

- Firstly, we aim to leverage GANs to generate synthetic temporal sequences and up-sample the minority class. The effectiveness of the proposed temporal GANs is determined by deploying the up-sampled temporal data on the conventional machine learning algorithms widely applicable in the Learning Analytics community and comparing the results with previous versions of GANs and other up-sampling techniques. In contrast to the baseline conventional up-sampling approaches, our proposed models outperform by 3.57% and 5.33% validation accuracy.

- Secondly, we intend to extend the existing conditional GANs for temporal data such that each new generated temporal sequence is formed by learning the data distribution of its previous temporal sequence. We propose two extensions of C-GANs: Educational Discrete (EDD)-GAN and Educational Continuous (EDC)-GAN. Both of the proposed models implement a different variation of the condition, with the inclusion of courses information as a condition in EDD-GAN and EDC-GAN’s condition comprises of an additional element; that is the student engagements.

- Lastly, the openly accessible Open University Learning Analytics dataset is utilized as a case study to leverage the effectiveness of our proposed GAN-based models. The dataset has a class imbalance with majority of the students passing a course as compared to those who tend to fail. The proposed models are deployed on the transformed sequential sequences to up-sample the ‘fail’ sequences.

We propose a novel method based on adversarial networks to eliminate the class imbalance in the openly accessible OULA dataset. The data is first transformed into a temporal quarterly format and
Actual students passed

Up-sampled failed students

Figure 6.1: Up-Sampling Time-Series Data
then fed to the adversarial networks. Each student’s interaction with the VLE is thus presented in quarters. In order to up-sample this temporal data, the generated sequences should be associated with each of their previous quarter, encapsulating the overall behavior for each student. The proposed temporal educational GANs implement this association by the inclusion of the required attributes in the condition. For each new quarter, the previously generated sequence for that $i^{th}$ student is included in the condition and thus the network learns the behavior of that student from its previous quarters. The proposed networks are evaluated in comparison to some conventional non-GAN based and GAN-based up-sampling approaches. A detailed explanation of the proposed approaches with their evaluation is presented below.

### 6.2 Dataset for Balancing Students Logs

Firstly, this section provides the research flow of this study, as presented in Fig. 6.2. Data processing steps are followed to transform the student interactions into quarters. Each course duration is divided into four quarters and thus each quarter consists of the engagement activities of some specific days [112]. The quarterly information is decomposed to train and test set, where training set consists of both actual and generated data and test set constitutes of the actual data only. From the actual data, quarterly interactions with some random noise are fed as input to the generator $G$. Neural network dense layers are used in the generator $G$ and discriminator $D$. $G$ generates rows following the specifications as mentioned in the next sub-sections. The discriminator $D$ intends to differentiate the generated rows from the real ones. Once the rows are generated through the proposed method the effectiveness of the up-sampled data is analysed through its machine learning efficacy. The ‘failed’ instances from the generated data are sampled to form the train set for the classifiers, eliminating the class imbalance between pass and fail performances, whereas the test set constitutes of a subset consisting of original ‘pass’ and ‘fail’ instances. The results for the prediction models are compared to evaluate the efficacy of the classifiers on the up-sampled balanced dataset.

Further, the sections provide the details of the procured OULA dataset and its transformation into temporal format. Moreover, an in-depth analysis is presented on the proposed methodology based on adversarial networks and each component of the network is described in the following subsections.

The dataset for this study is acquired from OULA which is freely available and accessible online, and
Figure 6.2: Data Pipeline for Temporal GAN-based Up-Sampling Methods.
encompasses log data of 32,593 students over a course duration of 9 months, from 2014-2015 [176].

The log data constitutes of their demographics, clickstream behavior including students’ engagement with the VLE and assessment related information such as their submissions and scores. The data is comprised of seven courses (modules) with each course being assigned a code and presented through a VLE at different intervals in a year. The duration of each course varies, with the maximum duration ranging from 259-264 days. The combination of course name and its presentation at different intervals throughout the year results in 20 number of courses. Four distinct performance classes are provided in the dataset: distinction, pass, fail and withdrawal. From the provided classes distinction and pass are merged to form one class of 15,385 students against the ‘fail’ class, comprising of 7,052 students. The students’ engagement with the VLE is characterized into 20 different activity types with each activity referring to a specific action, such as downloading lectures, viewing course contents, attempting quizzes, and logging in to the homepage. The names of each of these activity types, as provided in the dataset are as follows: dataplus, dualpane, externalquiz, folder, forumng, glossary, homepage, htmlactivity, oucollaborate, oucontent, ouelluminate, ouwiki, page, questionnaire, quiz, repeatactivity, resource, sharedsubpage, subpage and url; and the description of each feature is available in [54] study. Overall, the OULA dataset, acquired in a raw form consisting of several log files, should be transformed into a structured format in order to understand student behavior and obtain conclusive results. Feature engineering steps were performed to obtain relevant clickstream information for each student according to their engagement with the VLE. The activity related features, depicting student engagement, were formulated in oracle 11g where several inter-related log files were joined to obtain the required clickstream information for each student, depicting their interaction with the learning platform. The student engagements were computed in a quarterly manner such that each course was divided into four quarters. Each quarter consisted of the log data for the mentioned activities.

6.3 ED-GAN for Temporal Synthetic Data Generation

To synthetically generate tabular data, a data synthesizer $G$ is trained on the original table $T$ to learn the distributions of $T$ and generate a synthetic table $T_{syn}$ that encapsulates the behavior of $T$. The table $T$ contains $N_D$ discrete columns $D_1, \ldots, D_{(N_d)}$ and $N_C$ continuous columns $C_1, \ldots, C_{(N_c)}$ where $N_D$ follows a consistent discrete pattern consisting of each students performance (pass/fail), course taken (course name) and session of the course (spring/fall), and $N_C$ consists of the columns depict-
ing the interaction of students with the VLE hence each \(N_C\) column follows a different probability distribution. \(G\) is trained on a partition of \(T\) referred to as \(T_{\text{train}}\) and then \(T_{\text{sync}}\) is constructed by individually sampling rows for each student using \(G\). The performance of the generator \(G\) is evaluated through machine learning classifiers by validating the classifiers on the \(T_{\text{test}}\) partition of the \(T\) table where the train set constitutes of random sampling from \(T_{\text{train}}\) and \(T_{\text{sync}}\).

Though customizing the condition may not seem to be a significant change, however some intrusive modifications are required for implementing this customization. There are certain challenges associated with integrating customized conditions in the architecture of a GAN [208]. For instance, a) devising a representation to feed a vector in the generator \(G\) and preparing an input for \(G\); b) \(G\) should learn the real data conditional distribution such that \(T_{\text{sync}}\) data is conditioned on the original distribution. Conventionally, the input of the synthesizer \(G\) of a GAN constitutes of a vector sampled from a standard multivariate normal distribution (MVN), the output of it is then fed to the discriminator which eventually enables the model to learn the data distribution. However, this method of training does not cater the imbalance in the categorical columns and hence hinders the learning of the model. The GAN will learn only the distribution on which it is trained and thus will be trained incorrectly. If new combinations of categorical columns are introduced during the validation phase, the model will produce biased results. To eliminate this biasedness, the trained data is resampled such that all the distinct categories from the discrete attributes are uniformly sampled in the training phase and the discrete attributes not sampled are recovered in the validation phase.

We assume that \(i^*\) is the value from the \(D_{(k^*)}\) discrete columns and \(C_{(k^*)}\) continuous columns that should be aligned with the generated samples \(s^\land\), where \(C_{(k^*)}\) represent the \(k^{th}\) continuous column in the dataset conditioned on the \(k^{th}\) discrete column represented by \(D_{(k^*)}\) in the ground truth. Now the generator \(G\) can be defined as the conditional distribution of rows such that for each \(k^{th}\) discrete column each row of \(G\) can be defined by the following equation 6.1:

\[
s^\land \sim P_G(\text{row}|D_{(k^*)} = i^*)
\]  

(Equation 6.1)

Since our \(G\) is formulated through the conditional distribution of rows, hence the GAN built upon it is referred to as a conditional-GAN. It is also important for the generator \(G\) to learn the conditional
distribution of the real data such that the generated data is conditioned on the data passed as a condition in the generator by the following equation 6.2. Therefore, the equation shows that the probability of a row occurring in generated data is the same as the probability of the same row occurring in real data.

\[
\sum_{i \in D_k^*} P_G(\text{row} | D_{k^*} = i^*) = P(\text{row} | D_{k^*} = i^*)
\]  
(Equation 6.2)

The C-GAN in this study follows a similar behavior as formulated by Xu [208] consisting of some major components; devising a conditional vector, regulating the loss of the generator and training the model through sampling discrete rows from the dataset. However, extending this work further, the conditional generator also incorporates continuous values to learn the data distribution. Each of these components is described in the following section.

### 6.3.1 Conditional Vector

The conditional vector constitutes of a combination of \(N_D\) discrete and \(N_c\) continuous columns where \(N_D\), comprising of \(D_1, \ldots, D_{(N_D)}\) columns, is represented as one-hot vectors \(d_1, \ldots, d_{(N_d)}\), where \(i\) is the \(i^{th}\) feature of \(D\), and \(k\) is the value in the \(i^{th}\) feature, and \(i^*\) is the \(i^{*th}\) feature of \(d\) and \(k^*\) is the value in the \(i^{*th}\) feature of \(d\), such that the \(i^{th}\) one-hot vector is \(d_i = d_i^{(k)}\) where \(k=1, \ldots, |D_i|\). So the vector condition constituting of the discrete columns can be represented through the following equation 6.3. where discrete columns

\[
\bar{d}_k = \begin{cases} 
1 & \text{if } i = i^* \text{ and } k = d_i^* \\
0 & \text{otherwise}
\end{cases}
\]  
(Equation 6.3)

where the three discrete columns in the dataset are the performance with pass, fail values, class code
with seven different values and module session with two different values. This makes a one-hot encoded vector of size 11. For instance, \( \vec{d}_k \) for final-result “fail”, class code “BBB” and module session “J” would be:

\[
\vec{d}_1 = [0, 1], \vec{d}_2 = [0, 1, 0, 0, 0, 0, 0, 0], \vec{d}_3 = [0, 1]
\]

\[
\vec{d} = [0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1]
\]

Hence, the vector condition for discrete columns can be described as represented in equation 6.4.

\[
\vec{d} = \vec{d}_1 \oplus \ldots \oplus \vec{d}_{N_d}
\]

(Equation 6.4)

where \( \oplus \) represents the concatenation operator. The \( N_c \) continuous columns represent the interaction activities of the students depicting their behavior. The continuous columns are appended with the discrete columns in the vector condition. \( N_c \) can be represented in the following equation 6.5.

\[
\vec{c} = \vec{c}_1 \oplus \ldots \oplus \vec{c}_{N_c}
\]

(Equation 6.5)

The condition vector \( y \), constituting of \( N_d \) and \( N_c \), can thus be represented as in equation 6.6.

\[
\vec{y} = \vec{d} \oplus \vec{c}
\]

(Equation 6.6)

where values for each one-hot encoded discrete column are concatenated with values for the continuous columns and passed as condition to the G. Thus each row represents the engagement of student with their particular attributes constituting of their performance, course taken and session.
6.3.2 Generator Loss

During the training phase if \( G \) is fed with a condition vector of \( D_i = k^* \), the feed forward mechanism in the generator allows the output of the generator to be conditioned on the discrete features. In order for the conditional \( G \) to generate a conditioned output, for each batch, average cross entropy is added between the generated and one-hot vectors to penalize the loss [208]. As the training phase progresses, \( G \) intends to learn the distribution of the original data and generates data conditioned on discrete features.

6.4 Results: Experimental Setup

This section discusses the experimental setup for up-sampling the data through our proposed models and the results are evaluated by up-sampling through conventional up-sampling techniques such as CGAN and others which are taken as baseline in this study. Different variations of CGANs were implemented to up-sample the ‘fail’ students and in this section each up-sample methodology is discussed in detail.

6.4.1 Baseline Methods

For evaluating our proposed methodology conventional CGANs and CTGANs are taken as baseline studies. The following section describes each of the up-sampling techniques, including both GAN-based techniques and other traditional up-sampling approaches.

6.4.1.1 C-GAN

For the first evaluation a basic implementation of C-GAN was implemented where the class labels ‘pass’ and ‘fail’ were passed as condition to the GAN model. The network thus learns the behavior of students on the basis of the passed condition. For this implementation the quarter data for each student was formulated as one time stamp such that each quarter was appended with its next quarter.
Such formulation enabled to capture each student’s interaction as a single record in the table. The condition is signified as y-columns and the input as x-columns. Each quarter consisted of 20 activities with three attributes signifying attributes for each student therefore a matrix of 20x4 with additional 2 features was passed as input to the generator and class labels (pass, fail) were passed as one-hot encoded to the condition vector. The resultant generated data was combined with original data to up-sample the ‘fail’ instances and machine learning classifiers were deployed to get the predicted results.

6.4.1.2 CTGAN

Conditional Tabular GAN models the distribution of a tabular data through a conditional generator, encompassing several characteristics of tabular data such as mixed data types [208]. It introduces techniques to handle Gaussian, non-Gaussian and multi-modal data distribution and the imbalance in the discrete variables. It introduces mode-specific normalization to handle the variables with complex distributions. It presents the training by sampling method to encapsulate maximum behavior of the discrete variables in order to improve the training procedure. The conditional vector in this approach only caters the discrete columns concatenated with each other.

6.4.1.3 Other Known Up-Sampling Approaches

In relation to the GAN-based approaches some other non-GAN based conventional up-sampling approaches were also deployed for evaluations, which are also prevalent in the existing learning analytics studies.

**ROS:** Random Over-Sampling [209] is a naïve method to duplicate the minority class instances in order to balance the data according to the majority class instances. This method does not take in accordance the heuristics of the data or their complex structure. Since in our data the minority class is twice less than the majority class hence the ‘fail’ instances will be doubled. Such an inclined behavior will enforce biasedness and leads the model towards over-fitting.

**SMOTE:** Synthetic Minority Over-Sampling Technique is one of the most widely applicable ap-
proaches applied in this field [210]. It tends to synthetically generate new minority instances based on the k-nearest neighbors of existing minority instances [211]. For each minority instance, its random nearest neighbor is selected and interpolation calculations are computed to create a new minority instance. Although widely applicable, it suffers through a lot of issues including noise amplification and within-class imbalance. Since it does not enforce a decision boundary, therefore new instances are rarely created in sparsely populated areas. Hence, it is unable to create minority instances with some diverse behavior.

**ADASYN:** Adaptive synthetic sampling approach (ADASYN) [212] is based on the implementation of SMOTE and therefore learns a wider distribution of data compared to SMOTE. This approach emphasizes on the selection criterion of the nearest neighbors for each minority instance, such that new synthetic minority instances are created by selecting the neighbors belonging to the majority class. It is based on the concept of density distributions and creates new samples by following the distributions of the data points. ADASYN provides a more balanced distribution of the minority samples by focusing on the hard examples of the minority class that are more difficult to learn.

**K-means SMOTE:** It is a hybrid of SMOTE and k-means clustering, consisting of three major steps [213]. The data is first clustered into groups using k-means clustering algorithm, clusters are filtered based on the number of minority instances existing in them and those with sparsely populated minority instances are assigned more synthetic instances, and lastly each cluster is over-sampled through an application of SMOTE such that between-class imbalance is eliminated. The within class imbalance is compensated in the filtering process where a few synthetic instances are allocated to the densely populated areas [214].

### 6.4.2 Proposed Methods for Up-Sampling Temporal Data

This study proposes two conceptual approaches following the guidelines provided by [208]. This section describes the proposed approaches for handling the class imbalance, in reference to our OULA data after its composition into quarters.
In our first proposed model, Educational Discrete GAN, the condition passed to the generator was improvised such that the performance of students, course taken and session were aligned as condition \( y \). Such a condition will enable the generator to learn a better pattern of students with their precise description and features such that the generator is able to learn the pattern of students in a particular course and session. In this experiment, GANs network were implemented for each new quarter to up-sample instances for that particular quarter. We refer it as the EDD-GAN. A detailed description of the proposed GAN is provided graphically in Figure 6.3, where the red lines depict the condition vector fed to the generator, and the black lines are the feature vectors passed as ground truth to the discriminator.

**Figure 6.3:** Data Formatting for Each Quarter in EDD-GAN: Black lines represent the Ground-truth, red lines depict the Condition fed to the Generator.

For the first quarter, 3 conditional discrete features were fed as condition to the generator and its 20 activity features were used as ground truth in the discriminator. Condition features are one-hot encoded, generating a vector of length 11 for each student and passed as input to the generator along with a vector of random noise. Activity features are included as the input data for the discriminator and fed to the GAN network in the discriminator. The output of the generator is constituted of the
activity features conditioned on the discrete features. So, for the first quarter when generator is fed a
condition of 11 vectors and random noise of 20 dimension, it outputs 20 activity features. The output
of the generator (20) is used as input for the discriminator, along with the ground truth of the same
quarter but with conditional discrete features as class labels. Ground truth constitutes of a vector of
20 continuous and 11 discrete features.

For the second quarter, condition vector is of same format (discrete features passed as one hot vector)
but the ground truth (for the discriminator) is created by concatenating synthesized Q1 data, generated
from trained Q1 network, with original Q2 features. But the issue here is that we cannot just simply
concatenate random rows as condition features play an important role i.e. Q1 features of fail student
should not be concatenated with Q2 features of fail students from a different course. To elevate this
phenomenon, the synthesized data is concatenated carefully with equal percentage of all 11 classes
(one-hot discrete features) and index-wise concatenated with feature vectors of Q2 for matching con-
dition vector. For the discriminator, it will have 20 synthetic continuous features generated for Q1,
20 input continuous features for the second quarter and a feature vector of 11 dimensions (discrete
features) as condition. The generator for Q2 will comprise of 11 one-hot discrete features as condition
Y and noise of 20 dimension as input Z.

Similarly, for the third quarter, condition vector is same for the generator but the ground truth for
the discriminator is composed of concatenated synthesized Q1 data and synthesized Q2 data with
original Q3 features according to same condition vector criteria as explained for Q2. So, it will have
60 input features in the discriminator (that is synthesized Q1, Q2 continuous features, and original
Q3 continuous features) and 11 dimensional one-hot encoded condition vector. For the last quarter,
condition vector consists of discrete attributes, and the ground truth is created by concatenating the
synthesized Q1, Q2 and Q3 data with original Q4 features according to the same condition vector
criteria as explained above. So, it will have a total of 80 input features (in the discriminator) and
11-dimensional condition vector.

6.4.2.2 EDC-GAN

In our second proposed model, Educational Continuous GAN, a continuous feature vector, consisting
of the student interactions, along with the previously selected discrete features (as in EDD) was passed
as a condition. In order to analyse the impact of students’ interactions on their performance and on the performance of the model their interactions and activities were fed to the generator as the condition. These interactions are the continuous columns and were appended with the discrete one-hot encoded features, as discussed in Chapter 4. For each quarter GAN was executed to generate instances for each quarter and the behavior of each student was encapsulated through the customized condition that was modified with respect to each student with respect to their course taken, session and their interactions with the VLE. A detailed description of the proposed GAN is provided graphically in Figure 6.4, where the red lines depict the condition fed to the generator, along with random noise and black lines depict the features passed as ground truth to the discriminator.

Figure 6.4: Data Formatting for Each Quarter in EDC-GAN: Black lines represent the Ground-truth, red lines depict the Condition fed to the Generator.

Since the data is converted into a time-series format by transforming it into quarters, hence a generator is trained for each quarter separately in order to observe the learning of the model and its effectiveness with each additional quarter. The condition for $G$ is set such that for each current quarter its $G$ is conditioned to have the original discrete columns in the dataset and the generated continuous columns of the previous quarter. So, for the first quarter the condition passed (to the generator) will consist of discrete columns only since no generated continuous columns exist at that time-stamp. However for each of the next quarters, the condition $y$ consists of the discrete columns and the synthetically
generated continuous columns of all the previous quarters. In each quarter synthetic instances are generated after the completion of training, by passing a count of batches, that we want to generate, and condition vector y for each batch.

The proposed ED-GAN step-by-step upsampling approach is presented from Figure 6.5 to Figure 6.27. Also an overall detailed architecture of the proposed approach is presented in Figure 6.29, where for the first quarter 11 one hot vector is passed as condition to the generator along with a noise vector of 20.

The generator is composed of two hidden layers of 256 neurons each, and it outputs a vector 20 continuous feature. The output of the generator is used as input for the discriminator, along with the ground truth. The ground truth is a vector of 11 one-hot and 20 continuous features of Q1, encompassing the activities/interactions of the students. To make sure that the generated conditions are aligned with the ground truth, only the common discrete combinations are chosen from the generated instances.

![Diagram](image)

**Figure 6.5:** Input of the Generator for the First Quarter
Figure 6.6: Input of the Generator for the First Quarter (Noise)

Figure 6.7: Input of the Generator for the First Quarter (One Hot Vector)
Figure 6.8: Input of the Generator for the First Quarter (Concat Noise and One-Hot Vector)

Figure 6.9: Generator for the First Quarter
For each quarter output of the generator is a vector of 20 dimension.

Figure 6.10: Generator’s Output for the First Quarter

Figure 6.11: Discriminator’s Input-1 for the First Quarter
Figure 6.12: First Quarter: One-Hot Vector as Input in the Discriminator

Figure 6.13: First Quarter: Generated Data from the Generator as Input for the Discriminator
Figure 6.14: Ground Truth of First Quarter as Input for the Discriminator

Figure 6.15: First Quarter: Generated Data and Ground Truth as input in Discriminator
Now, the data for the first quarter is generated. For further analysis, detailed step-wise procedure is also explained for the second quarter. The vector output from the first quarter (20) is passed as condition in the generator of the second quarter, along with random noise and one-hot vector of 11 dimensions. The generator for the second quarter outputs 20 activity/continuous features. This output of the generator is fed as input to the discriminator along with the ground truth, comprising of a condition of $11 + Q_1$ vector, and $Q_2$ continuous features as ground truth. Similarly for the third quarter, the generated data from the previous quarters is passed as a condition in the generator, so it will be a vector of $11 + Q_1 + Q_2$ features, with 20 dimension of random noise as input for the generator.
Figure 6.17: Second Quarter: Generator’s Input
Figure 6.18: Second Quarter: One-Hot Vector added as Generator’s Input
Figure 6.19: Second Quarter: Previous Quarter Generated Data passed in this Generator’s Input
Figure 6.20: Second Quarter: Overall Dimensions of Generator’s Input
Figure 6.21: Second Quarter: Generator’s Output
Figure 6.22: Second Quarter: Concatenated Vector as Discriminator’s Input

Figure 6.23: Second Quarter: Q2 Generator’s Output passed as Discriminator’s Input
Figure 6.24: Second Quarter: Q2 Generator’s Output passed as Discriminator’s Input

Figure 6.25: Second Quarter: Ground Truth of 2nd Quarter as Input for this Discriminator
The output of the generator is 20 which is fed as input to the discriminator along with the ground truth data. This process is illustrated in Figures 6.26 and 6.27, which show the interactions between the generator and discriminator in a generator network context.
truth, comprising of 11+Q1+Q2 as condition and ground truth of Q3 as input for the discriminator. For the last quarter, condition for the generator is 11+Q1+Q2+Q3 and output of the generator is 20. This vector is also passed to the discriminator along with condition comprising of 11+ Generated Q1+ Generated Q2+ Generated Q3 features and 20 features of actual (ground truth) Q4 as input for the discriminator.

For further quarters, Q3 and Q4, the dimensions for the generator and discriminator’s input/output are presented in Figure 6.28 and Figure 6.29, respectively.

### 6.4.3 Experimental Setup for Proposed Temporal Models

This section discusses the experimental set-up for our proposed GAN-based models. The conventional practice in GANs framework is to update the parameters of discriminator $D$ and generator $G$ based on the output of $D$. This enables the learning of the discriminator to distinguish between actual and generated sequences whereas the generator inhibits the learning of the discriminator and tries to fool it so that it may assign wrong labels to the actual and generated sequences [126]. After a number of iterations generator will learn the distribution of the actual data sequences and will be able to generate realistic sequences whereas the discriminator will also be trained to distinguish between actual and synthetically generated sequences. In a GANs framework, generator takes sufficient iterations to learn the multivariate distribution of the data whereas discriminators are rather quick to train. Therefore, in our proposed GAN-based models, for each epoch, the generator is trained ten times more than the discriminator to balance the slow learning of the generator with the rapid learning of the discriminator. For weights initialization a variant of Xavier initialization was applied, as shown in equation 6.7:

$$W^l = \sqrt{\frac{2}{\text{size}^{l-1}}}$$  \hspace{1cm} (Equation 6.7)

Some random noise, equal to the neurons of the input layers, was fed to the proposed networks with learning rate of discriminator at $3e^{(-5)}$ and that of the generator at $1.8e^{(-4)}$. These learning rates
Figure 6.28: Proposed Conditional GAN Architecture for Q3
Figure 6.29: Proposed Conditional GAN Architecture for Q4
were selected after extensive experiments. The networks were trained from 2000-3000 epochs. For each experiment, the loss of the network was visualized after every 100 epochs and the settings which provided the most balanced behavior between generator and discriminator are reported in this study. The discriminator and generator are implemented through neural networks.

### 6.4.3.1 EDD-GAN Network Structure

The generator network for EDD-GAN consisted of fully connected layers with depth two and 128 neurons each; between each fully connected layer a batch normalization layer with 128 neurons and an activation function of leaky ReLU is added. The discriminator network is rather simple with a single fully connected layer of 64 neurons and sigmoid activation. In EDD the generator was executed twice as the discriminator. The EDD’s generator $G$ can thus be described in the following manner:

\[ x = z \oplus \text{cond} \]

\[ h_0 = \text{LeakyReLU}(BN(FC_{\text{cond}}[+|z|\rightarrow128](x))) \]

\[ h_1 = \text{LeakyReLU}(BN(FC_{\text{cond}}[+|z|+|128|\rightarrow128](h_0))) \]

### 6.4.3.2 EDC-GAN Network Structure

For EDC-GAN, the aforementioned settings for EDD were applied however the generators’ learning was slow and the loss didn’t converge therefore some more experiments were performed to tune the EDC model where the learning rates and neurons in the fully connected layers were tuned. The EDC’s generator network thus consisted of fully connected layers with depth three and 256 neurons each; between each fully connected layer a batch normalization layer with 256 neurons and an activation function of Leaky ReLU is added. The discriminator network is simpler with fully connected layers of depth 2 and 64 neurons each and sigmoid activation. In EDC, the generator was executed ten times more than the discriminator for each epoch and can be described in the following way:

\[ x = z \oplus \text{cond} \]
\begin{align*}
h_0 &= \text{LeakyReLU}(BN(FC_{\text{cond}} + x|\rightarrow 256(x))) \\
h_1 &= \text{LeakyReLU}(BN(FC_{\text{cond}} + z + |256| \rightarrow 256(h_0))) \\
h_2 &= \text{LeakyReLU}(BN(FC_{\text{cond}} + z + |256| + |256| \rightarrow 256(h_1)))
\end{align*}

Both the networks (EDD and EDC) were trained using binary cross-entropy loss with Adam activation. The synthetic data is generated by passing the count of batches and the discrete condition for each batch. In order to observe the performance of our proposed models in-comparison to the gan-based baseline approaches, the total number of clicks for each feature where normalized according to the number of the instances in each class and plotted during different stages of the training phase. Fig. 3 shows the number of features on the x-axis and y-axis represents the normalized total clicks for each feature. It was observed that in the initial stages the proposed GAN network produces poor results compared to the actual trend in the original data. During the later stages, as the network is updated, it learns the multivariate distribution of the actual data and thus produces a trend similar to that of the actual data. In comparison to the other techniques, EDC produces a trend similar to that of the actual data. For the CTGAN model, the trend in the start and end is similar to that of the actual data however in the middle it deviates from the original behavior. Similarly, EDD captures the actual trend for a few features whereas CGAN performs poorly compared to other models and only captures the actual trend for a few features. This behavior of the models demonstrates the significance of the discrete and continuous features for capturing the behavior of each student.

Since EDC consists of a one-hot vector constituting of each student’s course information and their interactions with the VLE, it more aptly encapsulates the behavior of each student. CTGAN performs inferior compared to our proposed approaches because the data consists of a few number of discrete columns therefore adding only these columns in the condition inhibits the model to learn the multivariate distribution of the data. Also in the educational datasets the continuous columns, consisting of the student interactions and their engagement patterns, are more inclined to determine the behavior of each student along with their course information. EDD performs better than CTGAN because EDD is quarter-specific, encapsulating the behavior of each quarter and forwarding it to the next quarters, whereas CTGAN is not quarter-specific and not for time-series data.
6.4.4 Evaluation of Proposed Temporal Models using OULAD

To evaluate the performance of our proposed GAN-based methods, the generated data from OULA was fed to the neural network and results were compared with the data generated from other class balancing techniques, as provided in Table 6.2. To tune ANN architecture, various experiments were performed with different optimizers and their default values. Further, the default learning rate values were then varied between the range of 0.1 to 0.0001, in various batches ranging from 64 to 264, activations relu, tanh, and sigmoid, with a combination of three hidden layers of 200, 100, 50 and 164, 128, 64. A learning rate of 0.01 with relu in the hidden layers (164, 128, 64 units) and sigmoid in the dense layer on a batch size of 64 produced optimal results. The setting providing optimal results was selected and was kept consistent for all the synthetically generated datasets.

Moreover, mean L1 Norm is also taken of all the GAN variants to demonstrate the effectiveness of the proposed approaches compared to the conventional ones. The mean L1 norm score is taken as:

\[ ||\text{GeneratedData} - \text{GroundTruth}|| \]

The results on all the deployed GAN variants are presented in Table 6.1. This mean score represents the difference in the generated and ground-truth values. The lower the mean, the more robust the generation method. It can be observed that our proposed models have a low mean L1 score compared to other GAN variants.

Table 6.1: Mean L1-Norm for all GAN-Variants

<table>
<thead>
<tr>
<th>GAN Variants</th>
<th>Mean L1 Norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGAN</td>
<td>44.77</td>
</tr>
<tr>
<td>CTGAN</td>
<td>23.09</td>
</tr>
<tr>
<td>One-Vector EDD-GAN</td>
<td>22.13</td>
</tr>
<tr>
<td>EDD-GAN</td>
<td>16.21</td>
</tr>
<tr>
<td>EDC-GAN</td>
<td>13.03</td>
</tr>
</tbody>
</table>
6.4.4.1 Evaluation Metrics

The performance of machine learning algorithms is typically evaluated using predictive accuracy. But for imbalanced datasets, where the costs of different errors vary significantly, it provides biased results, for instance it can give 99% accuracy if the model classifies all instances as the majority label. Therefore, for an imbalanced dataset, accuracy is not a good measure to signify the performance of a classifier; instead recall and precision are used to evaluate such imbalanced classification. Accuracy refers to the proportion of the correct predictions by the classifier. Precision is defined as the ratio of correctly classifying the students at-risk of failure from the overall students predicted as on-risk by the classifier, whereas recall measures the effectiveness of the classifier incorrectly predicting at-risk students from the total students that were actually on-risk. F1-score is a measure that reports a balance between precision and recall and is a much better metric to report the performance of the deployed classifiers. Similarly, AUC (area under the curve) is also an important metric for classifiers evaluation. It is a trade-off between recall and the false-positive rates and gives the probability of ranking a randomly chosen positive instance above a random negative instance. The values of AUC range from 0.5 to 1, with 1 indicating a perfect classifier that discriminates positive instances from negative ones [204].

6.4.4.2 Training & Validation Data

The training data constituted of a subset of the actual OULA data and the data generated through our proposed GAN-based model. Since in the actual data ‘pass’ instances exceed the ‘fail’ instances therefore the data was balanced by including fail instances from the generate data. The test set constituted of the actual OULA data instances that were not included in the training data. This was done in order to avoid overfitting issues and evaluate the performance of the proposed GAN-based model in a more rigorous manner since a random split is observed to produce biased results and inhibit robust predictions [205, 118]. The training set included 11,000 pass instances, 3,500 fail instances from the OULA dataset and 7,500 generated fail instances. The training set had equal students belonging to each class (i.e., pass and fail). The equal proportion of instances from both classes enabled the classifier to make predictions without bias. For testing, Train on Synthetic and Test on Real (TSTR) phenomenon was followed, with 4,000 real pass instances and 3,552 fail instances were used. Train-
ing and testing was done for each of the quarters separately and instances used in training were not repeated in the testing phase.

6.4.5 Results & Discussion

This section discusses the results of the various experiments performed using different baseline scores and their comparison with the proposed deployed methodologies. The results of the proposed methods are presented and discussed comprehensively to highlight the significance of the proposed methods.

Baseline Scores: To compare our results, firstly a baseline score was evaluated, comprising of the original OULA data. To get a baseline score the quarterly computed data was appended with each of the next quarters. This data was then used to train the neural network and the resulting F1-score, AUC value and accuracy were marked as baseline score. Since it caters the original OULA data with class imbalance, hence it is referred to as IM in the comparison Table 6.2. Table 6.2 also provides a comparison of various conventional up-sampling algorithms with CT-GAN and the proposed methods. For more baseline techniques, some conventional methods were deployed on the OULA data; Random Over-sampling (ROS), SMOTE, ADASYN, Borderline SMOTE2, K-means SMOTE and SVM SMOTE. These conventional methods for class balancing were deployed and compared for evaluating our proposed method.
Table 6.2: Comparison of the Proposed GAN-based Models with Various Over-sampling Techniques on Neural Network

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Testing Accuracy</th>
<th>F1-score</th>
<th>AUC Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DN</td>
<td>0.8036</td>
<td>0.7534</td>
<td>0.7934</td>
</tr>
<tr>
<td>IM</td>
<td>0.8144</td>
<td>0.7213</td>
<td>0.7332</td>
</tr>
<tr>
<td>IM-GF</td>
<td>0.7231</td>
<td>0.6342</td>
<td>0.6537</td>
</tr>
<tr>
<td>ADASYN</td>
<td>0.8096</td>
<td>0.7334</td>
<td>0.8256</td>
</tr>
<tr>
<td>BorderLine</td>
<td>0.8178</td>
<td>0.7416</td>
<td>0.8352</td>
</tr>
<tr>
<td>Kmeans-SMOTE</td>
<td>0.8477</td>
<td>0.7435</td>
<td>0.8270</td>
</tr>
<tr>
<td>SVM-SMOTE</td>
<td>0.8336</td>
<td>0.7391</td>
<td>0.8375</td>
</tr>
<tr>
<td>CGAN</td>
<td>0.8624</td>
<td>0.7494</td>
<td>0.8550</td>
</tr>
<tr>
<td>CTGAN</td>
<td>0.8801</td>
<td>0.7916</td>
<td>0.8906</td>
</tr>
<tr>
<td>One-Vector EDD-GAN</td>
<td>0.8821</td>
<td>0.7942</td>
<td>0.8926</td>
</tr>
<tr>
<td>EDD-GAN</td>
<td>0.8932</td>
<td>0.8299</td>
<td>0.9132</td>
</tr>
<tr>
<td>EDC-GAN</td>
<td>0.9084</td>
<td>0.8566</td>
<td>0.9378</td>
</tr>
</tbody>
</table>

**Down-Sampled OULAD:** In the OULA data, pass instances are twice then the fail instances, therefore the pass instances were downsampled to balance with the fail instances. The balanced data was fed to the neural network and results were compared with other deployed class balancing techniques. This is referred to as DN in the comparison table (Table 6.2).

**Imbalanced-Generated Fail:** In the original OULA dataset, the real fail instances were replaced with generated fail instances. The ‘IM’ algorithm in Table 6.2 represents the prediction results on the original OULA dataset with 15,000 pass and 7,052 fail instances. The IM-GF method uses generated fail instances to analyse the quality of the generated instances. It can been observed that the generated instances perform a little less compared to the original instances.

**Conventional CGAN:** A conventional CGAN was trained on time-series data. This is labeled as CGAN in the comparison table (Table 6.2) and has been discussed previously.

**Sequential Conditional GAN:** Sequential Conditional GAN (SC-GAN), follows the same behavior as that of a baseline conventional CGAN, however for each additional quarter the discriminator inputs for the network are updated with the previously generated instances of the quarters. The condition for each quarter is the class label (pass/fail) however the input for the discriminator is modified according to each quarter.
CTGAN on Time-Series Data: CTGAN [208] does not cater the time-series functionality of tabular data, therefore all four quarters were generated in a single model, by appending one quarter after another. The generated data was saved in files and fed to the neural network to evaluate the efficiency of CTGAN on time-series data compared to our proposed method.

One-Vector EDD-GAN: To analyse the performances of our proposed EDD-GAN, we also experimented with generating all the four quarters in one vector with the discrete variables passed as a condition to the generator.

In order to observe the performance of our proposed GAN variants for the early prediction, quarter wise analysis was also conducted. For each quarter, upsampled instances were used in the training data and for testing real/original data was used. Table 6.3 shows the quarter-wise analysis on the data upsampled from EDD-GAN and Table 6.4 presents the quarter-wise analysis results for EDC-GAN.

Table 6.3: Quarter-wise Machine Learning Efficacy for EDD-GAN

<table>
<thead>
<tr>
<th>EDD GAN</th>
<th>Testing Accuracy</th>
<th>F1-Score</th>
<th>AUC Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>0.791</td>
<td>0.743</td>
<td>0.824</td>
</tr>
<tr>
<td>Q1-Q2</td>
<td>0.832</td>
<td>0.782</td>
<td>0.846</td>
</tr>
<tr>
<td>Q1-Q3</td>
<td>0.854</td>
<td>0.802</td>
<td>0.862</td>
</tr>
<tr>
<td>Q1-Q4</td>
<td>0.893</td>
<td>0.829</td>
<td>0.913</td>
</tr>
</tbody>
</table>

Table 6.4: Quarter-wise Machine Learning Efficacy for EDC-GAN

<table>
<thead>
<tr>
<th>EDC GAN</th>
<th>Testing Accuracy</th>
<th>F1-Score</th>
<th>AUC Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>81.42</td>
<td>0.764</td>
<td>0.841</td>
</tr>
<tr>
<td>Q1-Q2</td>
<td>85.29</td>
<td>0.805</td>
<td>0.883</td>
</tr>
<tr>
<td>Q1-Q3</td>
<td>88.32</td>
<td>0.821</td>
<td>0.906</td>
</tr>
<tr>
<td>Q1-Q4</td>
<td>0.908</td>
<td>0.856</td>
<td>0.937</td>
</tr>
</tbody>
</table>

Comparative Analysis of Proposed Methods: We compared all the above variations with our proposed adversarial methods. These are labeled as EDD and EDC in the comparison table (Table 6.2) and have been explained in the previous sections.

Table 6.2 shows a summary of the classification results obtained from the datasets generated through various up-sampling techniques including our proposed methodologies. As provided in Table 6.2,
it can be observed that our proposed methodologies perform better than both GAN-based and non-GAN based baseline methodologies. The imbalanced dataset shows improved validation accuracy compared to the balanced downsampling (DN) dataset. However, accuracy is a biased measure, and therefore F1-measure and AUC are more important metrics for evaluation. The F1-measure for DN demonstrates its better performance compared to the IM classifier. GAN-based baseline methods perform better compared to the non-GAN based methods, highlighting the superiority of the deep learning approaches compared to the conventional approaches. Also, since the baseline scores do not cater the time series functionality of the data, therefore our proposed methodologies perform better than others.

To demonstrate the effectiveness of our proposed methods, Fig. 6.30 presents the comparison graphs with GAN-based and some traditional non-GAN based techniques. As illustrated in Fig. 6.30a, the AUC, precision, and recall measures for our proposed methods are better than other comparable oversampling methods. EDC outperforms CTGAN, CGAN, SMOTE, and ROS with an AUC of 0.8566, precision of 0.9169, and recall of 0.8038. Similarly, EDD outperforms others with an AUC of 0.8299, precision of 0.8902, and recall of 0.7774. The graph also represents the effectiveness of the EDC method in comparison to EDD with EDC outperforming EDD with 2.69% AUC, 2.99% precision and 3.39% recall. Fig. 6.30b illustrates the comparison of various up-sampling approaches in regards to learning accuracy, validation accuracy, and F1-measure. As presented in the graph, the proposed methods outperform the other comparable approaches. EDC outperforms all other methods with a validation accuracy of 0.9084 and an F1 measure of 0.8566, next in line is EDD outperforming all other comparable GAN-based and traditional non-GAN based approaches with a validation accuracy of 0.8932 and an F1 measure of 0.8299. EDC outperforms EDD with an F1 measure of 3.217%.

It can be observed that the proposed methodologies perform better than the baseline scores. GAN-based approaches perform better than other conventional up-sampling techniques. This phenomenon emphasizes the significance of the deep learning methods in learning the behavior of the data. GAN-based approaches perform better at learning the behavior of each student compared to other traditional non-GAN based techniques, as provided in Table 6.2 and illustrated in Fig. 6.30. Similarly, our proposed methodologies that encapsulate the behavior of students for each quarter, through their course information and their engagement patterns, unlike the other deployed up-sampling approaches, perform better at capturing and learning the behavior of students. Moreover, EDC outperforms EDD, signifying that the engagement pattern of students and their interactions with the VLE are important.
Figure 6.30: Performance Comparison for the Various Deployed Proposed and Conventional Up-Sampling Techniques
contributors in determining their overall performance along with their courses information. This study presents a solution for the time-series sequential data where each new time stamp encapsulates the behavior of the previous time stamp for each student such that each students interaction behavior in one quarter is related to its behavior in the previous quarter. The existing studies lack this association of the behavior across different time-stamps for each student. The conventional and other GAN-based approaches deployed do not associate the behavior of students with their previous time stamp behavior in the same course duration; therefore our proposed methodologies are novel in this regard.

This section examined the problem of class imbalance in the Learning Analytics community using the deep learning based adversarial networks. The first objective of the study was to replicate a conventional adversarial network on the educational data constituting of students interactions with the VLE. The study deploys a conventional conditional GAN on the OULA data, where the overall performance (pass/fail) of each student is fed as condition to the network. Secondly, the study aims to transform the class imbalance problem in a sequential time-series setting, where each instance is associated with its previous time-stamp/sequence. The existing studies related to time-series data lack this association of the instances therefore our study is first of its kind that makes the use of adversarial networks to not only balance out the class imbalance problem but also to incur an association of each instance with its previous time stamps such that the behavior of each student is encapsulated throughout the sequences. The results demonstrate that discrete and continuous both features are significant predictors of their behavior. Our first proposed model EDD encapsulates the behavior through the use of courses and its subsequent performance information. The second proposed model EDC, moving a step further, also includes the students’ interactions with the VLE to capture their behavior. The condition for this model is formulated by concatenating the given discrete and continuous attributes in the OULA dataset.

The performance for the generated balanced datasets is evaluated through different non-GAN and GAN based baseline approaches. The results demonstrate the effectiveness of the proposed approaches compared to the baselines with EDD and EDC outperforming the conventional CGAN by 3.57% and 5.33% validation accuracy, respectively and surpassing the CTGAN by 1.48% and 3.22% validation accuracy, respectively. Similarly, both the models also outperform in their F1-measure by 5.07% and 14.30% respectively for CGAN and 3.34% and 4.43% respectively for CTGAN. Similarly, results for AUC also demonstrate proposed models to outperform both the non-GAN and GAN-based up-sampling methods, with EDD outperforming CGAN and CTGAN by 6.81% and 2.53% AUC and
EDC outperforming CGAN and CTGAN by 9.68% and 5.29% AUC, respectively. This study presents a novel method to encapsulate the behavior of students in a time-series sequence such that each student’s previous sequence is associated with its next sequence, capturing its behavior throughout the sequences. This study also highlights the importance of the students’ interactions and its impact on the overall classifier. To capture the behavior of each student, its interactions with the VLE should also be catered to improve the classifier.

6.5 Summary

This chapter comprehensively described the deployed Adversarial based approaches proposed to eliminate the class-imbalance issue, in a temporal setting for early prediction of students’ academic performances. The results of the proposed up-sampling approach in a temporal setting are presented and the significance of the deployed adversarial-based approach is discussed. Through the inclusion of engagement information of students, the proposed GANs are trained to generate instances of minority class and augmenting the training set to eliminate class imbalance. The proposed networks are evaluated by deploying a series of conventional non-GAN based and some GAN-based up-sampling techniques. The results are compared with neural networks to demonstrate the effectiveness of the proposed GAN-based models.
Chapter 7

Conclusion and Recommendations

“No book can ever be finished. While working on it we learn just enough to find it immature the moment we turn away from it.”
— Karl R. Popper, The Open Society and Its Enemies

The overarching idea for this dissertation study revolves around the objectives for predicting students performance in an online environment, from the data collected from such virtual interactions between the students and the online learning platforms. In this chapter, we briefly summarize the results for each of our intended objectives, stated in Section 1.6. This chapter concludes the discussion of our objectives by highlighting their significant results and implications to the educational community. Some possible future directions are also presented in the prediction of academic performances of students, some limitations possessed with our study, along with some recommendations to the learning analytics research community.

7.1 Value Added

In the recent times, the pandemic has revolutionized the teaching mechanisms for both learners and instructors and providing corrective strategies to improve this mechanism is the need of the hour. This has also enabled the accumulation of sufficient repositories pertaining to the interaction of students, thus exploring this dimension in the educational setting is of utmost importance and globally researchers are now more than ever vigilant to devise the tools and techniques for improved and reliable measures of online education and exploring the implications of this mode of study.

To present the important potential dimensions of the Learning Analytics discipline, firstly a bibliometric study was conducted. The bibliometric study presented the research productivity in learning analytics using Scopus database over 2000 to 2017. Its research landscape was examined and explored to analyse it at various levels, including investigating the prominent countries, institutions and sources to visualize current trends in this field. Overall, learning analytics resonates strongly with
educational datasets, performing data-mining techniques on it, developing strategies to assess and evaluate learners’ performance, and assisting higher education decision-making. Learning analytics’ influence on higher education and its decision-making is a relatively new area, therefore no significant clusters for it were observed. In the coming years, with new and innovative research in this domain, more rational strategies will be developed to aid higher education decision-making.

Furthermore, this dissertation presented a contribution to knowledge in early prediction of students at-risk of a failure, determining students likely to withdraw from modules and ascertaining significant features that enable a student to outperform others. Results reveal demographic characteristics and students’ clickstream activity, after the module initiation, as having a significant impact on their academic performances. It also determines the effectiveness of the deep learning models in the early prediction of academic performances, enabling instructional interventions by the university to implement corrective strategies for students support and counseling. Due to the class imbalance problem in ‘distinction’ instances, a discrete pattern for such students was not observed, a limitation of our study. However, demographic and geographic characteristics tend to significantly impact performance. The performance evaluation model shows a sensitivity of 69%, a precision of 93% and overall accuracy of 88% in predicting at-risk students; a sensitivity of 86%, a precision of 96% and overall accuracy 93% in predicting early withdrawals. Similarly, ascertaining ‘distinction’ students from ‘fail’, a sensitivity of 74%, and precision of 81% and overall accuracy of 85% is achieved.

Another dimension of our study, examined the successful prediction of students through an application of conventional classifiers by monitoring the engagement of students with a learning platform used by learners enrolled in fully online courses that were offered in self-paced learning mode. We, firstly provided a comprehensive comparison of deep learning algorithms with conventional machine learning algorithms to predict students at risk to fail a course based on log data about the students’ interaction with the learning platform. The study results indicate that a deep learning algorithm outperformed the convivial and commonly used machine learning options. The deep LSTM technique outperformed all other alternatives and achieved an accuracy of 84.57%, precision of 82.24%, recall of 79.43%, and AUC of 82%. This demonstrates a strong potential of deep algorithms to provide a foundation for higher education institutions to address challenges associated with self-paced education in fully online learning settings.

The interpretability of deep learning algorithms is often mentioned in the literature as one of the
main issues that could potentially prevent their use in education. Therefore, to contribute to the
debate around the interpretability of the deep learning algorithms in education, a part of this study
proposed and empirically validated an approach that can identify the most significant predictors of
at-risk students. The approach is based on the Shapely Additive Explanation and can identify patterns
of association between the significance index and the changes in values of individual features. This
can offer valuable insights both to i) learning analysts to create the most parsimonious models of pre-
diction of students at risk; and ii) instructors and educational designers to evaluate the value of some
of their design decisions and teaching practices as well as to inform future interventions. Although
it was beyond the scope of the current study, future research should also look for the most effective
approaches to the presentation of significance analysis to support sensemaking and decision-making
processes of different stakeholder groups involved in (support of) teaching.

Moreover, our study showed that the LSTM deep learning algorithm outperformed the conventional
machine learning algorithms in the early prediction of students at risk. In fact, the accuracy of the
LSTM algorithm measured through AUC with the data accumulated after the first 5 (0.71) and 10
(0.74) weeks of the course outperformed almost all other conventional algorithms even when they
are used on the complete dataset collected for the entire duration of the course. This comes from
the property of LSTM to use computed log data on a weekly basis to convert performance prediction
into a time-series classification problem for early prediction. In practical terms, a relatively high level
of prediction accuracy, only after five weeks of course duration, offers promising opportunities for
early interventions in self-paced education where courses are offered for an extended period (6 or 9
months).

Another dimension of our study was to examine the problem of class imbalance in the learning an-
alytics community using the deep learning based adversarial networks. Firstly, we replicated a con-
ventional adversarial network on the educational data constituting of students interactions with the
VLE. The study deploys a conventional conditional GAN on the OULA data, where the overall per-
formance (pass/fail) of each student is fed as condition to the network. Secondly, the study aims to
transform the class imbalance problem in a sequential time-series setting, where each instance is as-
sociated with its previous time-stamp/sequence. The existing studies related to time-series data lack
this association of the instances therefore our study is first of its kind that makes the use of adver-
sarial networks to not only balance out the class imbalance problem but also to incur an association
of each instance with its previous time stamps such that the behavior of each student is encapsulated
through-out the sequences. The results demonstrate that discrete and continuous both features are significant predictors of their behavior. Our first proposed, class balancing model, EDD encapsulates the behavior through the use of courses and its subsequent performance information. The second proposed model EDC, moving a step further, also includes the students’ interactions with the VLE to capture their behavior. The condition for this model is formulated by concatenating the given discrete and continuous attributes in the OULA dataset. The performance for the generated balanced datasets is evaluated through different non-GAN and GAN based techniques. The results demonstrate the effectiveness of the proposed approaches compared to the baselines with EDD and EDC outperforming the conventional CGAN by 3.57% and 5.33% validation accuracy, respectively and surpassing the CTGAN by 1.48% and 3.22% validation accuracy, respectively. Similarly, both the models also outperform in their F1-measure by 5.07% and 14.30% respectively for CGAN and 3.34% and 4.43% respectively for CTGAN. Similarly, results for AUC also demonstrate proposed models to outperform both the non-GAN and GAN-based up-sampling methods, with EDD outperforming CGAN and CTGAN by 6.81% and 2.53% AUC and EDC outperforming CGAN and CTGAN by 9.68% and 5.29% AUC, respectively. This study presents a novel method to encapsulate the behavior of students in a time-series sequence such that each student’s previous sequence is associated with its next sequence, capturing its behavior through-out the sequences. This study also highlights the importance of the students’ interactions and its impact on the overall classifier. To capture the behavior of each student, its interactions with the VLE should also be catered to improve the classifier. Timely intervening the students is an important determinant of students’ success that incorporates the detrimental factors associated with students academic success. We highlighted this crucial aspect of the academic success of students that can be further utilized for constructive and formative pedagogical guidelines.

7.2 Limitations

A major limitation in this field is procuring the educational datasets since institutions are hesitant to share their students information. Therefore, only selected datasets are available publicly. Higher research committees should formulate some guidelines for the institutions so that they may share their educational datasets on the terms of anonymity.

The deployed OULA data only consisted of clickstream information of students, therefore no con-
tent analysis on student behavior was performed in this thesis. The data consists of performances of students associated with their clickstream information with the VLE, courses taken and assessment related information. It consists of the student participation with discussion forums in the form of clicks only. For a more thorough analysis of student behavior, a content analysis of students’ participation associated with their performances may be done, provided the data consists of such engagement interaction. Therefore, assessing student behavior through content analysis was beyond the scope of this study.

The proposed adversarial networks are mapped with the problem under-study, therefore for temporal class-imbalances in other disciplines, some elements of our methods may have to be tweaked.

7.3 Future Directions

The following are some of the most promising areas of research in the field of learning analytics, according to our perceptions through the detailed study of the literature analysis presented:

- Sophisticated enhancement of learning analytics student profiles, including data from different data streams, available from higher education institution data warehouses or social media.
- Advanced matching algorithms for skills and competencies building, based on maintained learning analytics ecosystems.
- Integration of learning analytics research with value adding services, both in the smart education and smart libraries context.
- Extensive use of learning analytics for sophisticated and machine-learning approaches to innovative and creating thinking capable of promoting students’ entrepreneurship and innovation capabilities.
- Dynamic, cognitive computing-based systems for the self-assessment and self-regulation of learning performance and allocation of learning resources of students in higher education.
- Development of indexes and KPIs related to the efficiency and the predictive capability of learning analytics for higher education learning outcomes.
Integration of administrative quality factors to learning analytics requirements throughout the entire learning process.

Extensive documentation and provision of learning analytics for disabled learners.

Advanced research on the visualization and use of learning analytics on a real-time basis on higher education analytics.

Promotion of mobile learning analytics ecosystems.

The underlying premise of the work presented in this thesis, focused on the measures to early predict the academic performances of students, that may assist in the formulation of pedagogical interventions. In the context of the academic performances of students, the issue of class imbalance is addressed to improve the prediction process of the machine learning models. An improved alarm system will assist in the development of corrective strategies and actions for timely instructional interventions. Given the challenges associated with existing approaches of students’ academic performance prediction, the objective of this thesis was to develop a class balancing technique that could assist the learning analytics community in building early alarm systems for pedagogical and instructional interventions by the academia.

Future research needs to explore additional avenues. First, the sensitivity of the deep learning approaches to instructional conditions, both in terms of the most significant predictors and early prediction, needs to be investigated. Second, future research should investigate if and if so, to what extent the most significant predictors change over time and if these predictors change with the nature of the courses offered. Third, the literature often indicates the importance of other data sources that are commonly used in student information systems. As this type of data was not available in the OULA, future research should explore how deep learning algorithms perform when built upon combinations of log data and other data sources.

Moreover, further research should also look for the most effective approaches to the presentation of significance analysis to support sense-making and decision-making processes of different stakeholder groups involved in (support of) teaching. The impact of these stakeholders on the performance of learners and devising mechanisms for introducing hybrid learning designs that influentially improve the learning process for students. This dissertation primarily emphasizes on the learner’s perspective of the learning analytics community, introducing models to improve their success prediction especially
in a self-paced environment. However, as a future avenue the instructor’s perspective should be studied comprehensively, devising tools and techniques to improve online teaching and measuring its impact from a learners perspective.
REFERENCES


