JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH AND ANALYTICS

Volume 18, No. 2
July 2025
ISSN: 1946-1836

In this issue:

iscap
Information Systems & Computing Academic Professionals

The **Journal of Information Systems Applied Research and Analytics** (JISARA) is a double-blind peer reviewed academic journal published by ISCAP, Information Systems and Computing Academic Professionals. Publishing frequency is three issues a year. The first date of publication was December 1, 2008. The original name of the journal was Journal of Information Systems Applied Research (JISAR).

JISARA is published online (https://jisara.org) in connection with the ISCAP (Information Systems and Computing Academic Professionals) Conference, where submissions are also double-blind peer reviewed. Our sister publication, the Proceedings of the ISCAP Conference, features all papers, teaching cases and abstracts from the conference. (https://iscap.us/proceedings)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%) and other submitted works. The non-award winning papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in JISAR. Currently the acceptance rate for the journal is approximately 35%.

Questions should be addressed to the editor at editor@jisara.org or the publisher at publisher@jisara.org. Special thanks to members of ISCAP who perform the editorial and review processes for JISARA.

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH AND ANALYTICS

Editors

2025 JISARA Editorial Board

Towards Adaptive Learning: A Review of Machine Learning on LMS Data

Cindy Zhiling Tu
cindytu@nwmissouri.edu
Northwest Missouri State University
Maryville, MO 64468

Gary Yu Zhao
zhao@nwmissouri.edu
Northwest Missouri State University
Maryville, MO 64468

Omar El-Gayar
Omar.el-gayar@dsu.edu
Dakota State University
Madison, SD 57042

Abstract

This study presents a literature survey on the application of machine learning (ML) in learning management system (LMS) data analytics, aiming to provide insights into adaptive learning development and propose an agenda for future research. The literature survey is based on a proposed adaptive learning framework and critically analyzes the results within this context. The results reveal that machine learning methods can be used to evaluate the effectiveness of instructional interventions and combining online behaviors with textual data can improve the outcome of performance prediction. Key findings also highlight several open issues, including using small datasets and the need for comprehensive ML methods and algorithm development. Future research directions include improving the accuracy of student performance prediction, supporting instructional interventions, enriching student engagement through multimodal LMS data analytics, and leveraging big data and ML approaches for learning behavior pattern detection.

Keywords: adaptive learning, machine learning, Learning Management System (LMS), data analytics, literature survey.

Recommended Citation: Tu, C., Zhao, G., El-Gayar, O., (2025). Towards Adaptive Learning: A Review of Machine Learning on LMS Data. *Journal of Information Systems Applied Research and Analytics* v18, n2, pp 4-19. DOI# https://doi.org/10.62273/XRHN6187.

Towards Adaptive Learning: A Review of Machine Learning on LMS Data

Cindy Zhiling Tu, Gary Yu Zhao and Omar El-Gayar

1. INTRODUCTION

The digital transformation of education has brought significant advancements in how learning is delivered and managed. Adaptive learning has emerged as a promising technology and a new teaching paradigm in higher education (Xie et al., 2019). Adaptive learning is a pedagogical approach that uses technology to provide corresponding educational experiences to individual learners' needs (Li et al., 2021). The adaptive learning environment is personalized to meet the unique needs of individual learners by dynamically adjusting the instruction based on real-time data to optimize the learning process and make it more effective and efficient (Cavanagh et al., 2020). Adaptivity occurs in instructional activities such as the content, the assessment, and the instruction sequence (Castro, 2019) based on the learner's learning performance and characteristics. Higher education institutions need to use instructional content and students' learning data to conduct adaptive learning systems.

A Learning Management System (LMS) is a software application for administering, documenting, tracking, reporting, and delivering educational courses, training programs, or learning and development programs (Elfeky & Elbyaly, 2021; Nizam Ismail et al., 2019). The data generated by an LMS includes learner-generated, teacher-generated, and system-generated data. LMS data contains a wealth of information about learning and teaching behavior and outcomes. As more LMS data becomes available, improving the capabilities for leveraging this data is essential to gain insights into learning and teaching activities (Tenzin et al., 2020; Zhu et al., 2022). Accordingly, learning analytics using machine learning (ML) techniques to analyze LMS data has gained significant attention in recent years. ML-based learning analytics can provide valuable insights and support for various learning theories and pedagogical interventions by analyzing data generated in educational contexts. Compared to traditional statistical analysis methods, ML methods can provide better accuracy and deal with complexity in data analytics, which offers powerful tools that can inform teaching practices and improve student learning experiences (Riestra-González et al., 2021; Villegas-Ch et al., 2020).

Research has been done on using ML in LMS data analytics to enhance adaptive learning, including delivering learning content, adapting to the individual learner's needs, and providing recommendations for learning paths (Kabudi et al., 2021). In addition, previous studies on ML-based LMS data analytics focus on predicting student performance and analyzing student interactions with LMS platforms to attain perspectives into student discourse in online discussions, identifying at-risk students, and improving student engagement and teaching practices (Gasevic et al., 2014; Korkmaz & Correia, 2019; Tenzin et al., 2020). However, the evidence regarding the potential connection between challenges experienced by students and teachers and the effectiveness of ML-based learning analytics and interventions in resolving these issues, the grounding in relevant theories, the appropriateness of various techniques, and the suitability of the data remains unclear.

This literature survey aims to provide a comprehensive overview of the current state of research at the intersection of machine learning, LMS data, and adaptive learning. This study addresses the following research questions: (1) Which ML methods and LMS data are used for various learning analytics objectives/outcomes in existing literature? (2) To what extent are ML-based LMS data analytics interventions grounded in adaptive learning? (3) What are the challenges and future research directions in leveraging ML in advanced learning analytics?

2. METHODOLOGY

Given the demonstrated potential of ML-based LMS learning analytics, we propose a literature survey framework adapted from Peng et al.'s (2019) personalized adaptive learning model. As shown in Figure 1, the adaptive learning route has three levels: "what to learn" - based on the learner's characteristics; "how to learn" - based on the learner's performance; and "how well learned" - based on the learner's personal development (Peng et al., 2019).

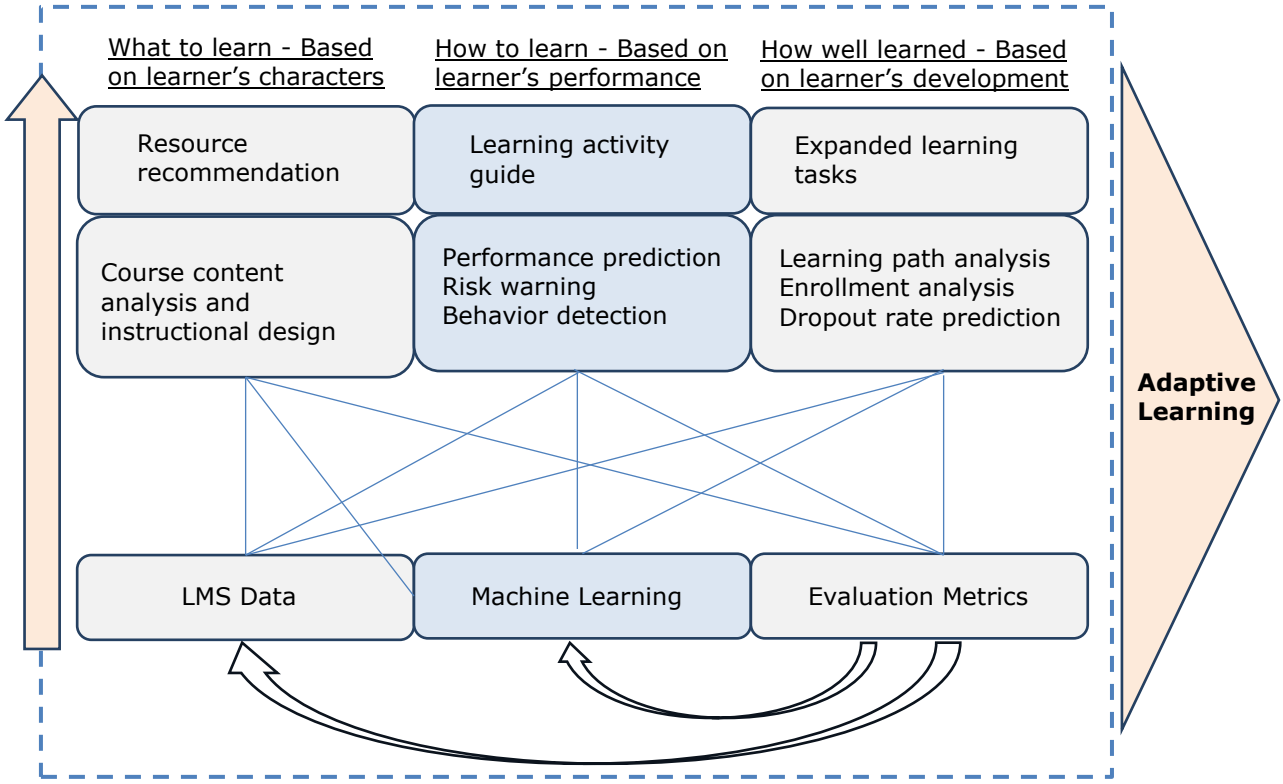

Figure 1. A Framework for Applying ML-based LMS Data Analytics to Adaptive Learning

In each level, three phases of data-driven pedagogical decisions based on ML-based learning analytics represent the ordinate. In the "what to learn" level, learning analytics focuses on learning content analysis and instructional design to tailor the learning resources that can match learners' characteristics. The ML-based analytics process must serve this objective, including LMS data collection, variable selection, ML model determination and training, model performance evaluation, and optimization. Moreover, the content may undergo continuous refinement through multiple iterations and incremental adjustments to accommodate the variations and the evolving individual characteristics of learners. In "how to learn" adaptive learning level, the data-driven pedagogies focus on guiding learning activity based on learners' performance (Peng et al., 2019). At this level, LMS data, ML algorithms, and evaluation metrics are determined by learning performance prediction, risk-warning, and learning behavior detection. In "how well learned" level, the data-driven pedagogies focus on expanded learning tasks based on learners' learning progress and personal development (Peng et al., 2019). To achieve this goal, the ML analytics processes need to provide learning path analysis, course enrollment analysis, and dropout rate prediction.

Notably, the three levels of adaptive learning paths have different weights and are not necessarily sequenced as in our model. Thus, our framework can be customized to fit various application contexts. Further, ML-based analytics are iterative processes, which means that based on the analytics outcomes, the LMS data, ML algorithms, and evaluation metrics must be adjusted and refined multiple times to achieve better performance.

We searched from five online databases: IEEE Xplore, ACM Digital Library, ProQuest Research Library, ABI/INFORM, and ScienceDirect (Elsevier) using two sets of keywords: ("machine learning" OR "ML" OR "analytics" OR "data analytics") and ("learning analytics" OR "learning management system" OR "LMS"). We combine these two keywords sets for each database as the search string. Studies must meet the following criteria: 1. Study of machine learning in LMS data analytics/learning analytics; 2. Full-text paper available; 3. Peer-reviewed paper; 4. Published between January 1, 2013 and January 31, 2023; 5. Written in English. Dissertations/theses, reviews, abstracts, books, book chapters, and reports are excluded from

this survey. Then, we manually scanned abstracts and filtered out irrelevant articles focusing on education curriculum, pedagogy, impacts, professional development, special external data sources, etc. In addition, we use the snowball technique to identify other relevant papers.

A total of 114 articles were extracted from all online databases. Two authors manually scanned titles and abstracts and filtered out irrelevant articles focusing on education curriculum, pedagogy, impacts, professional development, special external data sources, articles based on the inclusion/exclusion criteria, etc. Then, we conducted the full-text screening. Two authors cross-checked those included articles. In addition, we used the snowball technique to identify other relevant papers in the full-text screening stage. Finally, 52 peer-reviewed academic articles are selected for analysis. These articles are numbered for analysis purposes (see Appendix A).

3. LITERATURE SURVEY RESULTS

We identified the relevant information based on our survey framework and extracted it from each paper. For synthesizing the extracted data, we divided the data form into (i) demographic and contextual attributes and (ii) adaptive learning analysis. The first data set was analyzed through statistical techniques, and descriptive results were produced. The second set of data items was analyzed with a thematic analysis method.

Demographic Distribution

Figure 2 shows the number of selected papers published annually within the survey period. The number of published studies on the application of machine learning methods in LMS data analytics has been increasing since 2019 and reached a peak in 2020. 38 papers out of 52 (73%) were published in the past three years, signifying increasing interest, possibly due to the rise in online education since the pandemic and the increasing availability of learners' data.

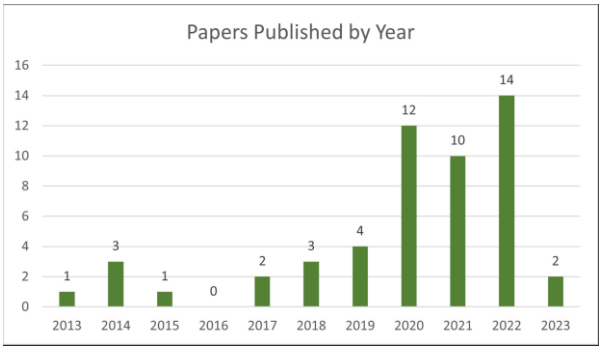

Figure 2: Papers Published by Year

As shown in Table 1, studies were reported from 27 countries. The United States accounted for most of the studies (17% or 9 studies), followed by the United Kingdom, Canada, India, Pakistan, and Greece (3 for each). Our findings are consistent with the prevalence of online education and technology in those countries.

Country	Paper Count	Country	Paper Count
United States	9	Bangladesh	1
Canada	3	Belgium	1
Greece	3	Ecuador	1
India	3	Hungary	1
Pakistan	3	Indonesia	1
United Kingdom	3	Kenya	1
China	2	Korea	1
Brazil	2	Malaysia	1
Croatia	2	Morocco	1
Japan	2	New Zealand	1
Spain	2	Philippines	1
Taiwan	2	Switzerland	1
Vietnam	2	Turkey	1
Australia	1		

Table 1: Distribution by Country

Analysis Based on Proposed Framework

Distribution of Papers by LMS Data Type. Table 2 shows the distribution of the 52 studies by LMS data type. Assessment data has been the most often utilized (27 papers or 52%). The following are learner's data (25 or 48%), user activity data (20 or 38%) and behavior log data (17 or 33%).

LMS Data	Articles	#	%
Assessment data - question and grade related to assignment, test, quiz, exam, etc.	P1,P4,P6,P7,P9,P10,P12,P14,P18,P19,P21,P23,P24,P27,P30,P31,P32,P33,P36,P39,P40,P41,P43,P44,P45,P46,P50	27	52
Leaner Demographic data and socioeconomic data - age, gender, location, device, enrolment, income, etc.	P4,P6,P7,P9,P10,P11,P12,P13,P15,P19,P25,P27,P30,P31,P32,P33,P36,P39,P40,P41,P43,P45,P46,P48,P52	25	48
Activity data - submissions, comments, posts, etc.	P1,P3,P4,P6,P8,P10,P13,P16,P18,P22,P23,P24,P26,P30,P32,P42,P44,P46,P48,P51	20	38
User behaviour log - navigation, page views, time spent on the platform, etc.	P6,P11,P14,P15,P17,P22,P25,P26,P28,P29,P33,P34,P37,P38,P45,P49,P51	17	33
Course Information - content webpage, instructor, start/end date, number of students, etc.	P4,P15,P16,P18,P20,P21,P26,P28,P30,P32,P33,P35,P37,P38,P50,P51	16	31
Interaction data - discussions, forum posts, announcements, messages, etc.	P2,P3,P9,P10,P22,P24,P26,P30,P44,P47,P48,P49,P51	13	25
Multi-modal data - audio, video, presentation, sensor signal, body posture and hand gesture, etc.	P2,P5,P20,P29,P47,P48,P52	7	13
Learning progress data - time spent on each module, the percentage of course completion	P13,P33,P35,P38,P49	5	10

Table 2. Distribution by LMS Data Type

Distribution of Papers by Machine Learning Approach. Table 3 shows the distribution of the 52 studies by machine learning method. Overall, SVM and Random Forest are the most often used machine learning methods (24 papers or 46% for each), followed by Logistic Regression (17 or 33%), Decision Tree (16 or 31%), and MLP Neural Network (14 or 27%).

Machine Learning	Articles	#	%
Support Vector Machine (SVM)	P3,P5,P6,P7,P9,P13,P15,P17,P22,P30,P31,P32,P34,P35,P36,P39,P40,P41,P43,P45,P46,P48,P52	24	46
Random Forest (RF)	P1,P6,P7,P9,P10,P11,P15,P17,P18,P20,P27,P29,P30,P32,P33,P35,P36,P38,P39,P40,P42,P48,P51,P52	24	46
Logistic Regression (LR)	P4,P6,P9,P12,P17,P22,P24,P25,P30,P33,P34,P40,P41,P42,P45,P46,P49	17	33
Decision Tree (DT)	P1,P3,P10,P11,P12,P15,P18,P26,P28,P34,P36,P39,P40,P46,P47,P51	16	31
MLP Neural Network (MLPNN)	P3,P9,P11,P13,P18,P19,P23,P28,P30,P34,P43,P44,P45,P46	14	27
KNN	P1,P7,P9,P11,P12,P22,P28,P30,P32,P33,P41,P46,P48	13	25
Naïve Bayes (NB)	P1,P6,P11,P18,P32,P33,P34,P40,P41,P42,P43,P49,P52	13	25
Clustering - K-means	P8,P16,P25,P27,P47,P49	6	12
Bayesian Network (BN)	P11,P21,P22,P28,P30	5	10
Gradient Boosting Machine (GBM)	P5,P11,P15,P40,P46	5	10
Linear Regression (LR)	P38,P39	2	4
BERT	P29,P50	2	4
Radial Basis Function Neural Network (RBFNN)	P3,P11	2	4
AdaBoost	P9	1	2
GPT3	P50	1	2
Long Short-term Memory (LSTM)	P46	1	2
Reinforcement Learning	P37	1	2
Bagging	P1	1	2
Convolutional	P15	1	2

Neural Network (CNN)			
RIPPER or JRIP	P18	1	2

Table 3. Distribution by ML Approach

Distribution of Papers by Analytics Outcome. Table 4 shows the distribution of selected papers by objectives/outcomes. Learning performance analysis/prediction is the most popular, being used in 27 studies (52%), followed by learning behavior/style detection and analysis, used in 17 studies (33%), learning path and recommendation (29% or 15 studies), then course delivery and instructional design (19% or 10 studies), and student enrollment, retention or dropout rate prediction, used in 13% of studies.

Analytics Outcome	Articles	#	%
Learning performance analysis/prediction	P1,P6,P7,P9,P10,P12,P13,P15,P17,P18,P20,P21,P22,P24,P25,P26,P27,P28,P30,P31,P32,P33,P34,P39,P43,P45,P46	27	52
Learning behavior/style detection and analysis	P2,P3,P8,P11,P14,P16,P23,P26,P29,P32,P43,P44,P47,P48,P49,P51,P52	17	33
Learning path and recommendation	P3,P9,P13,P14,P16,P21,P24,P29,P30,P35,p37,P38,P44,P48,P50	15	29
Course delivery and instructional design	P5,P14,P19,P20,P25,P35,P38,P42,P44,P50	10	19
Enrollment, retention, and dropout rate prediction/analysis	P4,P10,P17,P36,P40,P41,P46	7	13

Table 4. Distribution by Analytics Objective/Outcome

Distribution of Papers by Evaluation Method. As shown in Table 5, accuracy is the most often used evaluation metric in LMS data analytics with ML approaches (34 out of 52 or 65%). This is followed by the F1-score, used in 27% (14 of 52), then AUC-ROC (21% or 11 papers), then Precision and Recall (19% or ten papers), and RMSE/MSE/MAE metric (13%, 7 out of 52). Table 6 shows the use of analytics outcomes evaluation methods.

Evaluation Metric	Articles	#	%
Accuracy	P1,P3,P6,P7,P10,P11,P12,P13,P15,P18,P19,P20,P21,P23,P24,P26,P27,P28,P29,30,P32,P34,P35,P36,P38,P41,P43,P44,P45,P46,P47,P48,P51,P52	34	65
F1-score	P4,P7,P12,P17,P23,P29,P33,P36,P37,P40,P42,P48,P49,P51	14	27
AUC-ROC	P6,P7,P20,P23,p27,P30,P34,P41,P46,P48,P51	11	21
Precision and Recall	P7,P12,P17,P23,P27,P40,P46,P48,P49,P51	10	19
RMSE/MSE/MAE	P9,P22,P25,P31,P37,P39,P44	7	13

Table 5. Distribution by Evaluation Metric

Analytic Outcomes Evaluation Method	Articles	#	%
Benchmarking	P4,P14,P17,P21,P31,P37	6	12
Collecting learners' feedback	P2,P25,P44,P51	4	8
Specially designed assessment	P19,P48	2	4
Statistical analysis - ANOVA	P34	1	2
Prescriptive analysis	P32	1	2
No analytic outcomes evaluation methods clearly mentioned	P1,P3,P5,P6,P7,P8,P9,P10,P11,P12,P13,P15,P16,P18,P20,P22,P23,P24,P26,P27,P28,P29,P30,P33,P35,P36,P38,P39,P40,P41,P42,P43,P45,P46,P47,P49,P50,P52	38	73

Table 6. Distribution by Analytics Outcomes Evaluation Method

73% of studies do not clearly employ the analytic outcomes evaluation method. Benchmarking is the most used (12%, six papers), followed by collecting learners' feedback (8%, four papers) and designated assessment (4%, two papers). Further,

statistical and prescriptive analyses are used for analytic outcomes evaluation (one paper for each).

Distribution of ML Methods and LMS Data Used for Analytics Outcomes. As shown in Figure 3 and Figure 4, i) the most often used ML methods for learning performance prediction is SVM and Random Forest, and the most often used LMS data is learner demographic data and assessment data; ii) for learning behavior/style detection and analysis, Random Forest, MLP Neural Networks, and SVM are the most popular ML methods and the LMS data is learners' data and activity data; iii) for learning path and recommendation, top three most often used ML methods are SVM, Random Forest, and MLP Neural Network. The LMS data is assessment data, activity data, and course information; iv) for the outcome of course delivery/instructional design, Random Forest, MLPNN, LR, and SVM are commonly used with LMS assessment data, course information, and user behavior log data; v) for student retention/dropout rate prediction, SVM and LR are equally most popular used ML algorithms, and the LMS data is the combination of learner demographic data, assessment data, and activity data.

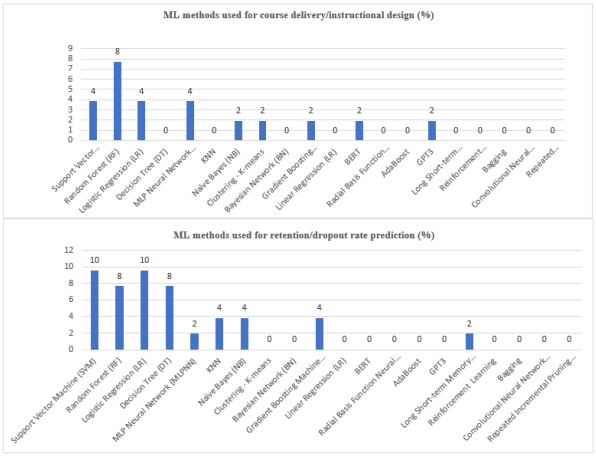

Figure 3. Distribution of ML Approaches Used for Analytics Outcomes

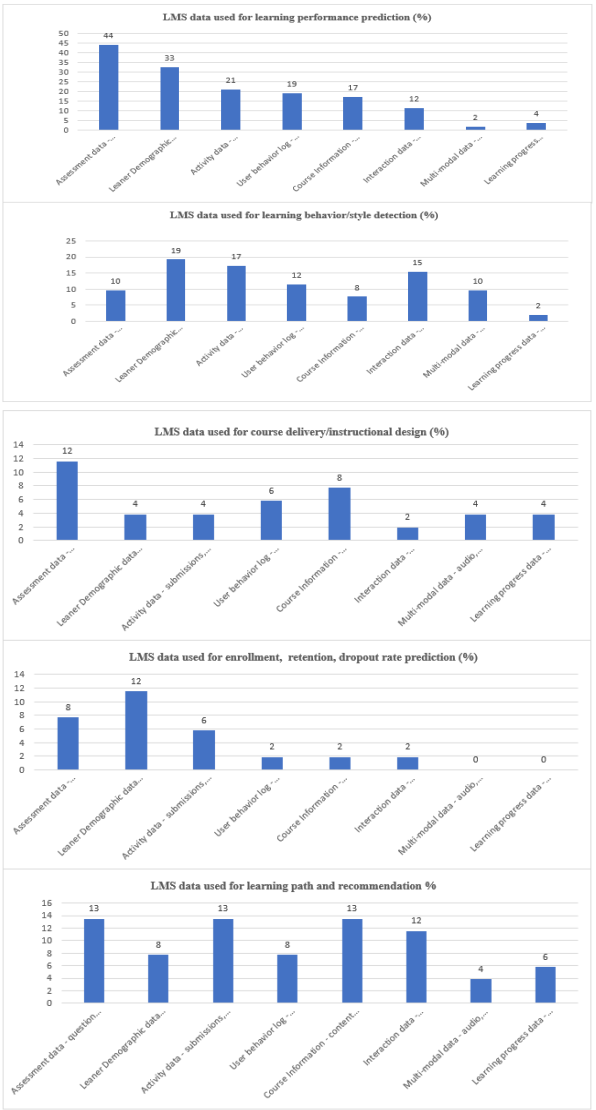

Figure 4. Distribution of LMS Data Used for Analytics Outcomes

Distribution of Evaluation Metrics by ML Methods. As shown in Table 7, accuracy is the most often used evaluation metric for all top six ML- algorithms.

ML Methods	Accuracy	Precision / Recall	F1- Score	AUC-ROC	RMSE / MSE	Cross-validatio
SVM	15	5	5	6	4	5
Random Forest (RF)	18	6	9	7	2	7
Logistic Regression (LR)	8	5	7	5	2	2
Decision Tree (DT)	14	4	4	3	0	2
MLP Neural Network (MLPNN)	13	2	1	4	2	2
KNN	12	4	4	5	2	4

Table 7. Distribution of Evaluation Metrics on Top Six ML Algorithms

In extant studies, researchers used multiple metrics instead of one to assess the effectiveness and efficiency of utilized ML approaches. Notably, AUC-ROC is the second most popular used for SVM, followed by RMSE, Cross-validation, and F1-Score. For random forest, logistic regression, decision tree, neural networks, and KNN methods, F1-score, followed by Precision/Recall, AUC-ROC is the most popular evaluation metric other than Accuracy.

4. DISCUSSION AND FINDINGS

Following the proposed review framework, we summarize the primary challenges and issues in the current literature, as shown in Figure 5. These findings cover the different levels of adaptive learning paths: LMS data (what to learn), machine learning methods (how to learn), analytics outcomes, and evaluation (how well learned). Most extant studies use relatively small datasets to train ML models. Such datasets primarily focus on course-level cross-section numeric data (Du et al., 2020). However, the data generated from the LMS platform nowadays is extensive, multimodal longitudinal data. Machine learning in the context of big data presents unique challenges, and overcoming these obstacles requires approaches that differ from traditional learning methods. Scalable, multidomain, parallel, flexible, and intelligent learning methods are preferred in this context (Qiu et al., 2016).

Secondly, existing literature lacks a comprehensive machine learning (ML) method or a combination of methods designed to achieve specific analytics outcomes (Islam & Mahmud, 2020). The focus has primarily been on using existing ML methods and comparing the performance with a lack of new algorithm development. For example, over 40% of reviewed studies use SVM but rarely combine it with other ML methods or optimize it to achieve better analysis accuracy. Notably, neural networks and some emerging ML methods such as LSTM, BERT, and GPT are gaining more and more attention and need deeper investigation in the domain (Pan et al., 2020; Yang, 2021). Furthermore, developing robust methods for feature selection and assessing their effectiveness is a promising direction in the domain (Coussement et al., 2020; Soleimani & Lee, 2021).

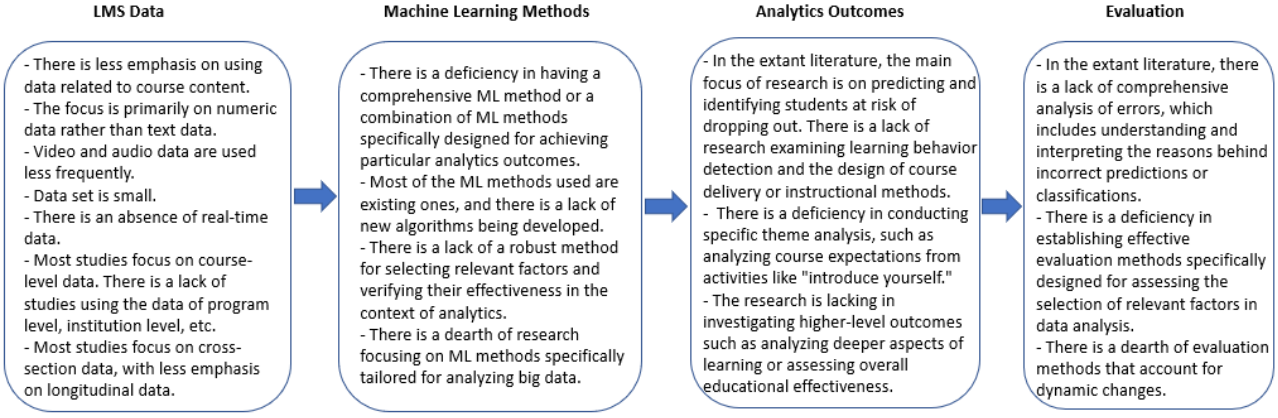

Figure 5. Challenges and Issues in Current Literature

Regarding the analytics objective/outcome, the primary emphasis is on forecasting learners' performance and detecting learners who are likely to discontinue their studies (Villegas-Ch et al., 2020). Less attention has been given to investigating the identification of learning behaviors and developing instructional techniques for course delivery. There is a shortage of research regarding the text analysis of specific themes, such as evaluating course expectations through activities like "introduce yourself." Additionally, the current body of research is limited regarding exploring various outcomes, including examining more profound facets of learning and assessing overall educational effectiveness (Yang, 2021).

Concerning evaluation metrics, extant literature lacks a comprehensive examination of errors, including a thorough understanding and interpretation of the underlying causes for inaccurate predictions or classifications. Moreover, there is a gap in defining evaluation methods that effectively assess the selection of pertinent features in data analysis (Lan et al., 2014). Further, there is a scarcity of evaluation methods that adequately consider dynamic changes (Yang, 2021).

Based on the abovementioned challenges and issues, there are several important directions for future research in this domain. First, student performance prediction remains a viable research topic in the domain (Jiao et al., 2022; Riestra-González et al., 2021). Even though many studies have been done on the utilization of various ML methods in student performance prediction, there remains a need for robust predictive models with good generalizability for different courses, programs, and institutions.

A general predictive model can be developed using existing machine learning algorithms, or combining multiple algorithms, or newly created algorithms. For example, recent advancements in text mining with ML methods have prompted researchers to utilize social media to predict learning performance (Shahbazi & Byun, 2020). However, in the field of LMS data analytics, studies need to effectively integrate online behaviors with textual data to enhance prediction accuracy. Therefore, it is crucial to combine online behaviors with textual data to improve the outcome of performance prediction.

Second, machine learning methods can be used to evaluate the effectiveness of instructional interventions, including course content delivery and instructional design. Due to the lack of an educational framework, no consistent results can be extracted from the studies of instructional content and design (Lee, 2021; Tran et al., 2022). Well-designed instructional content has a significant impact on enhancing learning effectiveness. Consequently, it is anticipated that more researchers will focus on identifying content design patterns in the future. However, current studies have not emphasized automated support for teachers and learners to improve their teaching and learning experiences, such as offering automatic suggestions for instructional design or adjustments to learning strategies (Du et al., 2020).

Third, newer LMS platforms include many user interaction features, such as discussion boards, announcement portals, conversation tools, collaboration tools, etc., aimed at improving student engagement and teaching performance. These advanced tools generate large volumes of text, video, and audio data. Extant literature is limited in regard to the use of ML methods with text or other types of data (Shahbazi & Byun, 2020). One of the important directions for future studies is to investigate the utilization of appropriate ML methods for LMS multimodal data analytics. For example, one study examined students' motivation and predicted their learning performance using video-viewing data in a flipped statistic course (Liao & Wu, 2023).

Fourth, the application of big data ML approaches in detecting learning styles and behavior. LMS platforms have been used for more than ten years. An individual institution possesses a large volume of longitudinal learners' activity and log data. Also, this big data can include various types and formats. An attractive direction for future research is how to gain valuable insights into learners' behavior patterns by using ML methods in this big data analytics context. For example, how many learning patterns are needed to train a classifier depends on balancing cost and accuracy when dealing with overfitting issues (Bird et al., 2022).

5. CONCLUSIONS

This study identified, organized, and discussed challenges and issues related to the application of ML in analyzing data from LMS into four perspectives according to the proposed literature analysis framework. Findings indicate that extant research often uses small datasets and

focuses on numeric data, while LMS platforms generate extensive multimodal longitudinal data. Future research directions include student performance prediction, instructional intervention analysis, multimodal data analytics, and big data ML approaches for learning style and behavior detection.

Overall, the application of ML in LMS data analytics has significant potential to improve teaching and learning outcomes. Institutions can implement adaptive learning platforms that adjust content delivery based on student data collected from the LMS. ML models can dynamically suggest content adjustments (e.g., additional resources for struggling students or advanced materials for high-performing students). The proposed research agenda focuses on a range of research questions and machine-learning methods that can be used to advance the field. By addressing these questions, researchers can develop a deeper understanding of the utilization of machine learning approaches in learning analytics and the development of advanced solutions. While there is a natural inclination to rely on existing methods in the field, it is essential to pursue a parallel line of research that develops new methods and systems.

6. REFERENCES

Ahmed, M. R., Tahid, S. T. I., Mitu, N. A., Kundu, P., & Yeasmin, S. (2020, 7/2020). A Comprehensive Analysis on Undergraduate Student Academic Performance using Feature Selection Techniques on Classification Algorithms. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT),

Anderson, K. (2022, November 7, 2022). Real-time Feedback for Developing Conversation Literacy.*ICMI '22*

Babić, I. Đ. (2017). Machine learning methods in predicting the student academic motivation. *Croatian Operational Research Review*, *8*(2), 443-461. https://doi.org/10.17535/crorr.2017.0028

Bird, K. A., Castleman, B. L., Song, Y., & Yu, R. (2022). Is big data better? LMS data and predictive analytic performance in postsecondary education. *EdWorking Paper*. https://doi.org/10.26300/8XYS-YM74

Bognár, L., & Fauszt, T. (2022). Factors and conditions that affect the goodness of machine learning models for predicting the success of learning. *Computers and Education: Artificial Intelligence*, *3*, 100100. https://doi.org/10.1016/j.caeai.2022.100100

Castro, R. (2019). Blended learning in higher education: Trends and capabilities. *Education and Information Technologies*, *24*(4), 2523-2546. https://doi.org/10.1007/s10639-019-09886-3

Cavanagh, T., Chen, B., Lahcen, R. A. M., & Paradiso, J. R. (2020). Constructing a design framework and pedagogical approach for adaptive learning in higher education: A practitioner's perspective. *International review of research in open and distributed learning*, *21*(1), 173-197. https://doi.org/10.19173/irrodl.v21i1.4557

Chen, L., Leong, C. W., Feng, G., & Lee, C. M. (2014, November 12, 2014). Using Multimodal Cues to Analyze MLA'14 Oral Presentation Quality Corpus: Presentation Delivery and Slides Quality.*MLA '14*

Coussement, K., Phan, M., De Caigny, A., Benoit, D. F., & Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decision Support Systems*, *135*, 113325. https://doi.org/10.1016/j.dss.2020.113325

Deeva, G., De Smedt, J., Saint-Pierre, C., Weber, R., & De Weerdt, J. (2022). Predicting student performance using sequence classification with time-based windows. *Expert Systems with Applications*, *209*, 118182. https://doi.org/10.1016/j.eswa.2022.118182

Dervenis, C., Kyriatzis, V., Stoufis, S., & Fitsilis, P. (2023, January 30, 2023). Predicting Students' Performance Using Machine Learning Algorithms.*ICACS '22*

Du, X., Yang, J., Jui-Long, H., & Shelton, B. (2020). Educational data mining: A systematic review of research and emerging trends. *Information Discovery and Delivery*, *48*(4), 225–236. https://doi.org/10.1108/IDD-09-2019-0070

Elfeky, A. I. M., & Elbyaly, M. Y. H. (2021). The use of data analytics technique in learning management system to develop fashion design skills and technology acceptance. *Interactive Learning Environments*, *31*(6), 3810–3827. https://doi.org/10.1080/10494820.2021.1943688

Elliott, R., & Luo, X. (2022, 2022-10-8). Learning Management System Analytics to Examine the Behavior of Students in High Enrollment STEM Courses During the Transition to Online Instruction. 2022 IEEE Frontiers in Education Conference (FIE),

Gasevic, D., Rose, C., Siemens, G., Wolff, A., & Zdrahal, Z. (2014). Learning analytics and machine learning. *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, 287–288. https://doi.org/10.1145/2567574.256763

Gkontzis, A. F., Kotsiantis, S., Tsoni, R., & Verykios, V. S. (2018, November 29, 2018). An effective LA approach to predict student achievement.*PCI '18*

Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education*, *131*, 22-32. https://doi.org/10.1016/j.compedu.2018.12.006

Imhof, C., Comsa, I.-S., Hlosta, M., Parsaeifard, B., Moser, I., & Bergamin, P. (2022). Prediction of Dilatory Behavior in eLearning: A Comparison of Multiple Machine Learning Models. *IEEE Transactions on Learning Technologies*, 1-15. https://doi.org/10.1109/TLT.2022.3221495

Islam, S., & Mahmud, H. (2020, June 18, 2020). Integration of Learning Analytics into Learner Management System using Machine Learning.*ICMET '20*

Jiao, P., Ouyang, F., Zhang, Q., & Alavi, A. H. (2022). Artificial intelligence-enabled prediction model of student academic performance in online engineering education. *Artificial Intelligence Review*, *55*(8), 6321-6344. https://doi.org/10.1007/s10462-022-10155-y

Joseph, B., & Abraham, S. (2022, 2022-2-12). Analyzing the Cognitive Process Dimension and Rate of Learning to Identify the Slow Learners in e-Learning. 2022 International Conference on Innovative Trends in Information Technology (ICITIIT),

Jui-Long, H., Rice, K., Kepka, J., & Yang, J. (2020). Improving predictive power through deep learning analysis of K-12 online student behaviors and discussion board content. *Information Discovery and Delivery*, *48*(4), 199-212. https://doi.org/10.1108/IDD-02-2020-0019

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, *2*, 100017. https://doi.org/10.1016/j.caeai.2021.100017

Kokoç, M., Akçapınar, G., & Hasnine, M. N. (2021). Unfolding Students' Online Assignment Submission Behavioral Patterns using Temporal Learning Analytics. *Journal of Educational Technology & Society*, *24*(1).

Kondo, N., Okubo, M., & Hatanaka, T. (2017, 7/2017). Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data. 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI),

Korkmaz, C., & Correia, A.-P. (2019). A review of research on machine learning in educational technology. *Educational Media International*, *56*(3), 250–267. https://doi.org/10.1080/09523987.2019.1669875

Kumar, M., Sharma, C., Sharma, S., Nidhi, N., & Islam, N. (2022, 2022-3-23). Analysis of Feature Selection and Data Mining Techniques to Predict Student Academic Performance. 2022 International Conference on Decision Aid Sciences and Applications (DASA),

Lagman, A. C., Alcober, G. M. I., Fernando, M. C. G., Goh, M. L. I., Lalata, J.-a. P., Ortega, J. H. J. C., . . . Claour, J. P. (2020, 2020-06-14). Integration of Neural Network Algorithm in Adaptive Learning Management System. ICRSA 2020: 2020

3rd International Conference on Robot Systems and Applications,

Lamb, R., Neumann, K., & Linder, K. A. (2022). Real-time prediction of science student learning outcomes using machine learning classification of hemodynamics during virtual reality and online learning sessions. *Computers and Education: Artificial Intelligence*, *3*, 100078. https://doi.org/10.1016/j.caeai.2022.100078

Lan, A. S., Waters, A. E., Studer, C., & Baraniuk, R. G. (2014). Sparse factor analysis for learning and content analytics. *The Journal of Machine Learning Research*, *15*(1), 1959–2008.

Le, M.-D., Nguyen, H.-H., Nguyen, D.-L., & Nguyen, V. A. (2020, July 13, 2020). How to Forecast the Students' Learning Outcomes Based on Factors of Interactive Activities in a Blended Learning Course.*ICFET '20*

Lee, A. V. Y. (2021). Determining quality and distribution of ideas in online classroom talk using learning analytics and machine learning. *Journal of Educational Technology & Society*, *24*(1), 236-249. https://www.proquest.com/pqrl/docview/2515025065/abstract/D46CC4C0E1E44CCDPQ/24

Li, F., He, Y., & Xue, Q. (2021). Progress, challenges and countermeasures of adaptive learning. *Educational Technology & Society*, *24*(3), 238-255. https://www.jstor.org/stable/27032868

Liao, C.-H., & Wu, J.-Y. (2023). Learning analytics on video-viewing engagement in a flipped statistics course: Relating external video-viewing patterns to internal motivational dynamics and performance. *Computers & Education*, *197*, 104754. https://doi.org/10.1016/j.compedu.2023.104754

Lwande, C., Link to external site, t. l. w. o. i. a. n. w., Oboko, R., & Lawrence, M. (2021). Learner behavior prediction in a learning management system. *Education and Information Technologies*, *26*(3), 2743-2766. https://doi.org/10.1007/s10639-020-10370-6

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588-599. https://doi.org/10.1016/j.compedu.2009.09.008

Nguyen, V. A., Nguyen, Q. B., & Nguyen, V. T. (2018, August 13, 2018). A Model to Forecast Learning Outcomes for Students in Blended Learning Courses Based On Learning Analytics.*ICSET 2018*

Nizam Ismail, S., Hamid, S., & Chiroma, H. (2019). The utilization of learning analytics to develop student engagement model in learning management system. *Journal of Physics: Conference Series*, *1339*(1), 012096. https://doi.org/10.1088/1742-6596/1339/1/012096

Omiros, I., Link to external site, t. l. w. o. i. a. n. w., Savvas, I. K., Panos, F., & Gerogiannis, V. C. (2021). A two-phase machine learning approach for predicting student outcomes. *Education and Information Technologies*, *26*(1), 69-88. https://doi.org/10.1007/s10639-020-10260-x

Oreski, D., & Hajdin, G. (2019, 5/2019). A Comparative Study of Machine Learning Approaches on Learning Management System Data. 2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO),

Pan, Z., Li, C., & Liu, M. (2020, 2020-08-12). Learning Analytics Dashboard for Problem-based Learning. L@S '20: Seventh (2020) ACM Conference on Learning @ Scale,

Peng, H., Ma, S., & Spector, J. M. (2019). Personalized adaptive learning: an emerging pedagogical approach enabled by a smart learning environment. *Smart Learning Environments*, *6*, 9 (2019). https://doi.org/10.1186/s40561-019-0089-y

Pérez Sánchez, C. J., Calle-Alonso, F., & Vega-Rodríguez, M. A. (2022). Learning analytics to predict students' performance: A case study of a neurodidactics-based collaborative learning platform. *Education and Information Technologies*, *27*(9), 12913-12938.

https://doi.org/10.1007/s10639-022-11128-y

Pimentel, J. S., Ospina, R., Link to external site, t. l. w. o. i. a. n. w., Anderson, A., & Link to external site, t. l. w. o. i. a. n. w. (2021). Learning Time Acceleration in Support Vector Regression: A Case Study in Educational Data Mining. *Stats*, *4*(3), 682. https://doi.org/10.3390/stats4030041

Purwoningsih, T., Santoso, H. B., & Hasibuan, Z. A. (2020, 2020-11-3). Data Analytics of Students' Profiles and Activities in a Full Online Learning Context. 2020 Fifth International Conference on Informatics and Computing (ICIC),

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, *2016*(7), 67. https://doi.org/10.1186/s13634-016-0355-x

Ramaswami, G. S., Susnjak, T., Mathrani, A., & Umer, R. (2020, 2020-12-16). Predicting Students Final Academic Performance using Feature Selection Approaches. 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE),

Riestra-González, M., Paule-Ruíz, M. D. P., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers & Education*, *163*, 104108. https://doi.org/10.1016/j.compedu.2020.104108

Segarra-Faggioni, V., & Ratte, S. (2021, March 6, 2021). Computer-based Classification of Student's Report.*ICETC '20*

Sghir, N., Adadi, A., El Mouden, Z. A., & Lahmer, M. (2022, 2022-3-3). Using Learning Analytics to Improve Students' Enrollments in Higher Education. 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET),

Shahbazi, Z., & Byun, Y. C. (2020). Toward Social Media Content Recommendation Integrated with Data Science and Machine Learning Approach for E-Learners. *Symmetry*, *12*(11), 1798. https://doi.org/10.3390/sym12111798

Soleimani, F., & Lee, J. (2021, June 8, 2021). Comparative Analysis of the Feature Extraction Approaches for Predicting Learners Progress in Online Courses: MicroMasters Credential versus Traditional MOOCs.*L@S '21*

Suleiman, R., & Anane, R. (2022, 2022-5-4). Institutional Data Analysis and Machine Learning Prediction of Student Performance. 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD),

Tamada, M. M., Giusti, R., & De Magalhaes Netto, J. F. (2021, 2021-10-13). Predicting Student Performance Based on Logs in Moodle LMS. 2021 IEEE Frontiers in Education Conference (FIE),

Tenzin, D., Lemay, D. J., Basnet, R. B., & Bazelais, P. (2020). Predictive analytics in education: a comparison of deep learning frameworks. *Education and Information Technologies*, *25*(3), 1951-1963. https://doi.org/10.1007/s10639-019-10068-4

Tran, T. P., Jan, T., & Kew, S. N. (2022). Learning Analytics for Improved Course Delivery: Applications and Techniques. *Proceedings of the 6th International Conference on Digital Technology in Education*, 100–106. https://doi.org/10.1145/3568739.3568758

Umer, R., Mathrani, A., Susnjak, T., & Lim, S. (2019, March 30, 2019). Mining Activity Log Data to Predict Student's Outcome in a Course.*ICBDE '19*

Veluri, R. K., Patra, I., Naved, M., Prasad, V. V., Arcinas, M. M., Beram, S. M., & Raghuvanshi, A. (2022). Learning analytics using deep learning techniques for efficiently managing educational institutes. *Materials Today: Proceedings*, *51*, 2317-2320. https://doi.org/10.1016/j.matpr.2021.11.416

Villegas-Ch, W., Román-Cañizares, M., & Palacios-Pacheco, X. (2020). Improvement of an Online Education Model with the Integration of Machine Learning and Data Analysis in an LMS. *Applied Sciences*, *10*(15), 5371. https://doi.org/10.3390/app10155371

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*, 106189. https://doi.org/10.1016/j.chb.2019.106189

Waheed, H., Hassan, S.-U., Nawaz, R., Aljohani, N. R., Chen, G., & Gasevic, D. (2023). Early prediction of learners at risk in self-paced education: A neural network approach. *Expert Systems with Applications*, *213*, 118868. https://doi.org/10.1016/j.eswa.2022.118868

Worsley, M. (2018, March 7, 2018). (Dis)engagement matters: identifying efficacious learning practices with multimodal learning analytics.*LAK '18*

Wu, J.-Y. (2021). Learning analytics on structured and unstructured heterogeneous data sources: Perspectives from procrastination, help-seeking, and machine-learning defined cognitive engagement. *Computers & Education*, *163*, 104066. https://doi.org/10.1016/j.compedu.2020.104066

Xie, H., Chu, H.-C., Hwang, G.-J., & Wang, C.-C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Computers & Education*, *140*, 103599. https://doi.org/10.1016/j.compedu.2019.103599

Xing, W., & Goggins, S. (2015, March 16, 2015). Learning analytics in outer space: a Hidden Naïve Bayes model for automatic student off-task behavior detection.*LAK '15*

Yang, S. J. H. (2021). Precision Education - A New Challenge for AI in Education. *Journal of Educational Technology & Society*, *24*(1).

Zhao, F., Liu, G.-Z., Zhou, J., & Yin, C. (2023). A Learning Analytics Framework Based on Human-Centered Artificial Intelligence for Identifying the Optimal Learning Strategy to Intervene in Learning Behavior. *Journal of Educational Technology & Society*, *26*(1).

Zhou, J., Hang, K., Oviatt, S., Yu, K., & Chen, F. (2014, November 12, 2014). Combining empirical and machine learning techniques to predict math expertise using pen signal features.*MLA '14*

Zhu, M., Sari, A. R., & Lee, M. M. (2022). Trends and issues in MOOC learning analytics empirical research: A systematic literature review (2011–2021). *Education and Information Technologies*, *27*(7), 10135-10160. https://doi.org/10.1007/s10639-022-11031-6

APPENDIX A
Articles Coded for Literature Survey

Citation	Paper Title	#
(Ahmed et al., 2020)	A Comprehensive Analysis on Undergraduate Student Academic Performance using Feature Selection Techniques on Classification Algorithms	P1
(Anderson, 2022)	Real-time Feedback for Developing Conversation Literacy	P2
(Babić, 2017)	Machine learning methods in predicting student academic motivation	P3
(Bognár & Fauszt, 2022)	Factors and conditions that affect the goodness of machine learning models for predicting the success of learning	P4
(Chen et al., 2014)	Using Multimodal Cues to Analyze MLA'14 Oral Presentation Quality Corpus: Presentation Delivery and Slides Quality	P5
(Deeva et al., 2022)	Predicting student performance using sequence classification with time-based windows	P6
(Dervenis et al., 2023)	Predicting Students' Performance Using Machine Learning Algorithms	P7
(Elliott & Luo, 2022)	Learning Management System Analytics to Examine the Behavior of Students in High Enrollment STEM Courses During the Transition to Online Instruction	P8
(Gkontzis et al., 2018)	An effective LA approach to predict student achievement	P9
(Gray & Perkins, 2019)	Utilizing early engagement and machine learning to predict student outcomes	P10
(Imhof et al., 2022)	Prediction of Dilatory Behavior in eLearning: A Comparison of Multiple Machine Learning Models	P11
(Islam & Mahmud, 2020)	Integration of Learning Analytics into Learner Management System using Machine Learning	P12
(Jiao et al., 2022)	Artificial intelligence-enabled prediction model of student academic performance in online engineering education	P13
(Joseph & Abraham, 2022)	Analyzing the Cognitive Process Dimension and Rate of Learning to Identify the Slow Learners in e-Learning	P14
(Jui-Long et al., 2020)	Improving predictive power through deep learning analysis of K-12 online student behaviors and discussion board content	P15
(Kokoç et al., 2021)	Unfolding Students' Online Assignment Submission Behavioral Patterns using Temporal Learning Analytics	P16
(Kondo et al., 2017)	Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data	P17
(Kumar et al., 2022)	Analysis of Feature Selection and Data Mining Techniques to Predict Student Academic Performance	P18
(Lagman et al., 2020)	Integration of Neural Network Algorithm in Adaptive Learning Management System	P19
(Lamb et al., 2022)	Real-time prediction of science student learning outcomes using machine learning classification of hemodynamics during virtual reality and online learning sessions	P20
(Lan et al., 2014)	Sparse factor analysis for learning and content analytics	P21
(Le et al., 2020)	How to Forecast the Students' Learning Outcomes Based on Factors of Interactive Activities in a Blended Learning Course	P22
(Lwande et al., 2021)	Learner behavior prediction in a learning management system	P23
(Macfadyen & Dawson, 2010)	Mining LMS data to develop an "early warning system" for educators: A proof of concept	P24
(Nguyen et al., 2018)	A Model to Forecast Learning Outcomes for Students in Blended Learning Courses Based On Learning Analytics	P25

(Nizam Ismail et al., 2019)	The utilization of learning analytics to develop student engagement model in learning management system	P26
(Omiros et al., 2021)	A two-phase machine learning approach for predicting student outcomes	P27
(Oreski & Hajdin, 2019)	A Comparative Study of Machine Learning Approaches on Learning Management System Data	P28
(Pan et al., 2020)	Learning Analytics Dashboard for Problem-based Learning	P29
(Pérez Sánchez et al., 2022)	Learning analytics to predict students' performance: A case study of a neurodidactics-based collaborative learning platform	P30
(Pimentel et al., 2021)	Learning Time Acceleration in Support Vector Regression: A Case Study in Educational Data Mining	P31
(Purwoningsih et al., 2020)	Data Analytics of Students' Profiles and Activities in a Full Online Learning Context	P32
(Ramaswami et al., 2020)	Predicting Students Final Academic Performance using Feature Selection Approaches	P33
(Riestra-González et al., 2021)	Massive LMS log data analysis for the early prediction of course-agnostic student performance	P34
(Segarra-Faggioni & Ratte, 2021)	Computer-based Classification of Student's Report	P35
(Sghir et al., 2022)	Using Learning Analytics to Improve Students' Enrollments in Higher Education	P36
(Shahbazi & Byun, 2020)	Toward Social Media Content Recommendation Integrated with Data Science and Machine Learning Approach for E-Learners	P37
(Soleimani & Lee, 2021)	Comparative Analysis of the Feature Extraction Approaches for Predicting Learners Progress in Online Courses: MicroMasters Credential versus Traditional MOOCs	P38
(Suleiman & Anane, 2022)	Institutional Data Analysis and Machine Learning Prediction of Student Performance	P39
(Tamada et al., 2021)	Predicting Student Performance Based on Logs in Moodle LMS	P40
(Tenzin et al., 2020)	Predictive analytics in education: a comparison of deep learning frameworks	P41
(Umer et al., 2019)	Mining Activity Log Data to Predict Student's Outcome in a Course	P42
(Veluri et al., 2022)	Learning analytics using deep learning techniques for efficiently managing educational institutes	P43
(Villegas-Ch et al., 2020)	Improvement of an Online Education Model with the Integration of Machine Learning and Data Analysis in an LMS	P44
(Waheed et al., 2020)	Predicting academic performance of students from VLE big data using deep learning models	P45
(Waheed et al., 2023)	Early prediction of learners at risk in self-paced education: A neural network approach	P46
(Worsley, 2018)	(Dis)engagement matters: identifying efficacious learning practices with multimodal learning analytics	P47
(Wu, 2021)	Learning analytics on structured and unstructured heterogeneous data sources: Perspectives from procrastination, help-seeking, and machine-learning defined cognitive engagement	P48
(Xing & Goggins, 2015)	Learning analytics in outer space: a Hidden Naïve Bayes model for automatic student off-task behavior detection	P49
(Yang, 2021)	Precision Education - A New Challenge for AI in Education	P50
(Zhao et al., 2023)	A Learning Analytics Framework Based on Human-Centered Artificial Intelligence for Identifying the Optimal Learning Strategy to Intervene in Learning Behavior	P51
(Zhou et al., 2014)	Combining empirical and machine learning techniques to predict math expertise using pen signal features	P52

A Comparison of Large Language Models for Oncology Clinical Text Summarization

Chiazam Izuchukwu
ci02061@georgiasouthern.edu
Georgia Southern University
Atlanta, GA 30302

Hayden Wimmer
hayden.wimmer@gmail.com
Georgia Southern University
Atlanta, GA 30302

Carl Michael Redman Jr.
carlr@sandiego.edu
University of San Diego
San Diego, CA 92110

Abstract

The rapid growth of data in the health sector has made it crucial to communicate essential information quickly and succinctly. The vast amount of textual data from electronic health records tends to overwhelm healthcare professionals which reduces the time they can dedicate to patient care. This massive amount of complex qualitative data causes physicians to struggle with the decision-making process which had traditionally relied on human evaluation. This study addresses the urgent need for effective summarization of health records to improve patient outcomes and clinical decision-making. We highlight the use of large language models (LLMs) to produce concise summaries of patients' medical oncology reports. Specifically, we utilized pre-trained transformer models, including BART, T5, and Pegasus, to summarize patient clinical notes. The performance of these models was evaluated using BLEU, ROUGE, and BERT scores on CORAL expert-curated medical oncology reports that were de-identified using Philter. The results show that the BART and T5 models performed the best, with the generated summaries being shorter than the original oncology reports. This approach reduces information overload and enhances patient care by providing concise and informative summaries.

Keywords: Large Language Model (LLM), Text Summarization, Artificial Intelligence (AI), Coral, Medical, Oncology

Recommended Citation: Izuchukwu, C., Wimmer, H., Rebman Jr., C.M., (2025). A Comparison of Large Language Models for Oncology Clinical Text Summarization. *Journal of Information Systems Applied Research and Analytics* v18, n2 pp 20-29. DOI# https://doi.org/10.62273/IGMU6476

A Comparison of Large Language Models for Oncology Clinical Text Summarization

Chiazam Izuchukwu, Hayden Wimmer and Carl Michael Redman Jr.

1. INTRODUCTION

Supporting physicians and clinicians in the oncology field struggle during the decision-making process and the oncology field has been a subject of extensive research and debate. Medical professionals are frequently overwhelmed with an abundance of data and information. While quantitative data is well-suited for machine learning and statistical methods to aid in decision support, qualitative data is rich with explicit and tacit information. Processing this qualitative data to make it available and useful for decision support is a complex task, traditionally relying on human evaluation through qualitative techniques such as coding and analysis. During a visit with an oncology doctor, much qualitative data is extracted from the patient and added to their electronic health record. This can be seen as a semi-structured interview with both closed and open-ended questions.

Physicians often face time constraints and production pressures, limiting the time they can spend with each patient. Reading and processing the extensive textual data generated during medical visits is an overwhelming task, given these time constraints and the need for high patient turnover. A single medical chart or electronic health record can generate numerous pages of qualitative textual data. Over the course of a patient's stay in a medical facility or through routine visits to medical providers, the volume of data grows significantly. One promising method to assist physicians and doctors in processing this vast amount of text data is the use of Artificial Intelligence (AI), specifically large language models (LLMs).

This paper proposes the use of large language models (LLMs) such as BART, T5, and Pegasus for summarizing medical oncology reports, enhancing clinical decision-making by reducing information overload. BART is highlighted as the most effective model, consistently outperforming others across various metrics including ROUGE and BERTScore, despite similar BLEU scores among the models. The study underscores the potential of LLMs to support physicians by efficiently processing extensive qualitative data in electronic health records, thereby improving patient care and decision-making timelines. The remainder of this paper is organized as follows: next we present a review of some relevant literature useful in our work, followed by our methodology where we illustrate the LLMs and evaluation metrics. We then advance to our results which include standard evaluation metrics and a brief statistical analysis, then conclude with our discussion and future works.

2. LITERATURE REVIEW

Text summarization offers a significant advantage over manual summarization by condensing large data into meaningful summaries while preserving content. It can be classified into extractive summarization, which uses statistical and linguistic features to highlight important parts, and abstractive summarization, which generates summaries by understanding the entire document. One area that can benefit from LLM and text summarization methods is clinical text and articles in healthcare.

According to Allahyari et al. (2017), the increasing availability of documents has spurred extensive research in automatic text summarization, which aims to create concise and fluent summaries that preserve key information and overall meaning. Automatic text summarization is challenging because humans summarize text by fully understanding it first, a capability that computers lack. There are two main approaches to summarization: extractive, which selects and reproduces key sections verbatim, and abstractive, which generates new text conveying the essential information. Despite the naturalness of human-created summaries, research has predominantly focused on extractive methods, which often produce better results due to the complexities involved in semantic representation and language generation inherent in abstractive summarization. (Allahyari et al., 2017).

Batra, Chaudhary, Bhatt, Varshney, and Verma (2020) felt that an overwhelming amount of articles and links that people have to choose from and as this data grows, the importance of

semantic density does as well. They make the claim that more concise, meaningful communication is needed, and that text summarization might be a solution. Text summarization addresses this by condensing lengthy texts into short, informative sentences. Machine learning models can play a crucial role in this process by first understanding the document and then producing a summary. Their paper analyzed five different models in the literature from the years 2013-2019 to and present the argument that these models can provide a good summarization of large amounts of data.
Bhatia and Jaiswal (2015) noted how the rapid growth of World Wide Web data has made it increasingly difficult to manually gather and summarize information. Their study investigated trends in text summarization methods. They examined eight different approaches to extractive summarization and eight different approaches to abstractive summarization. Their study concluded that extractive summarizations deal with important sentences while abstractive summarization processes seek understanding of the text and articles and then proceed to build a summary. They also found that automated processes can saves time and efficiently retrieves information from large documents. (Bhatia & Jaiswal, 2015).

Van Veen et al. (2023), conducted a study evaluated methods for adapting large language models (LLM) to summarize clinical text. According to Van Veen et al. (2023), sifting through vast textual data and summarizing key information from electronic health records (EHR) imposes a substantial burden on clinicians' time. They analyzed eight models across a diverse set of summarization tasks including radiology reports and doctor-patient dialogue. Although large language models (LLMs) show promise in natural language processing (NLP) tasks, their efficacy in clinical summarization has not been rigorously demonstrated. They performed a quantitative assessment which revealed trade-offs between models and methods, with some LLM advances not improving results. More notably, their study demonstrated that LLM summaries are often preferred over human expert summaries due to higher scores for completeness, correctness, and conciseness.

Medical care and observational studies in oncology require a thorough understanding of a patient's disease progression and treatment history, often documented within clinical notes. Large language models (LLMs) have demonstrated impressive capabilities, but the standards for clinical applications are exceptionally high. As large language models (LLMs) are becoming more popular, it is essential to evaluate their potential in oncology.

Savova et al. (2019), noted that data produced during the processes of clinical care and research in oncology are proliferating at an exponential rate This prompted them to perform a study that reviewed the advances of natural language processing (NLP) and information extraction methods relevant to oncology based on publications from PubMed as well as NLP and machine learning conference proceedings in the last 3 years. The review highlighted significant advancements in NLP and information extraction that have the potential to improve the fidelity of oncology phenotypes and reduce errors derived from clinical texts. They also notated that summarization and information retrieval applications can reduce search burden and enable clinicians to spend more time with their patients. They surmised that advancements are critical for catalyzing clinical care, research, and regulatory activities by providing more detailed and accurate phenotype information from real-world data (Savova et al., 2019).

Singhal et al., 2023 noted that medicine is an endeavor where language is important for interactions between clinicians, researchers, and patients. They felt that today’s AI models for applications in medicine and healthcare have largely failed to fully utilize language and were mostly effective with single task systems. Current assessments of clinical knowledge in these models often rely on automated evaluations based which may not fully capture the complexity and nuances of clinical reasoning and knowledge.

To address this issue Singhal et al. (2023) created a study and introduced MultiMedQA, a comprehensive benchmark combining six existing medical question-answering datasets, and a new dataset of medical questions searched online, HealthSearchQA. The goal was to evaluate the capabilities of LLMs in the medical domain comprehensively. They used the Pathways Language Model (PaLM), a 540-billion parameter LLM, and its instruction-tuned variant, Flan-PaLM, on MultiMedQA and assessed the models' performance across various datasets, including MedQA, MedMCQA, PubMedQA, and MMLU clinical topics. Their results found that Flan-PaLM achieved state-of-the-art accuracy on all MultiMedQA multiple-choice datasets, including 67.6% accuracy on MedQA, surpassing the prior state-of-the-art by

over 17%. However, human evaluations revealed significant gaps in the models' performance, highlighting areas where the models still fall short. The resulting model, Med-PaLM, showed improvements in comprehension, knowledge recall, and reasoning with increased model scale and instruction prompt tuning (Singhal et al., 2023).

Sushil et al. (2024) objective was to assess the performance of three recent LLMs (GPT-4, GPT-3.5-turbo, and FLAN-UL2) in extracting detailed oncological information from clinical progress notes, using a newly curated, fine-grained, expert-labeled dataset of 40 de-identified breast and pancreatic cancer progress notes. They evaluated the models in zero-shot extraction from two narrative sections of clinical progress notes, using BLEU-4, ROUGE-1, and exact match (EM) F1-score metrics. Their team of oncology fellows and medical students manually annotated 9028 entities, 9986 modifiers, and 5312 relationships to support this evaluation. GPT-4 exhibited the best overall performance with an average BLEU score of 0.73, an average ROUGE score of 0.72, an average EM-F1-score of 0.51, and an accuracy of 68% based on expert manual evaluation. It excelled in extracting tumor characteristics and medications, and in inferring symptoms and future medication considerations. Common errors included partial responses and hallucinations (Sushil et al., 2024).

3. METHODLOGY

In our study, we conducted various analyses and experiments on a unique CORAL reports dataset to assess the performance of various Large Language Models performing abstractive summarization. These datasets serve as the basis for our comparison and analysis. Figure 1 below illustrates the framework of our method.

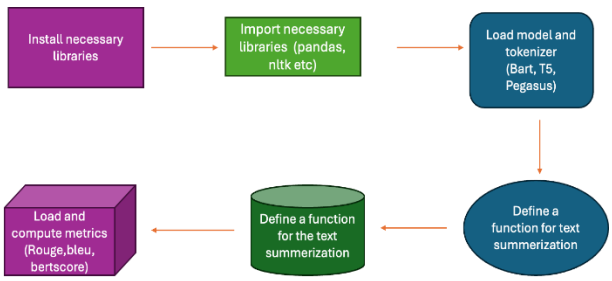

Figure 1 The architecture of oncology text summarization with LLMs

Dataset

The dataset used in this paper is CORAL expert-curated medical oncology reports (2024). The dataset comprises 100 pancreatic cancer notes, including demographic details and corresponding medical oncology notes for patients from the University of California, San Francisco (UCSF) Information Commons. This dataset, containing patient data from 2012 to 2022, has been de-identified using Philter (2023). Pancreatic cancer samples were collected while ensuring a diverse distribution of race/ethnicity. The race/ethnicity groups were either evenly distributed or limited to the maximum counts available in the UCSF dataset, whichever was smaller (Goldberger et al., 2000).

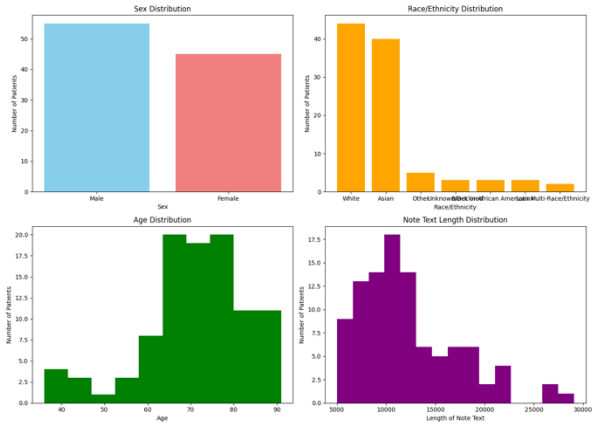

Figure 2 Composition of oncology dataset

Large Language Models

Large Language Models (LLMs) are transformative technologies in natural language processing (NLP). BART, Pegasus, and T5 are sequence-to-sequence models, also known as encoder-decoder models, primarily designed for natural language generation tasks. These models utilize vast datasets and advanced machine-learning techniques to accurately understand and generate human language. The foundational architecture behind LLMs typically involves deep learning techniques, such as transformers, which allow them to process and produce text that closely mimics human linguistic abilities (abstract approach) (2017). LLMs have revolutionized applications such as text generation, translation, and summarization, making interactions with machines more intuitive and seamless. They have become essential tools in various industries, enabling automated customer service, content creation, and data analysis.

BART Model
BART, or Bidirectional and Auto-Regressive Transformers, is a sophisticated LLM developed by Facebook AI. It combines the strengths of BERT's bidirectional encoding with GPT's autoregressive decoding, making it highly effective for a range of NLP tasks including text generation, machine translation, and summarization (2019). BART's architecture allows it to predict corrupted text and fill in missing information, enhancing its performance in generating coherent and contextually relevant text. This hybrid approach enables BART to excel in understanding and generating text, making it a versatile tool for various applications such as dialogue generation and language modelling.

Pegasus Model
Pegasus, created by Google Research, is another LLM specifically designed for abstractive text summarization. Pegasus employs a unique pre-training objective that involves masking entire sentences and then predicting them, which closely mimics the task of summarization (2020). This approach enables Pegasus to generate high-quality summaries that capture the essence of the original content, making it a powerful tool for condensing large volumes of information. Pegasus's ability to produce concise and informative summaries has significant implications for fields such as news aggregation, academic research, and legal document review.

T5 Model
T5, or Text-To-Text Transfer Transformer, is a versatile LLM from Google Research that treats every NLP task as a text-to-text problem (2020). This unified approach simplifies the model architecture and makes T5 applicable to a wide array of tasks, from translation to question answering to summarization. T5's ability to be fine-tuned for specific applications allows it to achieve state-of-the-art performance across various benchmarks. The model's versatility and effectiveness make it an essential tool for researchers and practitioners in NLP, enabling them to tackle a broad spectrum of language-related challenges with a single model framework.

The advancement of LLMs like BART, Pegasus, and T5 highlights the rapid progress in NLP. These models not only improve the accuracy and efficiency of language-related tasks but also pave the way for new applications in fields such as healthcare, education, and content creation. For instance, in healthcare, LLMs can assist in summarizing patient records, generating medical reports, and even supporting diagnostic processes through natural language understanding (2021). In education, these models can provide personalized tutoring, automated grading, and language translation services, enhancing the learning experience for students worldwide.

As LLM technology continues to evolve, it promises to further bridge the gap between human and machine communication, enabling more natural and productive interactions. The development and deployment of LLMs are expected to bring about significant changes in how we interact with technology, making it more accessible and efficient. However, the growth of LLMs also raises important ethical and societal questions, such as issues of bias, privacy, and the potential for misuse. Researchers and developers must address these challenges to ensure that the benefits of LLM technology are realized responsibly and equitably (2021).

4. RESULTS

Analysis and Evaluation Metrics
ANOVA tests and Tukey's Honest Significant Difference (HSD) test were used in the statistical analyses to ascertain the significance of the group differences. The following section analyzes the evaluation metrics employed in summarizing the clinical notes. These metrics are employed to measure the quality and effectiveness of the generated summaries, utilizing a range of well-known and widely accepted evaluation standards for various large language models (LLMs).

Rouge Score
According to Lin (2004) Rouge (Recall-Oriented Understudy for Gisting Evaluation) score is a set of metrics used to evaluate the quality of summaries by comparing them to reference summaries. ROUGE metrics are commonly used in natural language processing tasks to measure the similarity between a generated summary and a reference summary. Here are some of the most frequently used ROUGE metrics and their formulas:

ROUGE-N is a recall-based measure that calculates the overlap of n-grams between the generated summary and the reference summary (Lin, 2004).

$$ROUGE - N = \frac{\sum (Reference\ summaries) \sum_{gram_n} Count_{match}(gram_n)}{\sum (Reference\ summaries) \sum_{gram_n} Count(gram_n)} \#(1)$$

Where:

- $Count_{match}(gram_n)$ is the number of n-grams in the reference summary that match an n-gram in the generated summary.
- $Count(gram_n)$ is the total number of n-grams in the reference summary.
- ROUGE-1: Measures the overlap of unigrams (1-grams) between the generated summary and the reference summary.
- ROUGE-2: Measures the overlap of bigrams (2-grams) between the generated summary and the reference summary.
- ROUGE-L measures the longest common subsequence (LCS) between the generated summary and the reference summary.

$$ROUGE - L = \frac{LCS(\chi, \gamma)}{Lenght(\gamma)} \#(2)$$

Where:

- $LCS(\chi, \gamma)$ is the length of the longest common subsequence between sequences X and Y.
- $Length(\gamma)$ is the length of the reference summary.

Rouge1			
	Pegasus	**Bart**	**T5**
Mean	0.012752	0.0638	0.047
Median	0.01018	0.059105	0.04091

Table 1 Rouge 1 scores

Rouge2			
	Pegasus	**Bart**	**T5**
Mean	0.00669	0.05817	0.03959
Median	0.00391	0.05415	0.03509

Table 2 Rouge 2 scores

RougeL			
	Pegasus	**Bart**	**T5**
Mean	0.01079	0.06140	0.04502
Median	0.00912	0.05726	0.039560

Table 3 Rouge L scores

ROUGE metrics show that scores are typically between 0 and 1. Better translation quality is indicated by higher ROUGE scores, which show greater overlap between the generated summary and the reference summary. Smaller ROUGE scores indicate poorer translation quality since they suggest less precision or accuracy in the model's output when compared to the reference summary.

Bleu Score

According to Papineni, Roukos, Ward, and Zhu (2002), the BLEU (Bilingual Evaluation Understudy) score is a metric used to evaluate the quality of text which has been machine-translated from one natural language to another. The BLEU score compares the n-grams of the candidate translation with the n-grams of the reference translations and counts the number of matches. These matches are then used to calculate precision for the candidate translation. The BLEU score is calculated as follows:

Modified Precision for n-grams:

$$P_i = \frac{Count\ Clip(matches_i, max - ref - count_i)}{candidate - n - grams_i} \#(3)$$

Where:

- Count Clips is a function that clips the number of matched n-grams ($matches_i$) by the maximum count of the n-gram across all reference translations ($max - ref - count_i$).
- $matches_i$ is the number of n-grams of order i that match **exactly** between the candidate translation and any of the reference translations.
- $max - ref - count_i$ the maximum number of occurrences of the specific n-gram of order i found in any single reference translation.
- $candidate - n - grams_i$ is the total number of n-grams of order i present in the candidate translation.

Brevity Penalty (BP):

$$BP = exp\ (1 - \frac{r}{C}) \#(4)$$

Where:

- c is the average length of the reference translations.
- r is the length of the candidate translation

Geometric Mean of Precision Scores:

$$BLEU\ Score = BP * exp\left(\sum_{i=1}^{N} (w_i * ln\ ln\ (p_i))\right) \#(5)$$

Where:

- BP stands for Brevity Penalty
- w_i is the weight for n-gram precision of order i (typically weights are equal for all i)
- p_i is the n-gram modified precision score of order i.
- N is the maximum n-gram order to consider (usually up to 4)

Bleu			
	Pegasus	**Bart**	**T5**
Mean	1.34E-11	3.02E-07	1.92E-09
Median	2.99E-80	1.09E-15	7.71E-22

Table 4 Bleu scores

A BLEU score falls between 0 and 1. Better translation quality is indicated by higher BLEU scores, which show greater overlap between the generated summary and the reference summary. Smaller BLEU scores indicate poorer translation quality since they suggest less precision or accuracy in the model's output when compared to the reference summary.

BERT Score

BERT Score is a metric for evaluating text generation quality based on BERT embeddings. It calculates the similarity between the reference and generated text at the token level using contextual embeddings from a pre-trained BERT model (T. Zhang, Kishore, Wu, Weinberger, & Artzi, 2019).

BERTScore considers precision, recall, and F1 scores based on token similarity. Lee and Toutanova (2018) provides the formula which computes the cosine similarity between the generated and reference text.

- **Token Embeddings**: Compute the contextual embeddings for each token in the reference R and candidate C texts using BERT:

$$E_R = BERT(R), \quad E_C = BERT(C) \#(6)$$

- **Cosine Similarity**: Calculate the cosine similarity between all pairs of tokens from the reference and candidate texts:

$$S_{ij} = \frac{E_R[i] \cdot E_C[j]}{\| E_R[i] \| \| E_C[j] \|} \#(7)$$

where $E_R[i]$ and $E_c[j]$are the embeddings of the i-th and j-th tokens in the reference and candidate texts, respectively.

☐ **Precision:** For each token in the candidate text, find the most similar token in the reference text:

$$P = \frac{1}{|\ C\ |} \sum_{j} maxS_{ij} \#(8)$$

where |C| is the number of tokens in the candidate text.

☐ **Recall**: For each token in the reference text, find the most similar token in the candidate text:

$$R = \frac{1}{|\ R\ |} \sum_{j} maxS_{ij} \#(9)$$

where |R| is the number of tokens in the reference text.

☐ **F1 Score:** Combine precision and recall into an F1 score:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \#(10)$$

Bert			
	Pegasus	**Bart**	**T5**
Mean	0.795566	0.858640	0.848749
Median	0.803880	0.863125	0.850725

Table 5 Bert scores

Better translation quality is indicated by higher BERT scores, which show greater overlap between the generated summary and the reference summary. Smaller BERT scores indicate poorer translation quality since they suggest less precision or accuracy in the model's output when compared to the reference summary.

	Pegasus				
	Rouge1	Rouge2	RougeL	BERT	BLEU
Mean	0.012752	0.006693	0.010794	0.079557	1.34E-11
Median	0.01018	0.003915	0.00912	0.803998	2.99E-80
	BART				
	Rouge1	Rouge2	RougeL	BERT	BLEU
Mean	0.0638	0.058179	0.061409	0.85864	3.02E-07
Median	0.059105	0.05415	0.05726	0.863125	1.09E-15
	T5				
	Rouge1	Rouge2	RougeL	BERT	BLEU
Mean	0.047	0.039595	0.045025	0.848749	1.92E-09
Median	0.0491	0.03509	0.03956	0.850725	7.71E-22

Table 6 Comparison Table

Statistical Analysis

The following tables present the results of these statistical tests and shed light on the importance of the variations between the text summarization models. The results for the difference between groups from the ANOVA test is tabulated and summarized in Table 7.

ANOVA					
Metric	Sum of Squares	df	Mean Square	F	Sig.
Rouge1	0.134	2	0.067	125.983	0.0000
Rouge2	0.136	2	0.068	146.865	0.0000
RougeL	0.133	2	0.067	130.086	0.0000
Bert	0.23	2	0.115	236.834	0.0000
Bleu	0	2	0	1.908	0.1500

Table 7 ANOVA Results

A post-hoc analysis of the variance across groups has been done using the Tukey's Honest Significant Difference (HSD) Test. This test compares each pair of groups and provides the mean difference, standard error, significance level, and confidence interval. There were significant differences between all groups for rouge1, rouge2, rougeL, and bert; however, no statistical difference was found among groups based on the bleu score. The reason may be that the bleu scores were so small that statistical significance was not able to be reached.

5. DISCUSSION AND CONCLUSION

Overall, BART consistently outperforms Pegasus and T5 across all ROUGE metrics and BERTScore, indicating superior performance in capturing both content and semantic similarity in the summaries. While the BLEU scores are low for all models, BART still leads, suggesting a slight edge in n-gram precision. These results highlight the effectiveness of BART in summarizing oncology reports, with T5 performing moderately well and Pegasus lagging behind. The statistical analysis using ANOVA and Tukey's HSD provided further insight into the significance of these differences. The ANOVA results indicated significant differences between the three models, and the Tukey HSD test results show that:

- BART consistently outperforms Pegasus and T5 across ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore metrics.
- T5 also significantly outperforms Pegasus across these metrics.
- There are no significant differences in the BLEU scores between any of the models, indicating that all three models perform similarly in terms of n-gram precision.

These results align with the ANOVA findings and reinforce the conclusion that BART is the best-performing model in terms of ROUGE and BERTScore metrics, while the BLEU scores do not show significant differences between the models.

Supporting physicians and clinicians during the decision-making process is vital due to the astounding amount of both quantitative and qualitative data they encounter. While traditional methods rely heavily on human evaluation, the use of Artificial Intelligence (AI), particularly large language models (LLMs), offers a promising solution. LLMs can assist in processing the extensive qualitative data found in electronic health records, thus alleviating time constraints and enhancing the efficiency of medical professionals. This study represents a significant first step towards integrating LLMs in the processing of clinical notes, aiming to improve the overall decision-making timeline in medical practice.

6. FUTURE WORK

Future projects in this area will include a thorough validation to ensure accuracy and reliability of LLMs across oncology data and seamless integration into clinical workflows to complement existing architecture, systems and practices without disruption. Interoperability with various Electronic Health Record (EHR) systems is vital, requiring standardized interfaces for efficient data access. IT governance would be introduced to address ethical concerns and challenges that will arise from the use of AI to ensure client privacy is preserved. Exploring LLMs for real-time decision

support in clinical settings could revolutionize patient care by providing instant insights. AI/IT training would be introduced to aid in a faster and widespread adoption.

Continuous learning and improvement algorithms should be developed to keep LLMs updated with the latest oncology data and information.

7. References

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268. https://doi.org/10.48550/arXiv.1707.02268

Batra, P., Chaudhary, S., Bhatt, K., Varshney, S., & Verma, S. (2020). A review: Abstractive text summarization techniques using NLP. Paper presented at the 2020 International Conference on Advances in Computing, Communication & Materials (ICACCM). https://doi.org/10.1109/ICACCM50413.2020.9213079

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. Paper presented at the Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. https://doi.org/10.1145/3442188.3445922

Bhatia, N., & Jaiswal, A. (2015). Trends in extractive and abstractive techniques in text summarization. International Journal of Computer Applications, 117(6).

Chen, T., Allauzen, C., Huang, Y., Park, D., Rybach, D., Huang, W. R., . . . Moreno, P. J. (2023). Large-scale language model rescoring on long-form data. Paper presented at the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI: 10.1109/ICASSP49357.2023.10096429

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Socher, R. (2021). Deep learning-enabled medical computer vision. NPJ digital medicine, 4(1), 5. DOI: https://doi.org/10.1038/s41746-020-00376-2

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation, 101(23), e215-e220. DOI: https://doi.org/10.1161/01.CIR.101.23.e215

Lee, J., & Toutanova, K. (2018). Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 3(8).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. Paper presented at the Text summarization branches out.

Montgomery, D. C. (2001). Design and analysis of experiments, John Wiley & Sons. Inc., New York, 1997, 200-201. https://doi.org/10.1007/0-387-22634-6_15

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. Paper presented at the Proceedings of the 40th annual meeting of the Association for Computational Linguistics.

Radhakrishnan, L., Schenk, G., Muenzen, K., Oskotsky, B., Ashouri Choshali, H., Plunkett, T., Butte, A. J. (2023). A certified de-identification system for all clinical text documents for information extraction at scale. JAMIA open, 6(3), ooad045. https://doi.org/10.1093/jamiaopen/ooad045

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . . Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1-67.

Savova, G. K., Danciu, I., Alamudun, F., Miller, T., Lin, C., Bitterman, D. S., Warner, J. L. (2019). Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer research, 79(21), 5463-5470. https://doi.org/10.1158/0008-5472.CAN-19-0579

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Pfohl, S. (2023).

Large language models encode clinical knowledge. Nature, 620(7972), 172-180. https://doi.org/10.48550/arXiv.2212.13138

Sushil, M., Kennedy, V. E., Mandair, D., Miao, B. Y., Zack, T., & Butte, A. J. (2024). CORAL: expert-curated oncology reports to advance language model inference. NEJM AI, 1(4), AIdbp2300110. DOI: 10.1056/AIdbp2300110

Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Seehofnerova, A. (2023). Clinical text summarization: adapting large language models can outperform human experts. Research Square. doi: 10.21203/rs.3.rs-3483777/v1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems. Advances in neural information processing systems, 30(2017).

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. Paper presented at the International conference on machine learning.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675. https://doi.org/10.48550/arXiv.1904.09675

Stress and Driving Performance Evaluation through VR and Physiological Metrics: A Pilot Study

Rehma Razzak
rrazzak@students.kennesaw.edu
Kennesaw State University
Marietta, GA

Yi Li
joy.li@kennesaw.edu
Kennesaw State University
Marietta, GA

Estate Sokhadze
tato.sokhadze@louisville.edu
University of Louisville
Louisville KY

Selena He
she4@kennesaw.edu
Kennesaw State University
Marietta, GA

Abstract

This project explored physiological responses to driving stress using a Virtual Reality (VR) driving simulation, originally developed with the long-term goal of supporting stress management training in specialized populations, such as individuals with Autism Spectrum Disorder (ASD). For this pilot study, a control group was used to evaluate the system and analyze biometric data, including electrodermal activity (EDA), pulse rate, and temperature. The immersive VR environment provided a realistic yet controlled setting to induce and measure stress responses. Advanced statistical techniques, such as mixed linear models, ARIMA modeling, Mann-Whitney U tests, and quantile regression, revealed significant gender-based differences in stress-related biometric metrics, with female participants showing more pronounced changes in EDA and temperature compared to males. Feedback from participants also provided valuable insights for improving the VR simulation's design and user experience.

Keywords: Physiological Data, VR Driving Simulation, Gender Differences, Stress Responses, Driving Performance

Recommended Citation: Razzak, R., Li, Y., Sokhadze, E., He, S., (2025). Stress and Driving Performance Evaluation through VR and Physiological Metrics: A Pilot Study. *Journal of Information Systems Applied Research and Analytics* v18, n2, pp 30-51. DOI# https://doi.org/10.62273/LOKF8848.

Stress and Driving Performance Evaluation through VR and Physiological Metrics: A Pilot Study

Rehma Razzak, Yi Li, Estate Sokhadze and Selena He

1. INTRODUCTION

Driving is a complex activity involving cognitive and physical tasks, including visual and perceptual integration, decision making, vehicle control, and responding to dynamic environments (Caffò et al., 2020; Calvi et al., 2020). Learning to drive requires time and effort, and VR technology has emerged as a powerful tool to enhance this process. By simulating real-world experiences, VR helps new drivers grasp driving fundamentals in an engaging manner, increasing retention of critical information (Alonso et al., 2023). Additionally, VR simulators offer a safe, effective method for evaluating driving performance by integrating perceptual input, cognitive processing, and behavioral output, proving to be reliable and valid tools (Bédard et al., 2010; Davenne et al., 2012). Studies have also shown VR to be useful in examining driving behavior in various conditions, such as rural road intersections (Basu et al., 2022), and in assessing driver stress (Wickens et al., 2015).

Building on this, research has highlighted significant differences in driving behavior based on gender, which have implications for risk perception, traffic accident involvement, and driving performance. Studies indicate that female drivers often experience higher stress levels, and exhibit more pronounced physiological responses in stressful driving scenarios compared to men (Ferrante et al., 2019). For instance, female drivers tend to show lower HRV (Heart Rate Variability) under stress, indicating higher physiological stress levels that correlate with poorer driving outcomes (Arca et al., 2022). Additionally, women often report higher levels of stress and anxiety in driving situations, which leads to more significant physiological reactions such as increased heart rates and EDA Maxwell et al., 2021; Matthews et al., 1999).

Further evidence suggests that male and female drivers exhibit different behaviors during stopping maneuvers in urban environments, with men generally performing these maneuvers more carefully than women (De Blasiis et al., 2017). Driving simulator studies have also shown that female drivers are more likely to be involved in crashes due to errors in yielding, gap acceptance, and speed regulation (Ferrante et al., 2019). These findings highlight the importance of considering gender differences when designing and implementing VR driving simulations.

Incorporating these insights into VR driver training programs can enhance effectiveness by addressing specific stressors and tailoring interventions based on individual biometric profiles. Recognizing and accommodating the unique physiological and psychological responses of different genders can provide a more comprehensive training experience, ultimately contributing to safer driving practices.

As such, we implemented a VR driving game in address some of these findings. In our work, we attempt to address the following questions:

- RQ1: Were the participants less or more stressed as they played the VR Driving Game?
- RQ2: Which physiological metric was the most significant for the participants, and which were the most consistently statistically significant overall?
- RQ3: Were there any significant findings in terms of gender?
- RQ4: Did the VR Driving Game have a positive impact on the participants?

Section 2 of our paper discusses related works, while Section 3 details our experiment and VR driving game. Section 4 presents initial results from the participants' self-report questionnaires, and Section 5 covers data collection and preprocessing. Section 6 provides the analysis and results. Finally, Section 7 concludes the paper, and Section 8 discusses future work for our game.

2. RELATED LITERATURE

Using VR Driving Simulators to Measure Stress

Evaluating stress through physiological signals in a VR driving environment is a significant research area due to its profound impact on driving performance. Stress triggers physiological responses such as increased heart rate, elevated blood pressure, altered breathing patterns, and muscle tension, all of which can impair reaction time, decision-making, and overall driving performance (Kerautret et al., 2021).

In a VR driving environment, real-time monitoring and analysis of physiological responses provide valuable insights into how stress influences driving behavior. This understanding aids in developing interventions to manage stress, ultimately improving road safety (Antoun et al., 2017).

Building on this, a 2023 study by Mateos-García developed a system using biometric sensors in VR simulations to recognize driver stress. Using a PPG (Photoplethysmography) sensor, they found that heart rate closely correlates with stress levels, with ML (Machine Learning) algorithms classifying stress in real-time, demonstrating the feasibility of wearable devices for stress detection in driving scenarios (Mateos-García et al., 2023). Similarly, their 2022 study utilized PPG sensors to detect stress through HRV data, validated with VR experiments, further supporting the use of wearable devices for non-invasive stress detection (Mateos-García et al., 2022).

Expanding on this research, another study examined physiological responses such as GSR (Galvanic Skin Response), BVP (Blood Volume Pulse), and PR (Pupillary Response) in VR driving simulators. Testing 24 participants in five simulation environments revealed significant differences in GSR, highlighting how simulator environments affect stress levels. The study found that female participants exhibited higher stress levels, indicating gender as a crucial factor in physiological responses to driving simulations. Hybrid GA-SVM (Genetic Algorithm-Support Vector machine) and GA-ANN (Genetic Algorithm-Artificial Neural Network) approaches were used for data classification, providing insights into user engagement and stress responses (Liu et al., 2020).

Further exploring physiological responses, a study on individuals with ASD (Autism Spectrum Disorder) used EEG (Electroencephalography) data to classify affective states and mental workload during VR driving simulations. Twenty adolescents with ASD participated, with high classification accuracy achieved using k-nearest neighbors algorithm and univariate feature selection methods, supporting the feasibility of EEG-based models for recognizing affective states in driving contexts (Fan et al., 2018). Similar findings in earlier studies also found that other aspects, such as executive functioning and working memory, were noticeably worse in autistic individuals, and the incorporation of VR Driving simulations resulted in significant improvement (D.J Cox et al., 2017; S.M. Cox et al., 2015).

In the realm of therapeutic applications, a study on VR exposure therapy (VRET) for women with driving phobia demonstrated reduced anxiety and distorted thoughts after VRET sessions. Thirteen women participated and the findings suggested VRET can reduce anxiety and facilitate in vivo exposure for driving phobia without associated risks (Costa et al., 2018).
A cross-sectional study evaluated risky driving behavior across age groups using driving simulators. The sample included 115 drivers divided into young inexperienced (18-21 years), adult experienced (25-55 years), and older adult (70-86 years) groups. Participants were tested on scenarios with varying mental workloads. The study found that moderate scenario complexity highlighted differences in driving ability and elicited realistic behavior, with novel driving measures providing useful, non-redundant information (Michaels et al., 2017).

Investigating the impact of time pressure, one study involved 54 participants driving a 6.9-km urban track with and without time constraints. Measurements included driving performance, eye movement, pupil diameter, cardiovascular and respiratory activity. Under time pressure, participants drove faster, exhibited increased physiological activity, and altered their driving strategies. The findings emphasize the importance of managing stress to improve driving performance (Rendon-Velez et al., 2016).

Another study explored the relationship between flow states and HRV in driving simulations. Eighteen psychology students participated in tasks with varying demand levels to induce flow, anxiety, or boredom. HRV measures indicated that balanced skill-demand levels induced flow, while too high or low demands caused anxiety or

boredom. The study demonstrates how VR environments can effectively investigate psychological states and their impact on physiological responses (Tozman et al., 2015).

Remarks

These studies collectively underscore the significant role of VR driving simulators and physiological data in understanding and managing stress in driving. Leveraging advanced methodology and tools, we can develop effective interventions to enhance driver safety and performance. The versatility and effectiveness of VR driving simulators in enhancing driving skills, assessing driver behavior, and improving traffic safety are well-established.

Despite progress, notable research gaps remain:

- Personalized Models: Many studies develop models that are personalized to the individual subjects in the study. While this can improve the accuracy of stress detection for those individuals, it comes at a cost of generalizability.
- Realism of VR Simulations: The realism of the VR simulations used in these studies can also be a limiting factor. If the VR environment does not accurately reflect real-world driving conditions, the physiological responses observed may not accurately represent the stress responses of drivers in real-world situations.
- There is no standardized way to determine the appropriate complexity of driving scenarios, affecting stress levels and engagement.

Our work differs from previous studies by using more generalized scenarios, allowing our VR driving game to reach a wider audience. Additionally, our emphasis on statistical analysis provides deeper insights into our results, enhancing the overall understanding and applicability of our findings.

3. EXPERIMENT

The experiment was done at Kennesaw State University in an Experimental Studies Lab, that featured a Logitec Car Simulator, with a monitor hooked up to it. A total of 14 participants partook in the study (8 males: Mean = 22.89 years, STD = 2.67 years, 6 females: Mean = 21.20 years. STD = 1.30 years). All participants were 18 and over.

Participant Recruitment

Information about the study was disseminated via email, flyers, and the university's Reddit page. Interested students received a self-report questionnaire to gather basic information about their driving experience, health history, and general well-being, including their physical and mental health and experiences with driving and VR technology.

After completing the questionnaire, participants were emailed a consent form to fill out and return to the PIs. Session times for the study were then scheduled using the online tool Doodle. Each one-on-one study session lasted 45-55 minutes, with 3 to 5 minute breaks as needed.

Upon arrival, participants were asked about their current mental and physical health and familiarity with VR. They then received brief instructions on the game controls before beginning the game.

VR Driving Game

The VR Driving Game was developed by a team of four undergraduate students using the Unity 3D game engine during spring semester (January to April). The Researcher coordinated with the team through weekly meetings to ensure the game aligned with the study's objectives. The game featured low-poly textures for optimized performance and ran on a Windows 10 ASUS laptop with an NVIDIA 2060 GPU, Intel Core i5 processor, and 16 GB of RAM.

The VR Driving Game consisted of 3 levels, briefly explained below:

- Scenario 1: This takes place at a Grocery Store. Participants need to find and enter a parking space. As they reverse, a pedestrian or shopping cart unexpectedly appears behind the vehicle, requiring an abrupt stop to avoid a collision.
- Scenario 2: This also includes a scenario set in a grocery store. However, the participant must then leave the grocery store to navigate a moderate-traffic, daytime urban simulation. A key event during this simulation is a sudden stop by police for an alleged traffic violation.
- Scenario 3: Following a preset route, the key event is a sudden brake by the vehicle in front, causing a minor accident.

In the game, participants used Meta Quest 2 controllers for steering and menu navigation. A calming voice guided participants through the game, aiming to reduce stress.

At startup, participants navigated the main menu using Meta Quest 2 controllers, selecting levels by gently turning the steering wheel to the right, as see in Figure 1. Different sound effects and visuals represented each scenario.

As seen in Figures 2a, 2b and 2c, the participants were instructed to sit inside the car simulator to simulate the feeling of sitting in an actual vehicle.

Figure 2a: Car Simulator Set up

2b: Male Participant

2c: Female Participant

After clearing each scenario, participants were asked if they wanted a 3 to 5 minute break. If they declined, they continued immediately. Upon completing all three scenarios, they were questioned about their feelings on the game and the guiding voice, and asked for improvement suggestions. Participants then received a $30 Amazon gift card and filled out a post-study questionnaire.

First Level

In the first level, set in a grocery store, players are guided to drive into a parking space, with a voice praising their turns and reminding them to stay aware of their surroundings. Blue circular waypoints indicate where players need to go. As they approach, they are warned about a family putting away groceries and instructed to back up to give a car space to exit. They are also cautioned about a nearby child chasing a ball, prompting extra caution. Next, players are directed to a shopping cart waypoint to have enough room to back into a parking spot. While attempting to park, they encounter a pedestrian, requiring careful maneuvering to avoid hitting them. Figures 3a, 3b, and 3c display pictures of this grocery level.

If players successfully navigate the level, the voice praises their caution. If they fail, hitting the family, child, or pedestrian, the voice gently reminds them that accidents happen and encourages them to take a deep breath and try again. Notably, only one participant hit the pedestrian behind their car while backing into the parking spot. When this happens, the pedestrian shouts, "Watch it!"

Second Level

In Scenario 2, players are instructed to back out of their parking space to leave the grocery store, with reminders to check their surroundings and mirrors. Blue waypoints guide them on where to drive. Upon approaching a turn, they are instructed to make a right turn. Shortly after, a police siren is heard, prompting the player to pull over. The police officer explains the reason for the stop and then allows the player to continue driving. Figure 4 shows a snapshot of the second level.

Figure 4: 2nd Level - Policeman.

Third Level

The third level and last level of the game takes place after the second level. In the third level of the game, the players are instructed to drive on the road. At some point in the game, the player is warned that a car in front of them is breaking hard. A blue waypoint appears in front of the player, close to the car in front of them so that they may brake in time, not hitting the car. Figure 5 demonstrates a snapshot of this.

Figure 5: Level 3 scenario.

If players crash into the car ahead, they fail the level and are respawned to try again. After successfully braking, the car in front drives away. Shortly after, another car hits the player from behind. Players are reminded to stay calm and drive to the nearest gas station. A blue waypoint guides them to a parking spot. Upon parking, the car that hit the player arrives, and the driver apologizes and takes responsibility. Figure 6 illustrates this interaction.

Figure 6: Driver in Green shirt.

4. INTIAL ANALYSIS: SELF REPORT QUESTIONAIRE RESULTS

Upon reviewing the self-report questionnaire, participants identified several driving difficulties, categorized into Situational Awareness, Specific Maneuvers, Multitasking and Cognitive Load, and Distance and Speed Management. The number of responses for each category out of 14 participants is detailed in Figure 7. As shown in the pie chart, "Specific Maneuvers" received the highest number of responses, indicating it as the most cited difficulty among participants. Additionally, participants rated their driving skills on a scale from poor to excellent. The majority rated their skills as "good," as depicted in Figure 8. Comparing genders, male participants more often rated their driving skills as "good" or "excellent," while female participants were more likely to rate themselves as "average."

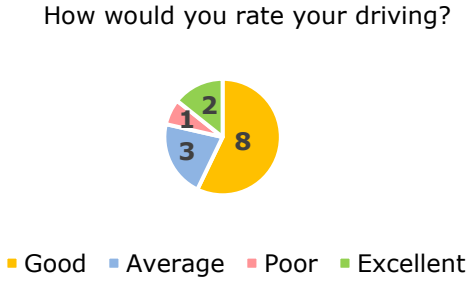

Figure 7: Driver Difficulty Category Responses.

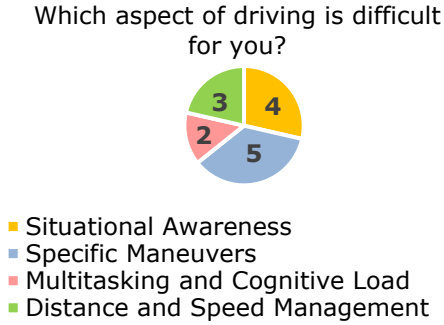

Figure 8: Driver Skill Ratings Distributions.

5. DATA COLLECTION AND PREPROCESSING

Data Collection

Physiological data was collected using the EmbracePlus Smartwatch while participants engaged with the VR game. The EmbracePlus, a medical-grade wearable, gathered various physiological parameters, which were transferred via Bluetooth to the Empatica CareLab app. The app analyzed the data, extracted digital biomarkers, and uploaded the information to the Empatica Cloud for secure storage and access via the Care Portal. This portal allowed team members to manage studies and visualize participants' biomarkers. Data was organized into a primary "participant_data" folder with subfolders for different dates.

Data Preprocessing

Preprocessing involved examining and modifying the data stored in a hierarchical directory structure. A Python script verified the directory, traversed subdirectories, and targeted 'digital_biomarkers' and 'aggregated_perminute.' It listed CSV files, loaded them into pandas DataFrames, converted timestamps to Eastern Time, and dropped 'missing_value_reason' columns. Cleaned data was saved back in a suitable format for analysis.

To address missing values, the script generated random values within specified ranges for EDA, pulse rate, and temperature, filling in missing data appropriately. Given the small dataset size, dropping rows was not a viable option as it would result in significant data loss and reduce the statistical power of the analysis. While regression-driven data imputation was considered, it was deemed less practical due to the limited data points and potential overfitting risks. Additionally, generating random values allowed for greater control over the imputation process, ensuring consistency and reliability in the data. This method ensured complete, properly formatted data covering the ideal time range for each participant.

Finally, the script identified all 'modified.csv' files, checked for remaining missing values, and confirmed data readiness for analysis by iterating over each DataFrame and reporting missing values. This quality assurance measure ensured comprehensive data for subsequent analysis.

6. DATA ANALYSIS AND RESULTS

Stationarity Testing and Initial Results

Our data analysis primarily focused on three

variables: EDA, pulse rate, and temperature, using the modified CSV files. Initial exploration revealed minimal outliers. To ensure reliable time series analysis, we conducted the Augmented Dickey-Fuller (ADF) test to check for stationarity in our data. We found some non-stationary data, which required rectification.

To address this, we implemented a script with a loop that, for each metric (EDA, pulse rate, temperature), performed the ADF test, checked if differencing was required, and applied the appropriate order of differencing. If a series remained non-stationary after first-order differencing, the script applied second-order differencing and rechecked for stationarity. This process continued until all series were stationary, ensuring our data was primed for accurate and meaningful analysis. First-order differencing reveals the rate of change between consecutive observations, making it easier to analyze seasonality and cyclical patterns. Second-order differencing is useful for addressing quadratic trends by removing the trend in the rate of change, highlighting any underlying seasonality or long-term cycles.

In our case, many series were initially non-stationary. EDA and pulse rate series became stationary after applying first-order differencing. Several temperature series required second-order differencing to become stationary.

After achieving stationarity, we analyzed how EDA, pulse rate, and temperature changed over time for all participants.

RQ1: Were the participants less or more stressed as they played the VR Driving Game?

Figures 9, 10, and 11 show a general trend (trend line shown in red) of reduced stress levels among participants playing the VR Driving Game. The methods used for the trendlines in Figures 9, 10, and 11 were LOESS (Locally Weighted Scatterplot Smoothing) for EDA with the frac parameter set to 0.20, Polynomial Regression for Pulse Rate with the degree set to 5, and Moving Average for Temperature with the window size being set to 10. LOESS is a non-parametric method that can flexibly fit curves to data by performing multiple localized regressions. This is particularly useful when the data exhibits non-linear patterns that a simple linear model cannot capture. Furthermore, by fitting a polynomial of a specified degree to the data, Polynomial Regression can model non-linear relationships. This allows the trend line to bend and fit the data more accurately than a straight line, capturing the underlying patterns more effectively. Given the non-linearity of the physiological responses, using regular linear regression modelling would have oversimplified these responses, leading to inaccurate trend readings.

Regarding RQ1, our analysis revealed that 12 out of 14 participants experienced a decrease in EDA, indicating reduced stress or arousal and suggesting increased relaxation over time. Similarly, pulse rates decreased in 10 out of 14 participants, further supporting the notion of relaxation. Additionally, 8 out of 14 participants showed a decrease in temperature, indicating physical cooling down as they played.

Further analysis of gender-specific trends revealed some differences. Two out of six female participants experienced a temperature decrease, and only one had an increased pulse rate. In contrast, among male participants, only one showed an increase in EDA, while three had increased pulse rates, and six experienced a decrease in temperature. These variations suggest that gender may influence physiological responses to stress, but overall, stress reduction was observed across both male and female participants.

Therefore, the answer RQ1 is that most participants, regardless of gender, experienced reduced stress, suggesting that the VR Driving Game had a generally calming effect over time.

RQ2: Which physiological metric was the most significant for the participants, and which were the most consistently statistically significant overall?

To answer RQ2, we calculated Cohen's D results for the three metrics (EDA, pulse rate, and temperature) for each participant, as shown in Table 1. Cohen's D measures effect size, interpreted as follows:

- Small effect size: $d \approx 0.2$
- Medium effect size: $d \approx 0.5$
- Large effect size: $d \approx 0.8$

As such, we can generalize the following findings:

- Pulse Rate vs. Temperature: Pulse rate generally shows a positive relationship with temperature across participants, meaning that higher temperatures tend to correlate with higher pulse rates. This aligns with the physiological response where increased body temperature can

lead to higher heart rates as the body works to regulate its internal temperature.

- Pulse rate generally shows higher values compared to temperature across participants, consistent with the expected physiological response where pulse rate increases in response to various stimuli or activities, whereas body temperature fluctuates within a narrower range under normal conditions.

- EDA vs. Pulse Rate: Across most participants, EDA tends to show either lower or higher activity compared to pulse rate. This suggests that in some individuals, changes in electrodermal activity might correlate positively with changes in pulse rate, indicating a potential physiological response pattern.

Next, we assessed the statistical significance of our results to ensure practical meaning behind our findings, as shown in Table 2. Statistically significant results were found for the following participants in terms of EDA, Pulse Rate and Temperature:

- EDA: Participants 1,2, 11, 12, 13
- Pulse Rate: Participants 7, 9, 10, 11, 13, 14
- Temperature: Participants 1, 5, 7, 10, 2, 11, 12, 13.

Significant changes in EDA were observed for five participants, while significant changes in pulse rate were noted for six participants, suggesting substantial changes in heart rate potentially related to stress. Significant changes in temperature were observed for eight participants.

To further validate our findings, we used bootstrapping for Cohen's D. Bootstrapping, a resampling technique, estimates statistics on a population by sampling a dataset with replacement. It is particularly useful when data normality is in doubt, or the sample size is small. In our case, the dataset is small, necessitating extra caution in interpreting findings. Bootstrapping can be particularly useful for the following reasons:

- Confidence Intervals: Bootstrapping can be used to construct confidence intervals around the Cohen's D statistic. This provides a range of plausible values for the population parameter and gives an indication of how precise the estimates are.

- Small Sample Sizes: Cohen's D is sensitive to the assumption of normality. When the sample size is small, this assumption may not hold, and the estimate of Cohen's D may be biased. Bootstrapping does not rely on the assumption of normality and can provide a more accurate estimate in these cases.

- Stability of the Estimate: By resampling the data multiple times and calculating Cohen's D for each sample, we can get a sense of the variability or stability of our estimate. If the bootstrapped estimates of Cohen's D vary widely, it suggests that the original estimate may not be reliable.

Significance across all participants was determined by examining the confidence intervals (CI Low and CI High) of Cohen's D values for each metric. A metric is considered significant if its confidence interval does not include zero, indicating a reliable effect size.

Based on this analysis, as shown in Table 3, temperature emerged as the most consistently significant metric, with significant results in nine participants. EDA was significant for six participants, and pulse rate for five participants. As such, we can conclude that for RQ2, temperature is likely the most reliable indicator of physiological changes, showing consistent significance across participants.

RQ3: Were there any significant findings in terms of gender?

The analysis next focused on potential gender-based differences, as detailed in Table 4. We found that pulse rate increases were more pronounced in male participants compared to females. Female participants showed mixed results, with some displaying positive effect sizes and others negative. Overall, significant differences between male and female participants were observed.

Notably, as shown in Tables 5 through 7, gender had a significant effect on both EDA ($P>|z| = 0.000$) and temperature ($P>|z| = 0.001$). These results indicate meaningful physiological differences between males and females for these metrics.

To further refine our understanding, we employed Quantile Regression in addition to

traditional mixed models. Unlike standard models, which assume normally distributed residuals, quantile regression does not require this assumption. This makes it better suited to handling non-normal data and outliers, allowing for a more nuanced exploration of the relationships between gender and physiological metrics such as EDA, pulse rate, and temperature. By estimating the conditional median or other quantiles, quantile regression provided insights that traditional models might have missed, especially in skewed distributions.

Specifically, we were interested in how the relationships between gender and outcomes (e.g., EDA, pulse rate, and temperature) varied across different parts of the distribution. While mixed models offered insight into the average effects of gender, quantile regression revealed how gender influenced different segments of the outcome distribution. This combined approach allowed us to capture both overall trends and the specific ways gender affected physiological responses, offering a more comprehensive understanding of its impact on EDA, pulse rate, and temperature.

The results of the quantile regression analysis showed that gender had a particularly significant effect on EDA for female participants, as demonstrated in Table 8. This suggests that the physiological response to the VR Driving Game, particularly in terms of EDA, differed notably by gender, with female participants exhibiting distinct patterns compared to their male counterparts. Thus, we can conclude that for RQ3, there were significant findings in terms of gender.

RQ4: Did the VR Driving Game have a positive impact on the participants?
After playing the game, participants completed a post-study questionnaire. This questionnaire included questions about which level they found most stressful and whether the guiding voice was helpful in calming them down and providing instructions.

As shown in Figure 12, Scenario 1 was the most stressful for both male and female participants. Interestingly, Scenario 3 was the second most stressful among female participants, while none of the female participants found Scenario 2 to be stressful.

When evaluating the effectiveness of the calming voice in terms of helpful hints, intervention, and overall appreciation, none of the participants found the voice annoying, and most found the voice's interventions effective and helpful. Additionally, none of the participants were dissatisfied with the game or the voice, finding it helpful and calm.

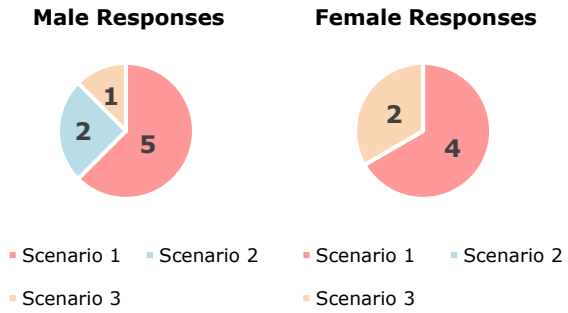

Figure 12: Stressful Scenario Responses – Male (left) and Female (right)

Figure 13 represents the participants' responses regarding their satisfaction with the VR game. Interestingly, three male participants rated the effectiveness of the voice's interventions as neutral. Similar findings were observed when evaluating whether the voice was helpful and when asked about the instructions and guidance provided by the voice. Overall, we can conclude that for RQ4, the VR driving game had a positive impact on the participants.

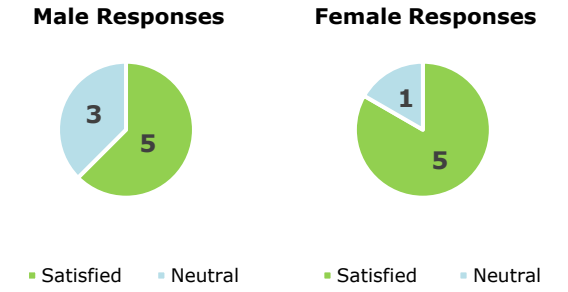

Figure 13: VR Game Satisfaction ratings – Male (left) and Female (right)

7. CONCLUSIONS

This project provided a detailed examination of biometric data from participants engaged in a driving simulation. By utilizing various statistical methods, significant insights were gained regarding the differences in biometric responses based on gender and other factors.

Our findings and analysis revealed that the effect of gender on biometric responses was significant. Our analysis revealed that female participants exhibited notable changes in EDA and temperature, suggesting a notable physiological response to the driving simulation

(Kerautret et al., 2021). This aligns with broader research findings indicating that women often show stronger physiological responses to stress in driving scenarios compared to men (Mostowfi & Kim, 2022).

Studies revealed that women often have faster, larger, and longer-lasting stress responses compared to men. For example, women have more receptors for stress-related neurotransmitters, and their stress responses, such as increased heart rate and electrodermal activity, can be more pronounced and prolonged (James et al., 2023; Wang et al., 2007). Studies also have observed that women often report higher levels of stress and anxiety in driving situations compared to men, leading to more significant physiological reactions such as increased heart rates and electrodermal activity (Arca, 2022; Antoun, 2017).

In our case, temperature and EDA have shown to be more reliable metrics for measuring driving stress compared to pulse rate. EDA directly measures sympathetic nervous system activity, providing real-time data on psychological or physiological arousal, while temperature changes reflect peripheral responses to stress.

While our study's sample size was limited, the observed trends are consistent with broader research findings on gender differences in physiological responses to driving stress.

8. FUTURE WORKS

Even though our VR driving game was successful, we plan to expand it based on participants' valuable suggestions. Participants recommended incorporating additional scenarios to increase realism and stress responses, such as inclement weather conditions like driving in the rain or nighttime driving. One participant suggested that the policeman in the simulation should be more aggressive, while two others recommended adding distractions such as music or phone calls to further simulate real-world driving. Enhancing the environment by adding more people or livelier scenery was another suggestion, as well as incorporating more interactive features with the Meta Quest 2 controllers, such as honking the horn or using turn signals.

In future iterations, we aim to add even more varied stress-inducing scenarios, such as receiving sudden instructions from a co-pilot, being cut off by another driver, engaging in a heated argument with passengers, or missing an exit due to a vehicle blocking the passing lane. These scenarios would provide opportunities to explore individual differences in responses to a wider range of stressful driving situations. For example, while some drivers might remain calm, others could experience heightened stress or road rage, giving us valuable insights into how personality traits influence stress responses.

Another area for improvement involves standardizing the breaks between game sessions. Consistent break durations will help control for reductions in adrenaline levels and ensure comparability across scenarios. Future studies will implement uniform breaks to maintain consistency in the physiological data collected.

We also recognize that the calming voice used in this study may have influenced participants' stress levels by reducing the emotional impact of stressful scenarios. Future studies could investigate how different voice tones—such as critical or chiding voices—affect participants' stress responses. Additionally, introducing scenarios where an accident is inevitable, without prior warning, could provide insight into how drivers react when failure is unavoidable.

Regarding data collection, expanding beyond the current reliance on wearables could yield richer insights. In future studies, we plan to collect additional data, such as facial expressions, eye movements, or galvanic skin response, to better understand the full range of participants' stress responses during driving simulations. If funding allows, these enhancements will help provide a more comprehensive view of stress dynamics.

9. ACKNOWLEDGEMENTS

We extend our gratitude to our undergraduate team for developing the game. The Principal Investigator (PI) supervised the game's quality, with refinements made by the PI's advisor. This project was sponsored by the Foundation for Neurofeedback and Neuromodulation Research (FNNR) through Mini-Grant #501-2023.

10. REFERENCES

Alonso, F., Faus, M., Riera, J. V., Fernandez-Marin, M., & Useche, S. A. (2023). Effectiveness of Driving Simulators for Drivers' Training: A Systematic Review. Applied Sciences, 13(9), 5266. https://doi.org/10.3390/app13095266

Antoun, M., Edwards, K. M., Sweeting, J., & Ding, D. (2017). The acute physiological

stress response to driving: A systematic review. PLOS ONE, 12(10), e0185517. https://doi.org/10.1371/journal.pone.0185517

Arca, A. A., Chouljian, E., & Mouloua, M. (2022). The Role of Gender Differences in Distracted Driving Behavior: A Psychophysiological Approach. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 66(1), 1280–1284. https://doi.org/10.1177/1071181322661453

Basu, S., Ferrante, C., & De Blasiis, M. R. (2022). Influence of Intersection Density on Risk Perception of Drivers in Rural Roadways: A Driving Simulator Study. Sustainability, 14(13), 7750. https://doi.org/10.3390/su14137750

Bédard, M., Parkkari, M., Weaver, B., Riendeau, J., & Dahlquist, M. (2010). Assessment of Driving Performance Using a Simulator Protocol: Validity and Reproducibility. The American Journal of Occupational Therapy, 64(2), 336–340. https://doi.org/10.5014/ajot.64.2.336

Caffò, A. O., Tinella, L., Lopez, A., Spano, G., Massaro, Y., Lisi, A., Stasolla, F., Catanesi, R., Nardulli, F., Grattagliano, I., & Bosco, A. (2020). The Drives for Driving Simulation: A Scientometric Analysis and a Selective Review of Reviews on Simulated Driving Research. Frontiers in Psychology, 11. https://doi.org/10.3389/fpsyg.2020.00917

Calvi, A., D'Amico, F., Ferrante, C., & Bianchini Ciampoli, L. (2020). A driving simulator validation study for evaluating the driving performance on deceleration and acceleration lanes. Advances in Transportation Studies, 50, 67-80. https://doi.org/10.4399/97888255317325

Costa, R. T. da, Carvalho, M. R. de, Ribeiro, P., & Nardi, A. E. (2018). Virtual reality exposure therapy for fear of driving: analysis of clinical characteristics, physiological response, and sense of presence. Revista Brasileira de Psiquiatria, 40(2), 192–199. https://doi.org/10.1590/1516-4446-2017-2270

Cox, D. J., Brown, T., Ross, V., Moncrief, M., Schmitt, R., Gaffney, G., & Reeve, R. (2017). Can Youth with Autism Spectrum Disorder Use Virtual Reality Driving Simulation Training to Evaluate and Improve Driving Performance? An Exploratory Study. Journal of Autism and Developmental Disorders, 47(8), 2544–2555. https://doi.org/10.1007/s10803-017-3164-7

Cox, S. M., Cox, D. J., Kofler, M. J., Moncrief, M. A., Johnson, R. J., Lambert, A. E., Cain, S. A., & Reeve, R. E. (2015). Driving Simulator Performance in Novice Drivers with Autism Spectrum Disorder: The Role of Executive Functions and Basic Motor Skills. Journal of Autism and Developmental Disorders, 46(4), 1379–1391. https://doi.org/10.1007/s10803-015-2677-1

Davenne, D., Lericollais, R., Sagaspe, P., Taillard, J., Gauthier, A., Espié, S., & Philip, P. (2012). Reliability of simulator driving tool for evaluation of sleepiness, fatigue and driving performance. Accident Analysis & Prevention, 45, 677–682. https://doi.org/10.1016/j.aap.2011.09.046

De Blasiis, M. R., Ferrante, C., Veraldi, V., & Moschini, L. (2017). Risk perception assessment using a driving simulator: a gender analysis. Road Safety and Simulation–RSS Proceedings. The Hague.

Edwards, J. D., Myers, C., Ross, L. A., Roenker, D. L., Cissell, G. M., McLaughlin, A. M., & Ball, K. K. (2009). The Longitudinal Impact of Cognitive Speed of Processing Training on Driving Mobility. The Gerontologist, 49(4), 485–494. https://doi.org/10.1093/geront/gnp042

Fan, J., Wade, J. W., Key, A. P., Warren, Z. E., & Sarkar, N. (2018). EEG-Based Affect and Workload Recognition in a Virtual Driving Environment for ASD Intervention. IEEE Transactions on Biomedical Engineering, 65(1), 43–51. https://doi.org/10.1109/tbme.2017.2693157

Ferrante, C., Varladi, V., & De Blasiis, M. R. (2019). Gender Differences Measured on Driving Performances in an Urban Simulated Environment. Advances in Human Factors and Simulation, 144–156. https://doi.org/10.1007/978-3-030-20148-7_14

Goedicke, D., Li, J., Evers, V., & Ju, W. (2018). VR-OOM. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3173574.3173739

James, K. A., Stromin, J. I., Steenkamp, N., & Combrinck, M. I. (2023). Understanding the relationships between physiological and psychosocial stress, cortisol and cognition. Frontiers in Endocrinology, 14. https://doi.org/10.3389/fendo.2023.1085950

Kerautret, L., Dabic, S., & Navarro, J. (2021). Sensitivity of Physiological Measures of Acute Driver Stress: A Meta-Analytic Review. Frontiers in Neuroergonomics, 2. https://doi.org/10.3389/fnrgo.2021.756473

Liu, Y.-H., Spiller, M., Ma, J., Gedeon, T., Hossain, M. Z., Islam, A., & Bierig, R. (2020). User Engagement with Driving Simulators: An Analysis of Physiological Signals. HCI International 2020 – Late Breaking Papers: Digital Human Modeling and Ergonomics, Mobility and Intelligent Environments, 130–149. https://doi.org/10.1007/978-3-030-59987-4_10

Mateos-GarcíaA, N., Gil-González, A. B., Reboredo, A. de L., & Pérez-Lancho, B. (2022). Driver Stress Detection in Simulated Driving Scenarios with Photoplethysmography. Distributed Computing and Artificial Intelligence, 19th International Conference, 291–301. https://doi.org/10.1007/978-3-031-20859-1_29

Mateos-GarcíaB, N., Gil-González, A.-B., Luis-Reboredo, A., & Pérez-Lancho, B. (2023). Driver Stress Detection from Physiological Signals by Virtual Reality Simulator. Electronics, 12(10), 2179. https://doi.org/10.3390/electronics12102179

Matthews, G., Joyner, L. A., & Newman, R. (1999). Age and Gender Differences in Stress Responses during Simulated Driving. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 43(18), 1007–1011. https://doi.org/10.1177/154193129904301802

Maxwell, H., Weaver, B., Gagnon, S., Marshall, S., & Bédard, M. (2021). The validity of three new driving simulator scenarios: detecting differences in driving performance by difficulty and driver gender and age. Human factors, 63(8), 1449-1464.

Michaels, J., Chaumillon, R., Nguyen-Tri, D., Watanabe, D., Hirsch, P., Bellavance, F., Giraudet, G., Bernardin, D., & Faubert, J. (2017). Driving simulator scenarios and measures to faithfully evaluate risky driving behavior: A comparative study of different driver age groups. PLOS ONE, 12(10), e0185909. https://doi.org/10.1371/journal.pone.0185909

Mostowfi, S., & Kim, J. H. (2022). Understanding Drivers' Physiological Responses in Different Road Conditions. HCI in Mobility, Transport, and Automotive Systems, 218–230. https://doi.org/10.1007/978-3-031-04987-3_15

Rendon-Velez, E., van Leeuwen, P. M. ., Happee, R., Horváth, I., van der Vegte, W. F., & de Winter, J. C. F. (2016). The effects of time pressure on driver performance and physiological activity: A driving simulator study. Transportation Research Part F: Traffic Psychology and Behaviour, 41, 150–169. https://doi.org/10.1016/j.trf.2016.06.013

Tozman, T., Magdas, E. S., MacDougall, H. G., & Vollmeyer, R. (2015). Understanding the psychophysiology of flow: A driving simulator experiment to investigate the relationship between flow and heart rate variability. Computers in Human Behavior, 52, 408–418. https://doi.org/10.1016/j.chb.2015.06.023

Wang, J., Korczykowski, M., Rao, H., Fan, Y., Pluta, J., Gur, R. C., McEwen, B. S., & Detre, J. A. (2007). Gender difference in neural response to psychological stress. Social Cognitive and Affective

Neuroscience, 2(3), 227–239. https://doi.org/10.1093/scan/nsm018

Wickens, C. M., Wiesenthal, D. L., & Roseborough, J. E. W. (2015). In Situ Methodology for Studying State Driver Stress: A Between-Subjects Design Replication. Journal of Applied Biobehavioral Research, 20(1), 37–51. Portico. https://doi.org/10.1111/jabr.1202

APPENDIX A – Figures

Figure 1: Main Menu of the game. The three floating rocks represent the levels of the game.

Figure 3a: Grocery Parking lot with Blue waypoint.

Figure 3b: Child with ball (circled in red).

Figure 3c: Backing into Pedestrian (circled in blue).

Figure 3: Grocery Level

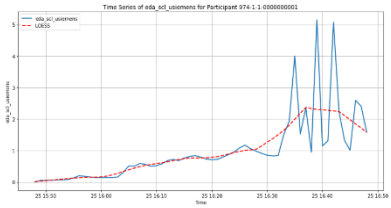

Figure 9a: Participant 1(M)

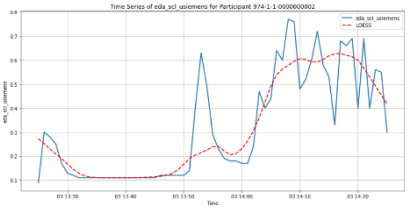

Figure 9b: Participant 2(M)

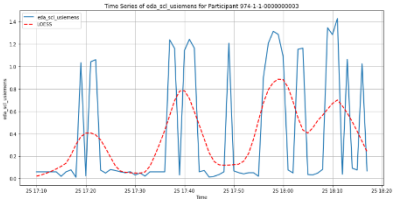

Figure 9c: Participant 3(M)

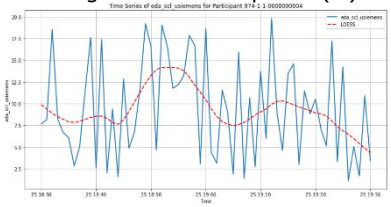

Figure 9d: Participant 4(F)

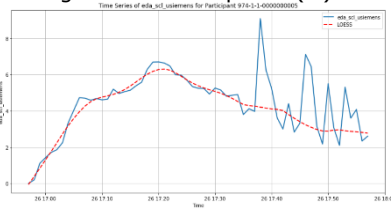

Figure 9e: Participant 5(F)

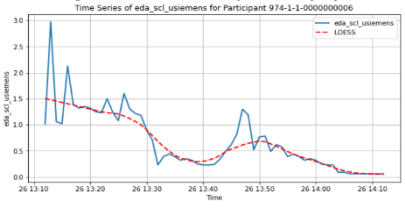

Figure 9f: Participant 6(F)

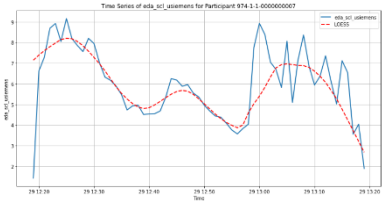

Figure 9g: Participant 7(M)

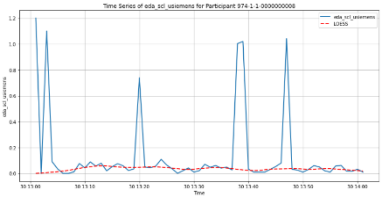

Figure 9h: Participant 8(F)

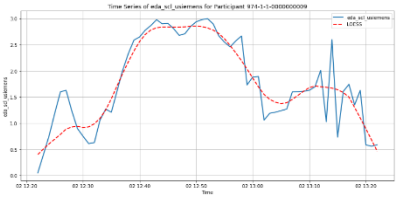

Figure 9i: Participant 9(M)

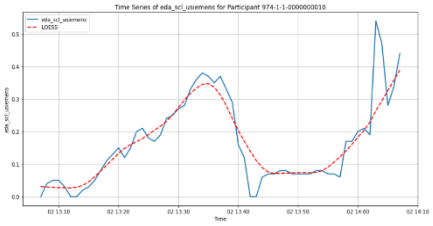

Figure 9j: Participant 10(M)

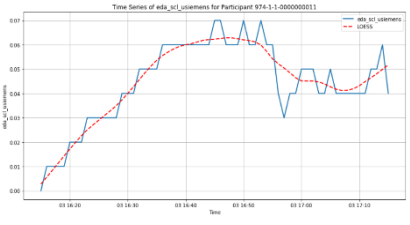

Figure 9k: Participant 11(F)

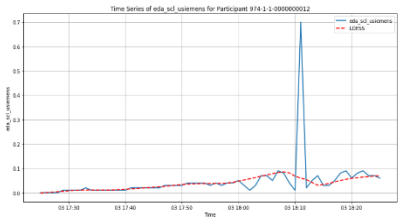

Figure 9l: Participant 12(M)

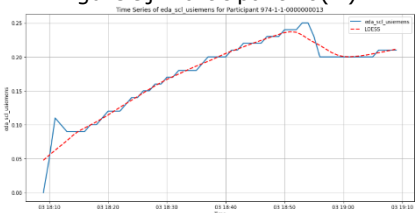

Figure 9m: Participant 13(F)

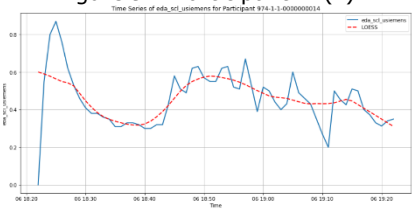

Figure 9n: Participant 14(M)

Figure 9: Participant's EDAs over time.

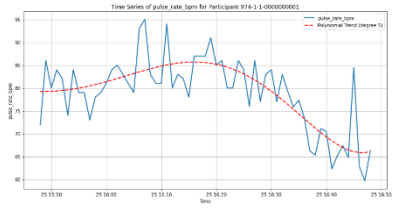

Figure 10a: Participant 1 (M)

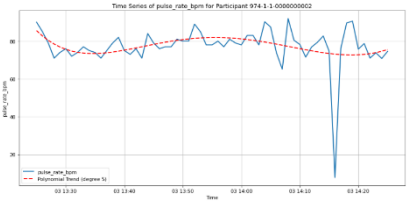

Figure 10b: Participant 2 (M)

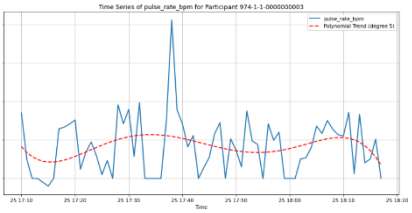

Figure 10c: Participant 3 (M)

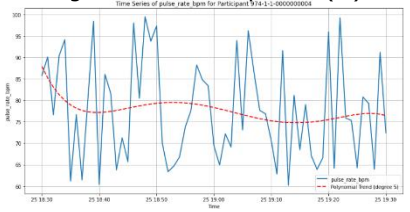

Figure 10d: Participant 4 (F)

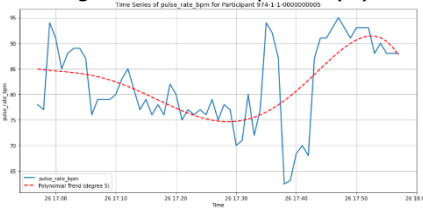

Figure 10e: Participant 5 (F)

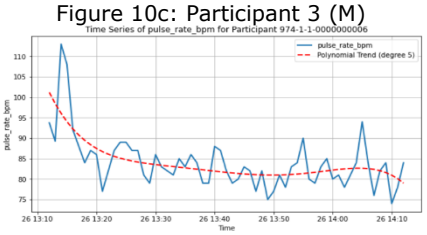

Figure 10f: Participant 6 (F)

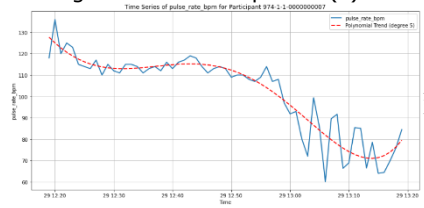

Figure 10g: Participant 7 (M)

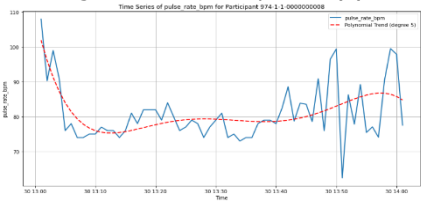

Figure 10h: Participant 8 (F)

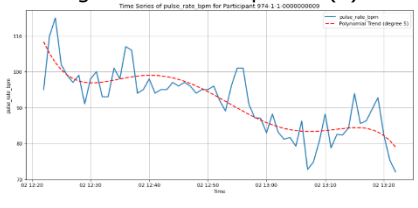

Figure 10i: Participant 9 (M)

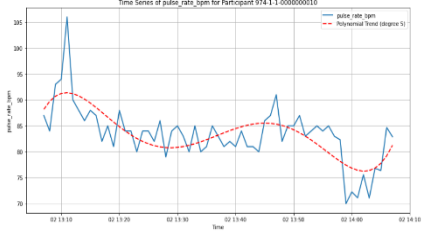

Figure 10j: Participant 10 (M)

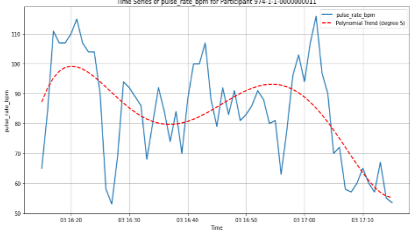

Figure 10k: Participant 11 (F)

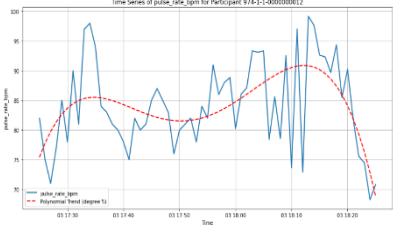

Figure 10l: Participant 12 (M)

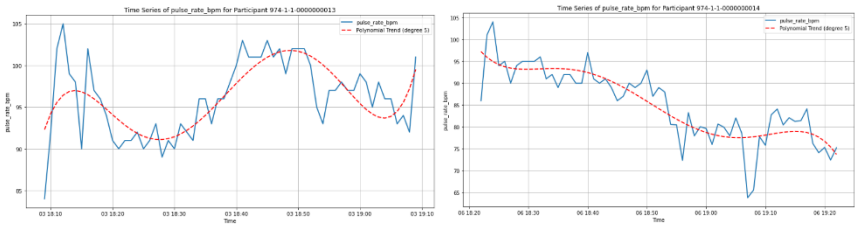

Figure 10m: Participant 13 (F) Figure 10n: Participant 14 (M)

Figure 10: Participants' Pulse Rates Over time.

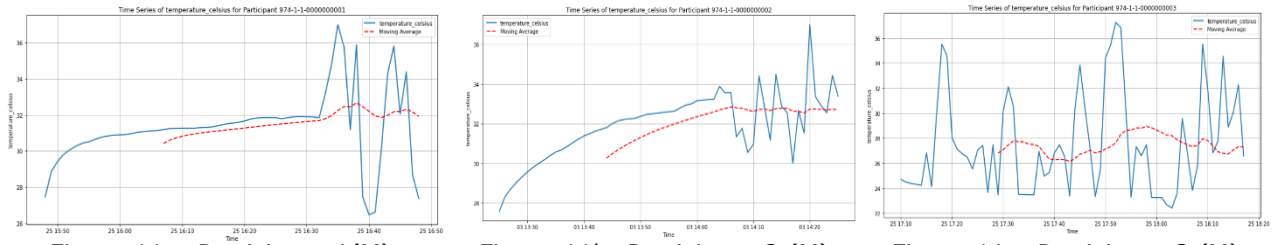

Figure 11a: Participant 1(M) Figure 11b: Participant 2 (M) Figure 11c: Participant 3 (M)

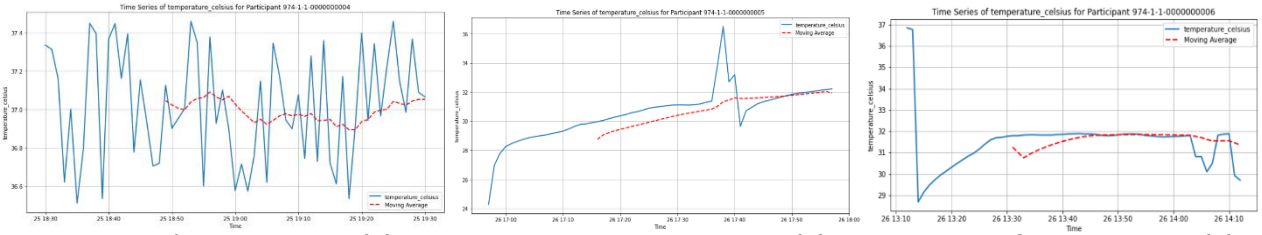

Figure 11d: Participant 4 (F) Figure 11e: Participant 5 (F) Figure 11f: Participant 6 (F)

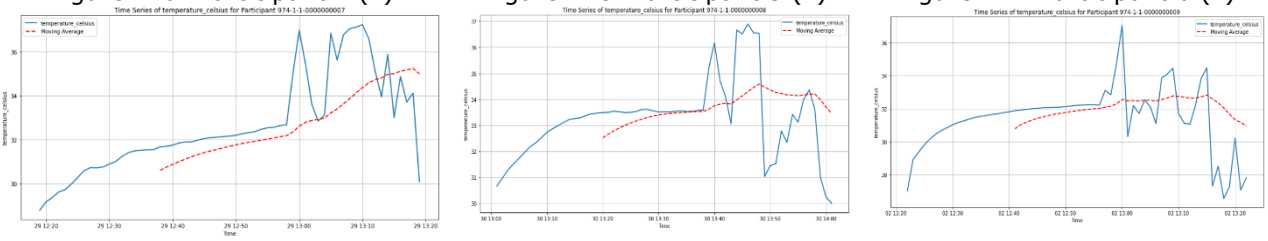

Figure 11g: Participant 7 (M) Figure 11h: Participant 8 (F) Figure 11i: Participant 9 (M)

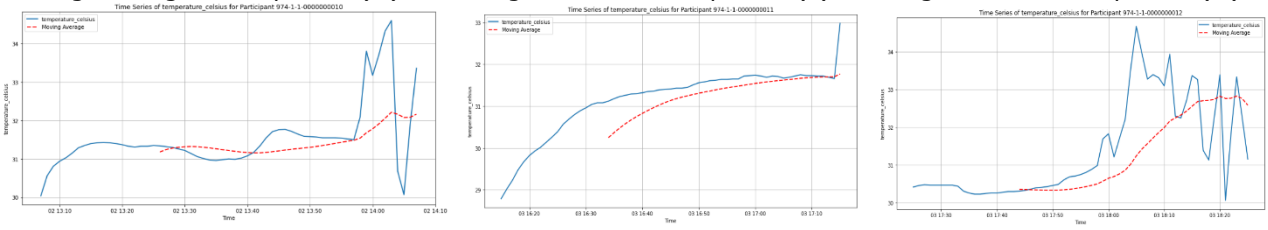

Figure 11j: Participant 10 (M) Figure 11k: Participant 11 (F) Figure 11l: Participant 12 (M)

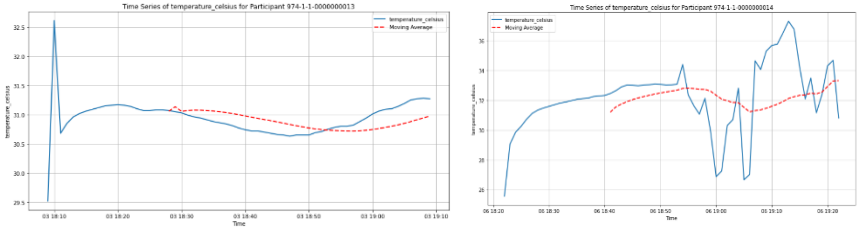

Figure 11m: Participant 13 (F) Figure 11n: Participant 14 (M)

Figure 11: Participants' Temperatures Over time.

APPENDIX B – Tables

Table 1. Participant Metrics with Cohen's D

Participant	Metric	Cohen's D
1	EDA	-1.438387
	Pulse Rate	0.838545
	Temperature	-0.587648
2	EDA	-0.344259
	Pulse Rate	-0.233768
	Temperature	-0.294475
3	EDA	0.336248
	Pulse Rate	0.298969
	Temperature	0.181656
4	EDA	-0.039809
	Pulse Rate	-0.227438
	Temperature	-1.901704
5	EDA	0.335254
	Pulse Rate	2.067079
	Temperature	-2.005618
6	EDA	0.033700
	Pulse Rate	-0.170715
	Temperature	-0.546009
7	EDA	0.188297
	Pulse Rate	1.972688
	Temperature	-0.214628
8	EDA	-0.079148
	Pulse Rate	0.791275
	Temperature	-0.918734
9	EDA	-1.774026
	Pulse Rate	0.103837
	Temperature	-1.393755
10	EDA	-0.796594
	Pulse Rate	0.610066
	Temperature	-1.870847
11	EDA	-0.649005
	Pulse Rate	-0.456534
	Temperature	-2.319676
12	EDA	-2.398581
	Pulse Rate	-1.229099
	Temperature	0.420771
13	EDA	0.030003
	Pulse Rate	2.768771
	Temperature	-0.314131
14	EDA	0.030003
	Pulse Rate	2.768771
	Temperature	-0.314131

Table 2: Statistical Analysis Results for Participants

Participant	Metric	Statistic	p-value	Adjusted p-value
1	EDA	16	9.68e-11	9.21e-10
	Pulse Rate	631	1.68e-02	3.28e-02
	Temperature	184	5.16e-05	1.34e-04
2	EDA	74	1.41e-08	5.51e-08
	Pulse Rate	424	5.58e-01	6.16e-01
	Temperature	116	4.96e-07	1.61e-06
3	EDA	494	3.04e-01	3.95e-01
	Pulse Rate	541.5	6.57e-01	6.92e-01
	Temperature	459	1.46e-01	2.27e-01
4	EDA	555	1.96e-01	2.84e-01
	Pulse Rate	539	2.88-01	3.88e-01
	Temperature	516	4.66e-01	5.51e-01
5	EDA	505	5.68e-01	6.16e-01
	Pulse Rate	386	2.56e-01	3.57e-01
	Temperature	18	1.18e-10	9.21e-10
6	EDA	298.4	1.96e−01	3.95e−01
	Pulse Rate	504.5	1.28e−02	3.18e−02
	Temperature	258.7	5.16e−05	5.51e−06
7	EDA	560	1.72e-01	2.59e-01
	Pulse Rate	908.5	1.57e-10	1.02e-09
	Temperature	24	2.08e-10	1.16e-09
8	EDA	533	3.29e-01	4.15e-01
	Pulse Rate	416.5	4.87e-01	5.59e-01
	Temperature	290	1.18e-02	2.55e-02
9	EDA	520	4.31e-01	5.26e-01
	Pulse Rate	862.5	9.92e-09	4.30e-08
	Temperature	316	3.21e-02	5.70e-02
10	EDA	445.5	7.83e-01	7.83e-01
	Pulse Rate	642	1.05e-02	2.54e-02
	Temperature	171.5	2.35e-05	6.57e-05
11	EDA	295	1.24e-02	2.55e-02
	Pulse Rate	626.5	2.01e-02	3.73e-02
	Temperature	0	2.01e-11	6.55e-10
12	EDA	96.5	8.23e-08	2.91e-07
	Pulse Rate	329.5	5.13e-02	8.70e-02
	Temperature	30	3.59e-10	1.75e-09
13	EDA	12	4.84e-11	6.55e-10
	Pulse Rate	172	2.30e-05	6.57e-05
	Temperature	641.5	1.10e-02	2.54e-02
14	EDA	444.5	7.72e-01	7.83e-01
	Pulse Rate	920.5	5.04e-11	6.55e-10
	Temperature	356	1.17e-01	1.90e-01

Table 3: Cohen's D Effect Size with 95% Confidence Intervals for EDA, Pulse Rate, and Temperature Bootstrap results

Participant	Metric	Cohen's D	CI Low	CI High
1	EDA	-1.552439	-2.10668	-1.25101
	Pulse Rate	0.870001	0.398781	1.367424
	Temperature	-0.643849	-1.26489	-0.10594
2	EDA	-0.345983	-0.86209	0.146409
	Pulse Rate	-0.171072	-0.49966	0.423803
	Temperature	-0.399335	-0.84821	-0.11594
3	EDA	0.347545	-0.1795	0.932074
	Pulse Rate	0.316637	-0.20479	0.853623
	Temperature	0.209698	-0.32162	0.737244
4	EDA	-0.02913	-0.52006	0.511049
	Pulse Rate	-0.251052	-0.86183	0.283087
	Temperature	-2.046014	-2.79041	-1.6159
5	EDA	0.333029	-0.20766	0.893306
	Pulse Rate	2.163959	1.705	2.756521
	Temperature	-2.102134	-2.6564	-1.62448
6	EDA	0.023903	-0.49015	0.534756
	Pulse Rate	-0.197559	-0.78951	0.302591
	Temperature	-0.589378	-1.18137	-0.1286
7	EDA	0.201216	-0.3232	0.752456
	Pulse Rate	2.073197	1.505642	2.764266
	Temperature	-0.249412	-0.829	0.300268
8	EDA	-0.077495	-0.5819	0.489859
	Pulse Rate	0.822014	0.405876	1.220347
	Temperature	-0.957564	-1.29143	-0.57238
9	EDA	-1.885303	-2.72735	-1.20656
	Pulse Rate	0.038468	-0.60327	0.463783
	Temperature	-1.457263	-1.95483	-0.98055
10	EDA	-0.83616	-1.38334	-0.35839
	Pulse Rate	0.63936	0.097418	1.2171
	Temperature	-1.937372	-2.36652	-1.59867
11	EDA	-1.100321	-2.3537	-0.56637
	Pulse Rate	-0.503784	-1.13235	0.016751
	Temperature	-2.443558	-3.18736	-1.8778
12	EDA	-2.511189	-3.2005	-1.97447
	Pulse Rate	-1.313049	-2.02982	-0.69726
	Temperature	0.468322	-0.01906	1.078587
13	EDA	0.035349	-0.46564	0.511139
	Pulse Rate	2.911906	2.362222	3.64434
	Temperature	-0.326803	-0.90044	0.210485
14	EDA	0.035349	-0.46564	0.511139
	Pulse Rate	2.911906	2.362222	3.64434
	Temperature	-0.326803	-0.90044	0.210485

Table 4: Cohen's D Results for Male and Female Participants

Participant	Gender	Metric	Cohen's D	CI Low	CI High
1	Male	EDA	-1.547	-2.057	-1.246
		Pulse Rate	0.866	0.381	1.372
		Temperature	-0.634	-1.223	-0.109
3	Male	EDA	-0.359	-0.863	0.14
		Pulse Rate	-0.169	-0.462	0.429
		Temperature	-0.39	-0.835	-0.119
7	Male	EDA	0.35	-0.157	0.87
		Pulse Rate	2.169	1.682	2.779
		Temperature	-2.094	-2.68	-1.614
9	Male	EDA	0.198	-0.304	0.78
		Pulse Rate	2.066	1.532	2.725
		Temperature	-0.235	-0.797	0.267
_10	Male	EDA	-0.09	-0.57	0.432
		Pulse Rate	0.811	0.388	1.225
		Temperature	-0.956	-1.303	-0.568
2	Male	EDA	-1.855	-2.596	-1.219
		Pulse Rate	0.044	-0.618	0.466
		Temperature	-1.45	-1.961	-1.009
12	Male	EDA	-1.096	-2.331	-0.565
		Pulse Rate	-0.502	-1.112	-0.002
		Temperature	-2.466	-3.254	-1.885
14	Male	EDA	0.027	-0.528	0.531
		Pulse Rate	2.936	2.348	3.676
		Temperature	-0.341	-0.91	0.186
4	Female	EDA	0.367	-0.162	0.932
		Pulse Rate	0.328	-0.211	0.86
		Temperature	0.189	-0.333	0.741
5	Female	EDA	-0.033	-0.519	0.53
		Pulse Rate	-0.248	-0.847	0.297
		Temperature	-2.031	-2.645	-1.62
6	Female	EDA	1.396	0.250	3.043
		Pulse Rate	-2.610	-2.75	-2.47
		Temperature	0.382	0.140	0.423
8	Female	EDA	0.034	-0.492	0.563
		Pulse Rate	-0.182	-0.759	0.324
		Temperature	-0.57	-1.104	-0.056
11	Female	EDA	-0.808	-1.348	-0.327
		Pulse Rate	0.632	0.071	1.232
		Temperature	-1.937	-2.346	-1.614
13	Female	EDA	-0.808	-1.348	-0.327
		Pulse Rate	0.632	0.071	1.232
		Temperature	-1.937	-2.346	-1.614

Table 5: Mixed Linear Model Results for Temperature

Mixed Linear Model Regression Results: Temperature						
Model: MixedLM Dependent Variable: temperature_celsius						
No. Observations: 7228 Method: REML						
No. Groups: 14	Scale:	84.7764				
Min. group size: 61	Log-Likelihood:	-26306.7				
Max. group size: 6428	Converged:	Yes				
Mean group size: 516.3						
	Coef.	Std. Err.	z	P>|z|	[0.025	0.975]
Intercept	31.798	0.374	85.049	0	31.065	32.531
Gender[T.girl]	0.748	0.221	3.383	0.001	0.315	1.182
Group Var	0.713	0.099				

Table 6: Mixed Linear Model Results for Pulse Rate

Mixed Linear Model Regression Results: Pulse Rate						
Model: MixedLM Dependent Variable: pulse_rate_bpm						
No. Observations: 7228 Method: REML						
No. Groups: 14	Scale:	472.8996				
Min. group size: 61	Log-Likelihood:	-32529.4233				
Max. group size: 6428	Converged:	Yes				
Mean group size: 516.3						
	Coef.	Std. Err.	z	P>|z|	[0.025	0.975]
Intercept	83.589	2.310	36.185	0.000	79.061	88.116
Gender[T.girl]	0.646	0.539	1.198	0.231	-0.411	1.703
Group Var	66.988	1.346				

Table 7: Mixed Linear Model Results for EDA

Mixed Linear Model Regression Results: EDA						
Model: MixedLM Dependent Variable: eda_scl_usiemens						
No. Observations: 7228 Method: REML						
No. Groups: 14	Scale:	11.2280				
Min. group size: 61	Log-Likelihood:	-19024.4145				
Max. group size: 6428	Converged:	Yes				
Mean group size: 516.3						
	Coef.	Std. Err.	z	P>|z|	[0.025	0.975]
Intercept	1.244	0.727	1.712	0.087	-0.180	2.668
Gender[T.girl]	1.634	0.083	19.576	0.000	1.470	1.798
Group Var	7.209	0.869				

Table 8: Quantile Regression Results for EDA

Results for eda_scl_usiemens (Quantile Regression)						
Dependent Variable: eda_scl_usiemens						
Model: QuantReg						
Method: Least Squares						
Date: Saturday, 13 July 2024	Psuedo R-Squared	0.005317				
Time: 05:11:55	Bandwidth:	0.7388				
	Sparcity:	1.681				
	No. Observations:	800				
	Df Residuals:	798				
	Df Model:	1				
	Coef.	Std. Err.	t	P>|t|	[0.025	0.975]
Intercept	0.4000	0.038	10.591	0.000	0.326	0.474
Gender[T.girl]	-0.2000	0.061	-3.270	0.001	-0.320	-0.080

A Comparison of Oversampling Methods for Predicting Credit Card Default with Logistic Regression

Dara Tourt
Dara.tourt@my.metrostate.edu
Department of Management Information Systems
Metropolitan State University
Minneapolis, MN 55403 USA

Queen E. Booker
Queen.booker@metrostate.edu
Department of Management Information Systems
Metropolitan State University
Minneapolis, MN 55403 USA

Carl Rebman
carlr@sandiego.edu
Department of Management Information Systems
University of San Diego
San Diego, CA 92110 USA

Simon Jin
Simon.jin@metrostate.edu
Department of Management Information Systems
Metropolitan State University
Minneapolis, MN 55403 USA

Abstract

In the era of big data, the prevalence of imbalanced datasets has emerged as a significant challenge in machine learning and data analytics. Analysts often employ two primary techniques - undersampling and oversampling - to overcome the imbalance problem. This study explores multiple oversampling techniques in addressing these imbalances, focusing on how appropriate sampling methods can enhance model performance, improve predictive accuracy, and facilitate better decision-making. The results affirm that oversampling can improve the predictive power for the minority class when compared to building a model with imbalanced data. However, the additional contribution is that the type of balancing technique matters to the overall performance and accuracy of the predictive model.

Keywords: Data Balancing, Predictive Modeling, Logistic Regression, Credit Card Fraud.

Recommended Citation: Tourt, D., Booker, Q., Rebman Jr., C.M., Jin, S.S., (2025). A Comparative Analysis of Oversampling Methods for Predicting Credit Card Default with Logistic Regression. *Journal of Information Systems Applied Research and Analytics* v18, n2 pp 52-63. DOI# https://doi.org/10.62273/GLAH4676

A Comparison of Oversampling Methods for Predicting Credit Card Default with Logistic Regression

Dara Tourt, Queen E. Booker, Carl Redman and Simon Jin

1. INTRODUCTION

In the era of big data, the prevalence of imbalanced datasets has emerged as a significant challenge in machine learning and data analytics. Imbalanced datasets occur when one class significantly outnumbers another, which is common in financial modeling when addressing issues such as credit decisions, fraud detection, and default predictions (He & Garcia, 2009). There are many instances where the majority of instances in a dataset are much greater than the instances for the minority such as stroke prediction, loan defaults, credit card default, and credit card fraud. In these and other cases, the minority instances can be as few as 1% of all instances. Classification algorithms often perform inadequately on imbalanced datasets because they tend to favor the majority class, leading to high overall accuracy but poor sensitivity for the minority class (Saito & Rehmsmeier, 2015).

Analysts employ two primary techniques - undersampling and oversampling - to overcome the imbalance problem. Undersampling involves reducing the number of instances in the majority class to create a more balanced dataset. This technique can lead to simpler models that generalize better, as it prevents the model from becoming overwhelmed by the sheer volume of majority class instances (Kotsiantis, 2006; Dube & Verster, 2023). However, undersampling carries the risk of losing potentially valuable information, which can negatively impact model performance (Batista et al., 2004).

Conversely, oversampling increases the number of instances in the minority class. Techniques such as SMOTE (Synthetic Minority Over-sampling Technique) generate synthetic instances based on the existing minority class data, helping to mitigate the risk of overfitting associated with simple duplication of minority instances (Chawla et al., 2002). By enhancing the representation of the minority class, oversampling can significantly improve the model's ability to learn relevant patterns.

Given the emphasis prior research has made regarding the significance of balancing imbalanced datasets, many balancing methods have been introduced to accomplish the goal. However, few studies have compared and contrasted the various methods of balancing datasets and measured the difference in the performance of the different approaches. This research study aims to address this gap, applying different oversampling methods to the credit card default problem using a logistic regression model as the comparative tool. The research questions for the study are:

1. *Does oversampling improve the performance of the logistic regression predictive model for identifying potential credit card accounts that default?*
2. *Is there an oversampling method that improves the performance of the logistic regression predictive model for identifying potential credit card accounts that default?*

We report the difference in Type I and Type II errors and significant difference between each model built using the different balancing methods.

This research contributes to the larger body of research in several ways. First, it provides a summary of current oversampling techniques. Next, it compares multiple balancing techniques using one modeling method to demonstrate that how the data is balanced matters when using the models for predictions. The study is limited in that it only examines one classification problem with one modeling technique. This limitation means the results are not generalizable but it provides analysts with a process for improving classification modeling.

The paper continues with the literature review of oversampling methods, followed by the research methodology, results, conclusions, limitations, and next steps.

2. LITERATURE REVIEW

In the field of machine learning, addressing class imbalance remains a critical challenge that can significantly impact the performance of predictive models. This literature review explores various oversampling techniques

reported in the literature and used in data analysis, their theoretical foundations, and their benefits and challenges.

Random Oversampling

Random oversampling serves as a fundamental technique for addressing class imbalance in machine learning. By increasing the representation of minority class instances, random oversampling helps mitigate bias and improve the performance of predictive models on imbalanced datasets. Random oversampling involves randomly duplicating instances from the minority class until a balanced distribution is achieved (Chawla et al., 2002). Random oversampling is easy to implement and does not require complex algorithms or parameter tuning compared to other oversampling techniques. Random oversampling also retains all instances from both classes, thereby preserving the overall information content of the dataset. While simple, it may lead to overfitting and increased computational costs. Random oversampling is based on the premise of increasing the minority class instances randomly until the class distribution is balanced with the majority class. Random oversampling preserves all instances from both classes but duplicates minority class instances. By doing so, it aims to provide the model with more examples of the minority class, thereby reducing bias and improving the model's ability to generalize to minority class instances. (Yang et al., 2024)

Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE generates synthetic instances for the minority class by interpolating between existing instances (Chawla et al., 2002). This technique preserves the underlying data structure better than random oversampling and reduces the risk of overfitting. SMOTE, proposed by Chawla et al. (2002), tackles class imbalance by generating synthetic instances for the minority class. It works by interpolating between existing minority class instances to create new synthetic samples in the feature space, thereby balancing the dataset without blindly duplicating existing data points.

This method is effective in improving the generalization ability of machine learning models by providing more balanced training data. Despite its advantages, SMOTE may struggle with datasets where the minority class is not uniformly distributed or when instances of the minority class overlap with those of the majority class. This can lead to synthetic samples that do not adequately represent the true characteristics of the minority class, potentially affecting model performance. (Kimbrell, 2014)

Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC)

SMOTE-NC is used for datasets that contain both numerical and categorical features. SMOTE-NC first identifies instances belonging to the minority class. For each minority instance, the algorithm calculates the k-nearest neighbors (typically using Euclidean distance). Synthetic samples are generated by taking a minority instance and one of its nearest neighbors. A random point is then created along the line segment joining the two instances. This process is repeated until the desired level of balance is achieved in the dataset. SMOTE-NC uses the most common value among the k-nearest neighbors for categorical attributes when generating synthetic samples. Instead of calculating a weighted average (as with continuous data), it identifies the mode (most frequent category) for categorical features. The synthetic instance is formed by combining the continuous attributes generated as described earlier and the most frequent values for nominal attributes. (Gök et al., 2021)

Safe-level-SMOTE (SL-SMOTE)

SL-SMOTE builds upon the original SMOTE framework by incorporating a safety level mechanism to control the generation of synthetic samples. SL-SMOTE evaluates the density of instances surrounding minority class samples. It utilizes the concept of safe regions, where synthetic samples can be generated without risking overfitting or misclassifying noisy data points. Instead of randomly selecting neighbors as in traditional SMOTE, SL-SMOTE identifies "safe neighbors" that lie within a certain threshold of distance from the minority instance.

This selective approach minimizes the chances of generating synthetic samples that may lead to decision boundary distortions (He & Ma, 2009). Once safe neighbors are identified, synthetic samples are generated similarly to traditional SMOTE, using linear interpolation. By focusing on safe regions for sample generation, SL-SMOTE significantly reduces the risk of overfitting to noisy or outlier data points that may exist within the minority class. Several studies have demonstrated that SL-SMOTE can lead to improved classification performance compared to traditional SMOTE, particularly in scenarios with extreme class imbalance (Shing et al., 2023).

Borderline-SMOTE (BSMOTE)
This variant of SMOTE focuses on generating synthetic instances near the decision boundary between classes (Han et al., 2005). It addresses classification errors that occur near the class boundaries. Borderline SMOTE focuses on generating synthetic samples specifically in the "borderline" areas where the minority class instances are most vulnerable to misclassification (Han et al., 2005). The algorithm first identifies minority class instances that lie close to the decision boundary between classes. These borderline instances are critical because they are often misclassified or underrepresented, making them essential for model training. For each borderline minority instance, the algorithm identifies its k-nearest neighbors within the minority class. The choice of k can be adjusted based on the dataset's characteristics. Synthetic samples are generated by interpolating between a borderline instance and its nearest minority neighbors. The interpolation is performed similarly to traditional SMOTE, using a weighted combination of the instances to create new synthetic samples in the feature space. (Han et al., 2004; Chen et al., 2023).

K-Means SMOTE
K-Means SMOTE is an approach that integrates K-Means clustering with SMOTE to enhance the representation of the minority class. By leveraging K-Means clustering, K-Means SMOTE generates synthetic instances that better capture the distribution and structure of the minority class, leading to improved model performance (Batista et al., 2004). The localized generation of synthetic instances helps reduceoverfitting by preserving the diversity within the minority class and avoiding excessive duplication of instances (Sun et al., 2007). Models trained on datasets augmented with K-Means SMOTE synthetic samples are better able to generalize to unseen data, as they have learned from a more balanced and representative dataset (He & Ma, 2013).

Support Vector Machine SMOTE (SVM SMOTE)
To enhance the performance of predictive models on imbalanced datasets, SVM SMOTE has emerged as an advanced approach that integrates Support Vector Machines (SVM) with SMOTE to improve the representation of minority class instances. SVMs are powerful supervised learning models used for classification tasks. SVM SMOTE integrates the strengths of SVM and SMOTE by selectively applying the oversampling technique to minority class instances that are support vectors or are close to the SVM decision boundary. By focusing on instances near the SVM decision boundary, SVM SMOTE generates synthetic samples that are more relevant to the SVM classifier's learning process, thereby enhancing its ability to generalize (He & Ma, 2013). SVM SMOTE aims to mitigate the impact of class imbalance on SVM classifiers, resulting in improved accuracy and robustness in predicting minority class instances (Sun et al., 2007). Empirical studies and applications across various domains, such as healthcare diagnostics, fraud detection in finance, and image classification, have demonstrated the efficacy of SVM SMOTE in addressing class imbalance and improving predictive model performance (Batista et al., 2004; Zhang & Mani, 2003).

Adaptive Synthetic Sampling (ADASYN)
ADASYN adjusts the density distribution of the minority class by focusing synthetic instance generation on instances that are harder to classify (He & Ma, 2013). It emphasizes regions of the feature space where the classifier performs poorly. ADASYN is another extension of Synthetic Minority Over-sampling Technique (SMOTE), which addresses class imbalance by oversampling the minority class. SMOTE generates synthetic samples along line segments joining minority class instances. However, SMOTE does not consider the distribution of minority class instances, potentially leading to overfitting in dense minority regions and underfitting in sparse regions. ADASYN improves upon SMOTE by adaptively generating synthetic samples based on the density distribution of minority class instances. Specifically, it focuses more on generating samples in regions where the class distribution is sparser, thereby making the classifier more robust and reducing the risk of overfitting. (Mitre et al., 2023)

Self-adaptive Oversampling (SAOM)
The Self-Adaptive Oversampling Method (SAOM) introduces a dynamic approach to the oversampling process, allowing it to adjust based on the characteristics of the data at hand. Unlike static oversampling techniques that apply a uniform strategy across the dataset, SAOM adapts its sampling strategy according to the local distribution of minority and majority classes, thereby enhancing the quality of the synthetic samples generated. SAOM continuously evaluates the data distribution and adjusts the oversampling strategy based on local density estimates.

By incorporating an adaptive mechanism, SAOM strikes a balance between exploring underrepresented regions of the feature space and exploiting areas where the minority class is already well-represented. This dual strategy improves the diversity of synthetic samples while ensuring they remain relevant to the underlying data distribution. Studies have shown that models trained using SAOM exhibit superior performance compared to those utilizing traditional oversampling methods. The self-adaptive mechanism allows SAOM to be tailored to a wide range of applications and datasets, making it a versatile tool in the machine learning toolbox. Its scalability ensures that it can be applied effectively in both small and large datasets. (Tao et al., 2023)

3. RESEARCH METHODOLOGY

The purpose of this study was to evaluate the performance of a predictive modeling technique using different oversampling techniques. The application for the study is logistic regression to predict customer default on credit card payments. There are many machine learning methods used to predict default behavior. Logistic regression was used for this study to explore the oversampling method because, according to Yeh and Lien (2009) and Sperandei (2014), logistic regression is specifically tailored for scenarios with a binary response variable and is typically the first or baseline technique to compare subsequent models for performance. Logistic regression's strength lies in its ability to offer a straightforward probabilistic framework for classification.

The study compares eight different oversampling methods to the imbalanced dataset. The oversampling techniques studied were Random Over-Sampling, SMOTE, SMOTENC, ADASYN, BSMOTE, SVM SMOTE, K-Means SMOTE, and SL-SMOTE.

The research questions for the study were:

1. *Does oversampling improve the performance of the logistic regression predictive model for identifying potential credit card accounts that default?*
2. *Is there an oversampling method that improves the performance of the logistic regression predictive model for identifying potential credit card accounts that default?*

Based on the literature, oversampling methods improve the performance of data mining algorithms. However, there was no indication through the literature review process that any method significantly outperformed another for the credit card default application. Thus, the hypotheses for the study are as follows:

H0: A logistic regression model for predicting credit card payment default built using an imbalanced dataset will not perform significantly better than a model built using a balanced dataset.

H1: When comparing oversampling methods to balance the dataset, no logistic regression model performs significantly better than another.

All models were built and evaluated using Python. To test the significant difference between each model, a t-test was performed comparing error rates. Each model was evaluated using standard suitability measures. According to the literature, there is general agreement how they are defined and are listed as follows (Chen et al., 2021; Demraoui et al., 2022; Karthiban et al., 2019; Lusinga et al., 2021; Li et al., 2017; Ndayisenga, 2021; Orji et al.; Peiris, 2022; Pimcharee & Surinta, 2022, Booker & Rebman, 2024):

- Accuracy score over 90%
- Specificity score over 85%
- Type I Error score under 10%
- Type II Error score under 10%
- Recall score over 85%
- Precision score over 85%
- F measure score over 85
- AUC near to 1

Dataset

The dataset used in the study contains information on customer default payments in Taiwan. Figure 1 illustrates that the number of accounts not expected to default the following month vastly outnumbers those that are at risk of default and shows the class imbalance between defaults and non-defaults, with 6,636 accounts classified as defaults and 23,364 as non-defaults. It is a multivariate dataset with 30,000 instances and 23 features, including both categorical and nominal data types. The dataset is hosted by the UCI Machine Learning Repository and can be accessed directly at https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

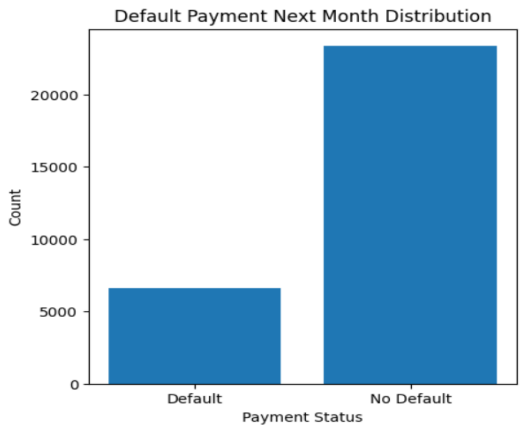

Figure 1: Default Instances in the Dataset

The creators of this dataset, led by I-Cheng Yeh, compiled it for business applications, specifically within the subject area of risk management associated with credit card default payments. The dataset does not contain any missing values. Variables were recoded as necessary to ensure categorical data was represented as binary variables. The decision variable was whether a customer defaulted with 1 for defaulted and 0 for non-default.

The variables in the dataset were:

- X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (1 = graduate school; 2 = university; 3 =high school; 4 = others).
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6–X11: History of past payment. We tracked the past monthly payment records (from April to September 2005) as follows:
    - X6 = the repayment status in September, 2005; X7=the repayment status in August, 2005;...;X11 = the repayment status in April, 2005.
        - The measurement scale for the repayment status is:
            - -1 =no pay delay
            - 1=payment delay for one month
            - 2 =payment delay for two months...;
            - 8 = payment delay for eight months;
            - 9 = payment delay for nine months and above.

  X12–X17: Amount of bill statement (NT dollar).
    - X12 =amount of bill statement in September, 2005; X13 =amount of bill statement in August, 2005;...;X17 = amount of bill statement in April, 2005.
- X18–X23: Amount of previous payment (NT dollar).
    - X18 =amount paid in September, 2005; X19 = amount paid in August, 2005;...;X23 = amount paid in April, 2005.

Model Development

Each model was built using 10,000 instances. For the imbalanced dataset, all the observations were drawn from the dataset. Following the recommendation of Gholamy et al. (2018), the training used 80% of the dataset. The remaining 20% was used for testing. All models were built using logistic regression which is a widely used statistical method for predictive modeling, particularly suited for binary classification tasks. It models the relationship between one or more independent variables (features) and a binary dependent variable (outcome) using a logistic function. Logistic regression is designed to predict the probability of a binary outcome, typically coded as 0 and 1. Each model, including the application of sampling techniques were built using the Python software application.

4. RESULTS

This section presents the results of the validation stage of the analysis. Each model was applied to the full dataset. Table 1 summarizes the performance metrics of accuracy, precision, recall, and F-measure across the different oversampling methods using the results from the validation of the models.

Based on the results, overall the imbalanced model appears to perform better than the oversampling models except for ROS. However, each of the oversampling methods had at least two measures that met the suitability standards, and each oversampling model performed better

than the imbalanced model in precision. Random oversampling met all four standards. In comparing the suitability measures, it would seem that the random oversampling model would provide the best predictive power.

Model	Precision	Recall	Accuracy	F1 Score
Imbalanced	0.8631	0.8917	0.8055	0.8771
Random Over-Sampling	0.9729	0.9649	0.9517	0.9689
SMOTE	0.9658	0.7771	0.8050	0.8612
SMOTENC	0.9677	0.7806	0.8089	0.8642
ADASYN	0.9662	0.7803	0.8076	0.8634
Borderline SMOTE	0.9665	0.7793	0.8070	0.8628
SVM SMOTE	0.9676	0.7774	0.8064	0.8621
KMeans SMOTE	0.9667	0.7757	0.8045	0.8607
SL-SMOTE	0.9066	0.7803	0.7662	0.8387

Table 1: Validation Data Results Logistic Regression on Various Over-sampling Methods to Deal with Imbalance Class (Default, Non-Default)

However, a review of the confusion matrices in Figures 2 through 10 show that the imbalanced model predicts the majority instances well but falters when predicting the defaults, providing a 50/50 predictive power. For credit card default, a client is likely interested in having more potential default cases predicted than fewer.

In examining the matrices, the imbalanced model has the worst performance with regards to correctly identifying default instances, predicting approximately 50% of the instances correctly. The best method of those tested was SVM-SMOTE, correctly identifying more than 90% of the default instances. However, the model with the best predictive power for the majority class was random oversampling.

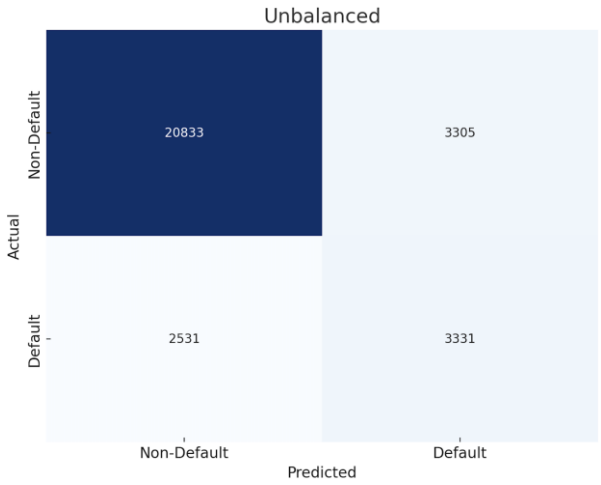

Figure 2: Imbalanced Confusion Matrix

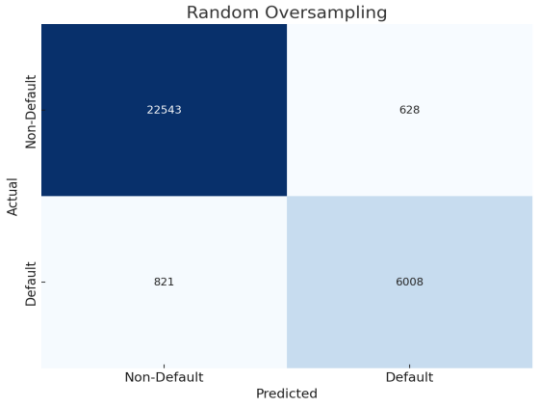

Figure 3: Random Oversampling Confusion Matrix

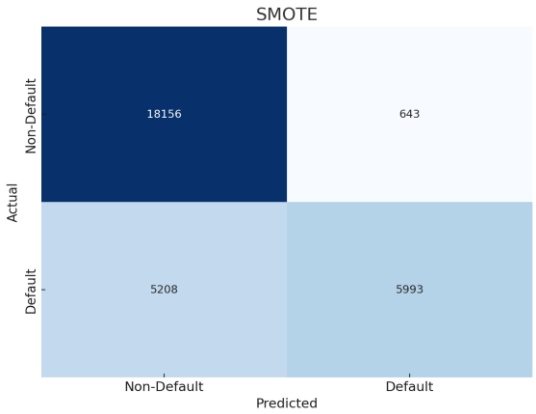

Figure 4: SMOTE Confusion Matrix

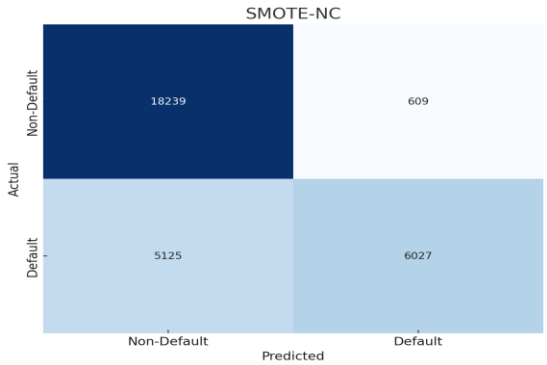

Figure 5: SMOTE-NC Confusion Matrix

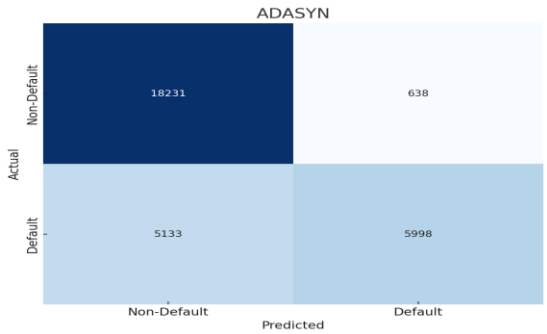

Figure 6: ADASYN Confusion Matrix

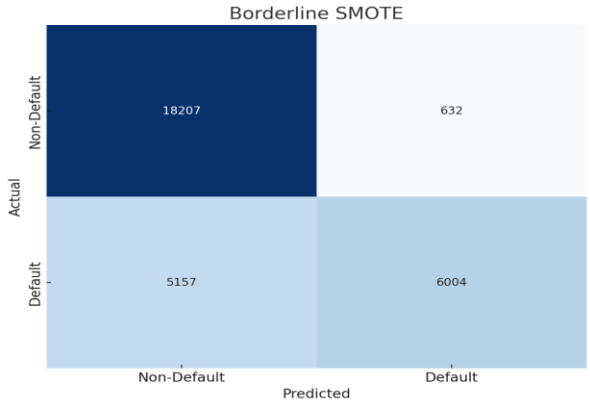

Figure 7: Borderline SMOTE Confusion Matrix

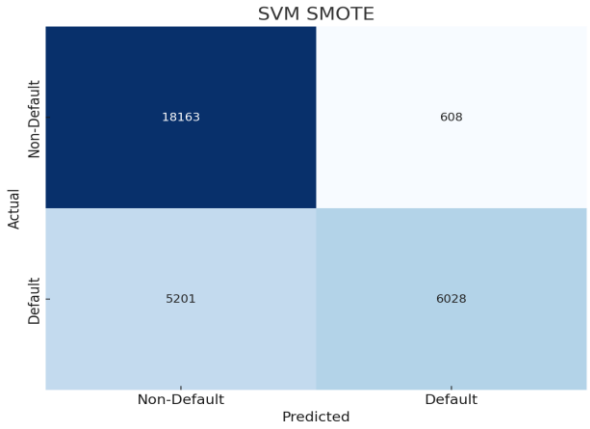

Figure 8: SVM SMOTE Confusion Matrix

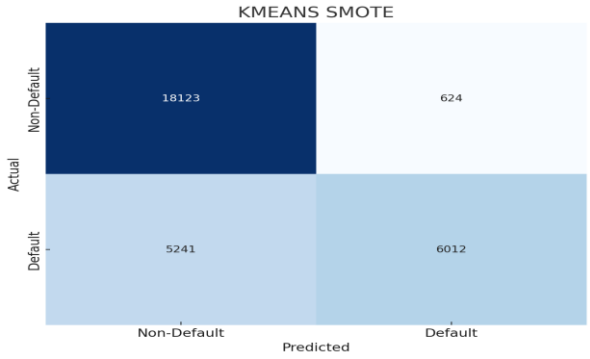

Figure 9: KMEANS SMOTE Confusion Matrix

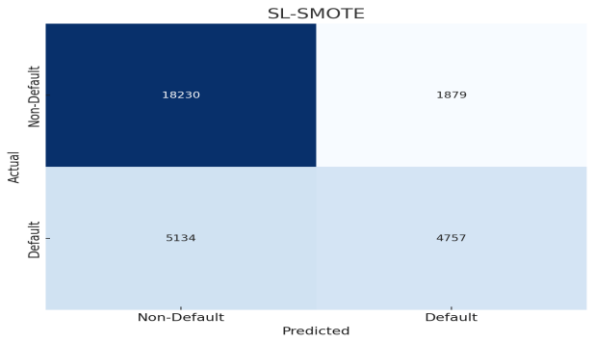

Figure 10: SL-SMOTE Confusion Matrix

The suitability measures and the confusion matrices indicate that the oversampling models perform better than the imbalanced model when predicting default instances. The next step was to determine if the differences were significant. Paired t-tests were performed for the imbalanced model and each of the oversampling models, and between each of the oversampling models. The results are shown in Table 2 in Appendix A.

Based on the t-tests, all the models that used oversampling methods performed with significant difference from the model using an

imbalanced dataset when evaluating the prediction for default. Within the oversampling methods, SL-SMOTE was significantly different from the other methods, with SL-SMOTE performing worse rather better.

The final step in the analysis was to evaluate the hypotheses and research questions. Recall the primary research hypotheses:

H0: *A logistic regression model for predicting credit card payment default built using an imbalanced dataset will not perform significantly better than a model built using a balanced dataset.*

H1: *When compared to other oversampling methods to balance the dataset, no logistic regression model performs significantly better than another.*

H0 is accepted because all the oversampling models performed better, based on the t-test results, on precision. H1 is partially accepted as the SL-SMOTE performed better than the other models when comparing correct vs incorrect predictions for both minority and majority instances.

When returning to the research questions RQ1 "*does oversampling improve the performance of the logistic regression predictive model for identifying potential credit card accounts that default?" and RQ2 "is there an oversampling method that improves the performance of the logistic regression predictive model for identifying potential credit card accounts that default?"*, the results indicate that oversampling does improve the performance of the logistic regression predictive model for identifying potential credit card accounts that default and of the oversampling methods tested, all the other models performed better than the SL-SMOTE method.

In summary, the results indicate that there is value in comparing data balancing methods when developing predictive modeling as such a comparison can improve the performance of the predictive model.

5. LIMITATIONS AND CONCLUSIONS

This study examined only oversampling methods in the context of predicting credit card default for a specific dataset using a specific modeling method-logistic regression. The results of the study cannot be generalized as there are many factors to consider when building predictive models including but not limited to the variables, data balancing methods, and predictive modeling techniques. Therefore, additional analysis is needed to determine the conditions best suited for each sampling method, dataset configuration, and predictive modeling tool.

However, oversampling techniques represent a critical approach to addressing class imbalance in data analysis. While each technique has its strengths and weaknesses, their application depends heavily on the specific characteristics of the dataset and the objectives of the analysis. Continued research and development in this area aim to improve the robustness, scalability, and applicability of oversampling methods across diverse domains and applications in machine learning and statistical modeling. Oversampling serves as a viable strategy to address the challenges posed by imbalanced datasets. The selection of the appropriate method hinges on the specific requirements of the task, the nature of the dataset, and the criticality of predictive accuracy in the minority class. As machine learning continues to evolve, ongoing research into sampling approaches that combining the strengths of multiple methods may provide further avenues for improvement in managing imbalanced datasets.

6. REFERENCES

Batista, G. E. A. P. A., Monard, M. C., & Silva, J. C. P. (2004). A study of data preprocessing and classifiers for imbalanced datasets. *Proceedings of the Brazilian Conference on Neural Networks*. https://doi.org/10.5220/0005201103820389

Batista, G.E., Prati, R. C. & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20–29. https://doi.org/10.1145/1007730.1007735

Booker, Q. & Rebman, C. (2024). Applying Heterogeneous Ensemble Models to Detect Credit Card Fraudulent Transactions. *Proceedings of the 2024 Southwest Decision Sciences Conference (SWDSI),* Galveston, TX.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced

problem. In Advances in knowledge discovery and data mining: 13th Pacific-Asia conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 proceedings 13 (pp. 475-482). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-01307-2_43

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2011) DBSMOTE: Density-based synthetic minority over-sampling technique. Applied Intelligence, vol. 36, pp. 1–21. https://doi.org/10.1007/s10489-011-0287-y

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

Chen, C., Shen, W., Yang, C., Fan, W., Liu, X., & Li, Y. (2023). A New Safe-Level Enabled Borderline-SMOTE for Condition Recognition of Imbalanced Dataset. IEEE Transactions on Instrumentation and Measurement, 72, 1–10. https://doi.org/10.1109/TIM.2023.3289545

Chen, R.C., Luo, S.T., Liang, X., and Lee, V. (2005). Personalized Approach Based on SVM and ANN for Detecting Credit Card Fraud. International Conference on Neural Networks and Brain, IEEE, 810-815. https://doi.org/10.1109/icnnb.2005.1614747

Demraoui, L., Eddamiri, S., & Hachad, L. (2022). Digital Transformation and Costumers Services in Emerging Countries: Loan Prediction Modeling in Modern Banking Transactions. Lecture Notes on Data Engineering and Communications Technologies, 627–642. https://doi.org/10.1007/978-3-030-90618-4_32

Dube, L. & Verster, T. (2023). Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models. Data Science in Finance and Economics. 3. 354-379. https://doi.org/10.3934/DSFE.2023021.

Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F. (2018). Foundations on Imbalanced Classification. In: Learning from Imbalanced Data Sets. Springer, Cham. https://doi.org/10.1007/978-3-319-98074-4_2

Gholamy, A., Kreinovich, V., and Kosheleva, O. (2018) "Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation" (2018). Departmental Technical Reports (CS). 1209. https://scholarworks.utep.edu/cs_techrep/1209

Gök, E. C., & Olgun, M. O. (2021). SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples. Neural Computing & Applications, 33(22), 15693–15707. https://doi.org/10.1007/s00521-021-06189-y

Han, H., Wang, W. Y., & Mao, B. H. (2005) Borderline-smote: A new over-sampling method in imbalanced data sets learning," in Advances in Intelligent Computing, (Hefei, China), vol. 3644, pp. 878–887, Springer-Verlag. https://doi.org/10.1007/11538059_91

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284. https://doi.org/10.1109/tkde.2008.239

He, H., & Ma, Y. (Eds.). (2013). Imbalanced learning: foundations, algorithms, and applications. https://doi.org/10.1002/9781118646106

Karthiban, R., Ambika, M., & Kannammal, K. E. (2019). A Review on Machine Learning Classification Technique for Bank Loan Approval. International Conference on Computer Communication and Informatics. https://doi.org/10.1109/iccci.2019.8822014

Kimbrell, J. (2014). Smote. Ploughshares, 40(1), 137–138. https://doi.org/10.1353/plo.2014.0004

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 30(3), 301-320. https://doi.org/10.1007/s10462-007-9052-3

Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject Inference in Credit Scoring Using Semi-Supervised Support Vector Machines. Expert Systems with Applications, 74, 105–114. https://doi.org/10.1016/j.eswa.2017.01.01

1

Lusinga, M., Mokoena, T., Modupe, A., & Mariate, V. (2021). Investigating Statistical and Machine Learning Techniques to Improve the Credit Approval Process in Developing Countries. AFRICON. https://doi.org/10.1109/africon51333.2021.9570906

Mitra, R., Bajpai, A., & Biswas, K. (2023). ADASYN-assisted machine learning for phase prediction of high entropy carbides. Computational Materials Science, 223, 112142-. https://doi.org/10.1016/j.commatsci.2023.112142

Naboureh, A.; Li, A.; Bian, J.; Lei, G.; Amani, M. A Hybrid Data Balancing Method for Classification of Imbalanced Training Data within Google Earth Engine: Case Studies from Mountainous Regions. *Remote Sens.* 2020, *12*, 3301. https://doi.org/10.3390/rs12203301

Ndayisenga, T. (2021). Bank Loan Approval Prediction Using Machine Learning Techniques. [Doctoral dissertation, University of Rwanda]. http://www.dr.ur.ac.rw/handle/123456789/1437

Orji, U. E., Ugwuishiwu, C. H., Nguemaleu, J. C. N., & Ugwuanyi, P. O. (2022). Machine Learning Models for Predicting Bank Loan Eligibility. 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON). https://doi.org/10.1109/nigercon54645.2022.9803172

Peiris, M. P. C. (2022). Credit Card Approval Prediction by Using Machine Learning Techniques [Doctoral dissertation, University of Colombo School of Computing]. https://dl.ucsc.cmb.ac.lk/jspui/handle/123456789/4593

Pimcharee, K., & Surinta, O. (2022). Data Mining Approaches in Personal Loan Approval. Engineering Access, 8(1), pp. 15-21. doi: 10.14456/mijet.2022.2. https://ph02.tci-thaijo.org/index.php/mijet/article/view/244392

Saito, K., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Shing, T. W., Sudirman, R., Daud, S. N. S. S., Razak, M. A. A., Zakaria, N. A., & Mahmood, N. H. (2023). Multistage Anxiety State Recognition based on EEG Signal using Safe-Level SMOTE. Journal of Physics. Conference Series, 2622(1), 12010-. https://doi.org/10.1088/1742-6596/2622/1/012010

Sperandei S. (2014). Understanding Logistic Regression analysis. Biochemia medica, 24(1), 12–18. https://doi.org/10.11613/BM.2014.003

Sun, L., Hu, N., Ye, Y., Tan, W., Wu, M., Wang, X., & Huang, Z. (2022). Ensemble stacking rockburst prediction model based on Yeo–Johnson, K-means SMOTE, and optimal rockburst feature dimension determination. Scientific Reports, 12(1), 15352–15352. https://doi.org/10.1038/s41598-022-19669-5

Tao,X., Guo,X., Zheng, Y., Zhang,X, & Chen, Z. (2023) Self-adaptive oversampling method based on the complexity of minority data in imbalanced datasets classification. Know.-Based Syst. 277, C. https://doi.org/10.1016/j.knosys.2023.110795

Yeh, I., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst. Appl*., 36, 2473-2480. https://doi.org/10.1016/j.eswa.2007.12.020

Yang, C., Fridgeirsson, E. A., Kors, J. A., Reps, J. M., & Rijnbeek, P. R. (2024). Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. Journal of Big Data, 11(1), 7–17. https://doi.org/10.1186/s40537-023-00857-7

Appendix A

Model	Imbalanced	Random Over-Sampling	SMOTE	SMOTENC	ADASYN	Borderline SMOTE	SVM SMOTE
Random Over-Sampling	56.544 *						
SMOTE	56.239 *	0.444					
SMOTENC	57.518 *	0.572	1.006				
ADASYN	56.357 *	0.295	0.146	0.864			
Borderline SMOTE	57.346 *	0.118	0.323	0.689	0.177		
SVM SMOTE	57.173 *	0.598	1.032	0.030	0.905	0.720	
KMeans SMOTE	56.745 *	0.118	0.569	0.453	0.418	0.238	0.578
SL-SMOTE	25.838 *	28.810 *	28.134 *	29.390 *	28.612 *	28.737 *	39.647 *

Table 2: T-test Results Comparing Accuracy of the Models

Emergent Technologies Production in the US: Exploratory Analysis of Motivations and Adverse Factors

Katarzyna Toskin
toskink1@southernct.edu
Southern Connecticut State University
New Haven, CT 06515

Marko Jocic
mjocic@kennesaw.edu
Kennesaw State University
Kennesaw, GA 30144

Abstract

New technologies can provide substantial business opportunities, and many firms are currently working on adopting them in their organizations. Prior literature provides insight and guidance to help firms navigate the technology adoption process, but there is limited information about companies that supply or produce newer technologies in the US market. Therefore, this study analyzed the extent to which US firms produce or use emergent technologies, the motivating factors to do so, and the reasons that impede their progression. Findings reveal that the share of technology producers is proportional to the share of users based on each technology group. Additionally, a majority of US companies that produce technologies report upgrading of goods and services, expanding the range of goods and services, and increasing or maintaining market share as the top motivating factors for producing emergent technology or products/services that include such technology. Furthermore, the producers reported high costs as the top adverse reason for generating emergent technology. These findings provide new insight into firms that produce technologies and have direct implications for business strategists, as well as policymakers.

Keywords: Artificial intelligence, cloud-based, robotics, technology, innovation, adoption, production

Recommended Citation: Toskin, K., Jocic, M., (2025). Emergent Technologies Production in the US: Exploratory Analysis of Motivations and Adverse Factors. *Journal of Information Systems Applied Research and Analytics* v18, n2 pp 52-63. DOI# https://doi.org/10.62273/DRXP6422.

Emergent Technologies Production in the US: Exploratory Analysis of Motivations and Adverse Factors

Katarzyna Toskin and Marko Jocic

1. INTRODUCTION

A significant effort has been devoted to examining factors contributing to new technology adoption or impeding organizational progress toward innovation. More specifically, prior studies have investigated motivations and barriers to various firms' using or adopting emergent technologies (Bunte et al., 2021; Cubric, 2020). Some studies have also assessed the current adoption and use level of advanced technologies within US businesses (Acemoglu et al., 2022, 2024). However, the primary focus of these studies has been on the early adopters or technology users only. Little attention has been given to the producers or the suppliers of such technologies in the US market.

Therefore, the purpose of this study is to expand the current literature by investigating the share of US technology producers and users as well as motivations, and factors that adversely affect the production or usage of emergent technologies by US firms. This group provides a unique insight as it represents US companies that have already progressed through the initial phase of the technology adoption curve and possess hands-on experience creating and delivering innovative technologies to the market. Understanding such factors will not only shed new light on this group of companies but also provide understanding and a chance to address tech suppliers' motives, opportunities, and needs.

2. LITERATURE REVIEW

The main reason organizations adopt or use advanced technologies such as AI is to increase performance through improved operational efficiency (Bhalerao, Kumar, Kumar, & Pujari, 2022). Prior literature reports that about 29.5 % of Small and Medium-sized Enterprises (SMEs) were open to adopting AI applications (Bhalerao et al., 2022). More specific reasons include innovation, increased productivity and efficiency of business processes, reduced human error, improved decision-making, and predictive capabilities (Cubric, 2020). However, Acemoglu et al. (2022, 2024) noted that current adoption in US firms is still minimal, especially for AI, with only 3.2% of companies using AI and only 2% of companies using robotics during the 2016–2018 time period. Although the low technology adoption rates have been highlighted in the current literature, along with the motivations and challenges, we know very little about US firms that produce or supply this technology.

The diffusion of innovation theory (Rogers, 1995) posits that adoption occurs after innovation is communicated in several phases through the social system. It consists of innovators, early adopters, early majority, late majority, and laggards, and it resembles an S-shaped distribution when the number of adopters from each category is mapped over time (Lai, 2017). Hall & Khan (2003) discussed the factors affecting technology diffusion. The authors highlighted that adoption is affected not only by demand but also by the suppliers of the new technology. The demand is primarily driven by the perceived benefits and the cost of adoption. On the other hand, the suppliers' role relates to addressing improvements of the new technology, which might initially be imperfect, and lowering the cost of new technologies over time (Rosenberg, 1972). Another important factor relates to "complementary inputs," which involves the suppliers' ability to offer training courses to upskill labor on the demand side (Hall & Khan, 2003). This information provision and knowledge transfer builds the firm's potential to use and ultimately adopt the new technology.

Considering that the adoption rate for AI and other emergent technologies within the US firms is still very low, we posit that this is in part due to the low share of supplying firms available in the US market. Hence, we form our first hypothesis as follows:

H1: The share of US firms that are suppliers is proportional to the share of adopters by each technology group.

Motivation to adopt emergent technologies can vary across companies and industries, but ultimately, the key reason for the adoption of innovative technology is to increase organizational performance (Hameed, Counsell, and Swift, 2012). Acemoglu et al. (2022)

estimated that the use of advanced technologies could lead to an 11.4% higher labor productivity because automation can replace manual labor and increase the efficiency of many internal processes. However, the motivations for companies that produce AI and other emergent technologies could be different primarily because selling or supplying innovative technology becomes part of the firm's value proposition. Hence, their business model is built around solving customers' problems with emergent technologies and establishing new customer segments and markets. As the motivations for producers of emergent technologies are likely to differ from technology adopters, we form our second hypothesis.

H2: The motivation of producers has a more strategic focus targeted externally towards the market, whereas users of emergent technologies focus on improving and upgrading internal processes.

The adoption of new technologies does not come without challenges. The most prominent adverse reasons identified by these firms were the lack of applicability and the high costs of deploying and integrating these technologies. Similar findings were reported by McElheran et al. (2024), who investigated the adoption and diffusion of technologies associated with AI, such as automated guided vehicles, machine learning, machine vision, natural language processing, and voice recognition. Their study revealed that fewer than 6 % of US firms utilized any of those technologies as of 2018.

Additionally, Cubric (2020) analyzed 30 published reviews involving AI adoption across various business sectors. The author reported that in addition to economic reasons such as high cost, AI adoption was negatively affected by technical and social aspects. Technical aspects included a lack of suitable data and limited reusability of models. Social reasons included a lack of expertise in this field, such as not understanding AI capabilities, which led to unrealistic expectations. Other social factors included stakeholder's perspective, distrust in the technology, fears related to its safety, and job insecurity. Similarly, Bunte et al. (2021) conducted interviews with 68 German companies and reported several challenges associated with the application of AI in SMEs. The reasons ranged from some participating companies feeling they were too small for AI to other companies evaluating the potential use of AI. Also, the lack of sufficient expertise, the extended amortization period, and the different priorities for capital expenditures were listed as challenges. Additionally, some companies felt that AI did not offer enough potential for organizational improvements. To alleviate some of these challenges, further efforts have been made by formulating guidelines, solutions, and best practices to help organizations expedite the technology adoption process (Bunte et al., 2021; McKinsey, 2019). Due to the financial reasons being highlighted as the most prominent adverse factor impacting technology adopters, in our final hypothesis, we posit that cost is also one of the largest challenges facing suppliers.

H3: The adverse factors for technology producers are similar to adverse factors of technology adopters, with high cost being one of the primary reasons.

Considering the critical role of technology suppliers in the adoption and diffusion process, this study extends current research by investigating the motivations and barriers of companies that produce emerging technologies or use them for their goods or services. This specific subset of companies offers unique insight into the motivations and challenges due to their firsthand experience with these technologies in the US market, which might shed additional light on the diffusion process and low adoption rates of the advanced technologies among US firms.

3. METHOD

This study used the US Census Annual Business Survey (ABS) data collected in 2019 (United States Census Bureau, 2019) and reported on three years from 2016 - 2018. This is the latest data set available that contains valuable information about both technology users and producers. We used the surveys regarding the extent to which US firms produced and used emergent technologies, the motivations for doing so, and the factors that adversely impacted technology production or adoption.

The exact names of the data set used for production were titled Annual Business Survey: Technology Production in Employer Firms by 2-digit NAICS for the United States and States: 2018, Annual Business Survey: Motivation to Produce Technology of Employer Firms by Sex, Ethnicity, Race, Veteran Status, and Employment Size for the United States: 2018," and "Annual Business Survey: Factors Adversely Affecting Technology Production in Employer Firms by Sex, Ethnicity, Race, Veteran Status, and Employment Size for the United States: 2018."

For usage, the datasets were titled "Annual Business Survey: Extent of Technology Use of Employer Firms by Sex, Ethnicity, Race, Veteran Status, and Employment Size for the United States: 2018," "Annual Business Survey: Motivation for Technology Use of Employer Firms by Sex." and "Annual Business Survey: Factors Adversely Affecting Technology use in Employer Firms by 2-digit NAICS in the United States and States: 2018."

They contained information about firms with paid employees and receipts of $1,000 or more grouped by industry using a 2-digit NAICS (North American Industry Classification System) code. The data set also included aggregate data for all sectors. Since the Census suppresses specific data to maintain confidentiality, only the aggregate-level data (totals) were available for analysis.

Measures

The ABS survey captured information about the following five technology groups:

- Artificial Intelligence
- Cloud-based
- Robotics
- Specialized Software
- Specialized Equipment

In the technology production part of the survey, the participants were asked whether their business sold one of the technologies or sold goods or services that included one of such technologies. The choices provided for respondents included yes, no, or don't know. Those participants who answered no or don't know were asked to skip the Motivations section and progress to the end of the survey to answer the questions regarding "Factors Adversely Affecting Technology Production." Otherwise, the participants who responded yes were directed to the next section, which captured the factors that motivated the technology production in these firms.

For the technology adoption portion of the survey, the participants were asked to what extent their business used one of the technologies in production processes for goods or services. The response options included: did not use, tested but did not use in production or service, low use, moderate use, high use, and don't know. Those participants who answered that they did not use, tested but did not use in production or service, or didn't know were asked to skip the Motivations section and progress directly to the "Factors Adversely Affecting Technology Adoption and Utilization" section. Responses with all forms of usage levels (i.e., low, moderate, and high) were counted towards the yes category to compute the technology usage metrics.

Table 1 lists the total number of employer firms that participated in each part of the production survey (dataset) broken down by technology group.

Technology Group	Production	Motivation*	Adverse Factors
Artificial Intelligence	4,740,855	43,515	4,872,086
Cloud-Based	4,771,077	290,800	4,836,407
Robotics	4,771,494	28,758	4,814,035
Specialized Equipment	4,771,936	204,661	4,825,232
Specialized Software	4,740,229	340,577	4,825,607

* The sample size is smaller because it applies only to firms that indicated they were producing the technology, thus filtering out non-producers.

Table 1: Total Number of Firms Reporting in the Production Survey by Technology Group

Table 2 provides the sample sizes for employer firms that participated in each part of the extent of use survey broken down by technology group.

Technology Group	Extent of Use	Motivation*	Adverse Factors
Artificial Intelligence	4,750,687	141,731	4,743,443
Cloud-Based	4,784,033	1,550,716	4,731,659
Robotics	4,785,415	88,657	4,748,446
Specialized Equipment	4,785,419	855,657	4,751,574
Specialized Software	4,750,559	1,821,368	4,733,331

* Sample size is smaller as it applies only to firms that indicated they were using the technology, thus filtering out non-users.

Table 2: Total Number of Firms Reporting in the Usage Survey by Technology Group

4. FINDINGS

The data was analyzed using Tableau Desktop version 22.2.0 and Microsoft Excel 365 software. We begin by reporting the breakdown of firms that produced, did not produce, or did not know whether they produced the corresponding technology in Table 3. The data shows that most businesses did not produce any of the emergent technologies or goods/services that included those technologies during the survey period. From the subset of companies that did produce the technology, specialized software represented the largest percentage (3.9), followed by cloud-based computing (3.2), specialized equipment (2.3), AI (0.4), and Robotics (0.3).

Technology Group	Yes	No	Don't know
Artificial Intelligence	19,789 (0.4)	4,470,228 (94.3)	250,838 (5.3)
Cloud-Based	152,386 (3.2)	4,356,220 (91.3)	262,471 (5.5)
Robotics	15,071 (0.3)	4,520,639 (94.7)	235,784 (4.9)
Specialized Equipment	108,675 (2.3)	4,393,931 (92.1)	269,330 (5.6)
Specialized Software	185,315 (3.9)	4,296,562 (90.6)	258,352 (5.5)

Table 3: Number of Firms Responding to Technology Production Questions by Technology Group (percentage of firms in parenthesis)

We then analyzed the number of firms that used each technology group in Table 4. We find that specialized software is the most used, followed by cloud-based technologies, specialized equipment, AI, and robotics. The share of firms using and producing, as well as their order in terms of size, is proportionate to one another, hence providing supporting evidence for hypothesis 1.

Technology Group	Yes	No	Don't know
Artificial Intelligence	141,731 (3)	4,336,113 (91.3)	251,786 (5.3)
Cloud-Based	1,550,716 (32.4)	2,931,192 (61.3)	264,342 (5.5)
Robotics	88,657 (1.9)	4,517,555 (94.4)	170,601 (4.9)
Specialized Equipment	855,657 (17.9)	3,658,991 (76.5)	256,238 (5.4)
Specialized Software	1,821,368 (38.3)	2,641,604 (55.6)	262,673 (5.5)

Table 4: Number of Firms Responding to Technology Usage Questions by Technology Group (percentage of firms in parenthesis)

We then compute a ratio of users to producers for each technology group. We find the highest ratio (10.1) for cloud-based computing, followed by specialized software (9.8), specialized equipment (7.8), AI (7.5), and robotics (6.3). Higher numbers indicate greater demand for this technology relative to producers, whereas lower numbers indicate higher competition for producers of that technology.

Technology Group	Users	Producers	Ratio of Users to Producers
Artificial Intelligence	141,731 (3)	19,789 (0.4)	7.5
Cloud-Based	1,550,716 (32.4)	152,386 (3.2)	10.1
Robotics	88,657 (1.9)	15,071 (0.3)	6.3
Specialized Equipment	855,657 (17.9)	108,675 (2.3)	7.8
Specialized Software	1,821,368 (38.3)	185,315 (3.9)	9.8

Table 5: Ratio of Users to Producers by Technology Group

To gain additional insight regarding the industries in which these technologies were produced, we provide the number of firms for each industry sector (by NAICS Code) and Technology Group in Table 6, posted in the Appendix (due to space limitations). The data shows that the largest sector that produces specialized software, specialized equipment, cloud computing, and AI is "Professional, scientific, and technical services." The largest sector that produces robotics is Manufacturing.

Next, we report the motivating factors for firms that produced the technologies in Figure 1. Using the highlighted table approach to emphasize the magnitude of each factor, we note that the foremost motivating factor across all technology groups was to upgrade goods and services. The second most important reason across the board was to expand the range of goods and services. The third top reason was different for AI with adapting existing products to new markets (at 41.2%). The third reason for all remaining technology groups included increasing or maintaining market share. Adopting standards and accreditation was the least reported factor, followed by "Some other reason".

	Artificial Intelligence	Cloud-Based	Robotics	Specialized Equipment	Specialized Software
Upgrade goods or services	56.50	50.70	45.40	52.50	50.80
Expand the range of goods or services	54.70	42.60	41.80	44.90	39.20
adapting existing products to new markets	41.20	27.30	32.70	25.50	24.30
Increase or maintain market share	36.10	30.40	35.00	31.20	29.70
Adopt standards and accreditation	16.50	18.40	12.40	14.70	16.90
Some other reason	14.80	21.50	23.50	19.50	22.90

Figure 1: Percentage of Firms by Motivation to Produce Technology Group

Subsequently, we report the motivating factors for firms that use the technologies in Figure 2. Using the same approach, we note that the foremost motivating factor was improving the quality or reliability of processes or methods, followed by upgrading outdated processes or methods. These results provide some support for hypothesis 2.

	Artificial Intelligence	Cloud-Based	Robotics	Specialized Equipment	Specialized Software
To improve quality or reliability of processes or methods	49.30	49.20	45.20	51.10	50.90
To upgrade outdated processes or methods	37.10	43.30	32.60	38.60	43.40
To automate tasks performed by labor	27.50	15.60	40.20	20.50	19.80
To expand the range of goods or services	23.80	14.10	24.40	25.60	15.30
Some other reason	18.20	23.50	21.10	17.80	19.40
To adopt standards and accreditation	11.70	10.40	8.60	10.90	11.90

Figure 2: Percentage of Firms by Motivation to Use Technology Group

Finally, we report factors adversely affecting technology production in Figure 3 and technology usage in Figure 4. These figures demonstrate that most respondents reported that no factors adversely affected the technology production or usage or that the technology did not apply to their business. However, the specific adverse factors for both users and producers included technology being too expensive, which provides support for hypothesis 3.

	Artificial Intelligence	Cloud-Based	Robotics	Specialized Equipment	Specialized Software
No factors adversely affect..	43.90	52.40	43.20	48.80	53.50
Technology not applicable to thi..	48.80	38.60	50.10	43.50	38.00
Technology was too expensive	5.50	5.40	5.40	5.80	6.00
Lacked access to capital	0.90	0.90	0.70	1.20	1.10
Concerns regarding safety..	0.60	2.00	0.40	0.50	0.80
Technology was not mature	1.10	0.70	0.60	0.40	0.50
Lacked access to required human ..	0.70	0.70	0.50	0.60	0.80
Laws and regulations	0.40	0.60	0.30	0.40	0.60
Lacked access to required data	0.50	0.50	0.30	0.30	0.40
Required data not reliable	0.30	0.30	0.20	0.20	0.20

Figure 3: Factors Adversely Affecting Technology Production by Technology Group

	Artificial Intelligence	Cloud-Based	Robotics	Specialized Equipment	Specialized Software
No factors adversely affected the adoption of this technology	43.10	55.30	43.30	50.70	57.90
Technology not applicable to this business	46.60	31.00	48.00	38.50	29.60
Technology was too expensive	7.70	7.00	7.20	7.90	8.50
Concerns regarding safety and security (physical security and/or cyber security)	1.10	4.20	0.50	0.60	1.40
Lacked access to capital	1.50	1.40	1.00	1.80	1.60
Technology was not mature	2.10	1.10	0.90	0.60	0.80
Laws and regulations	0.80	1.00	0.30	0.60	1.00
Lacked access to required human capital and talent	1.20	1.00	0.70	0.70	1.00
Lacked access to required data	0.90	0.80	0.40	0.50	0.70
Required data not reliable	0.50	0.50	0.30	0.30	0.40

Figure 4: Factors Adversely Affecting Technology Usage by Technology Group

Additionally, for AI, respondents noted that the technology was not mature, whereas, for cloud computing, the concerns regarding safety and

security were more prevalent. The percentage was more evenly distributed among the remaining factors for robotics, specialized equipment, and specialized software.

5. DISCUSSION AND CONCLUSION

This study explored the prevalence of US firms producing and using technology, factors motivating technology production or usage, and reasons hindering their progress among US firms.

Our first hypothesis investigated the proportions of US firm adopters to producers. Data shows that shares of adopters of emergent tech groups are proportionate to shares of producers. For example, the AI share of users was ranked as the second smallest category, which matched the order of AI producers, who also ranked as the second smallest share. Hence, we found supporting evidence for our first hypothesis. This finding, therefore, raises an important point regarding the role suppliers play in the diffusion of the innovation process. The ratio of users to producers offers insights into the technology groups that are more saturated with competition, like robotics, which has the lowest ratio, versus cloud-based computing, which has the highest ratio and thus less competition.

Our second hypothesis looked at the motivations of producers when compared to users, investigating their scope and reach. The results reveal that the majority of firms that produce emergent technologies do so to upgrade their goods or services and expand their range of goods or services. However, one key distinction between users and producers was that producers also focused on increasing market share and adapting existing products to new markets. These reasons demonstrate that companies that supply or produce technology focus on more strategic business reasons, whereas the companies that only use emergent technologies do so at an operational level and to gain efficiency through internal processes. Hence, it provides at least partial support to our second hypothesis.

Our last hypothesis looked at the adverse factors of producers, highlighting cost as the primary challenge for producers. This hypothesis was met based on data analyzed for this study. When the "technology non-applicable to this business" and "No factors adversely affecting the adoption…" were excluded from the sample, the main adverse reason for both users and producers was financial in nature, with both groups selecting technology as too expensive as the key adverse reason. This factor emphasizes that technology has been and continues to be one of the most expensive units within the organization. It is multifaceted and accounts for infrastructure, data, application development, security, and production support, to name a few. According to Bell's law, a new computer class is created each decade, imposing constant change and improvement (Bell, 2008). This ongoing change contributes to increased costs not only due to the need for new and improved hardware and software but also human resources and continuous upskilling and knowledge-sharing initiatives. Such factors might be compounded for suppliers who also focus on sales, marketing, customer service, and support.

This paper has important implications for both business strategy and policymaking. For businesses, the findings underscore the importance of innovation and technological advancement as critical drivers for competitiveness and market expansion. Firms that invest in producing new technologies are more likely to secure a stronger position in the market by continually improving their offerings and exploring new market opportunities. This strategic approach not only helps retain existing customers but also attracts new ones by meeting their evolving needs with advanced products and services.

From a policymaking perspective, the study highlights the need for supportive measures that encourage technology production. Government incentives, such as tax breaks, grants, and subsidies for research and development, can play a crucial role in fostering innovation. Additionally, creating a favorable regulatory environment that simplifies the process of bringing new technologies to market can significantly boost the efforts of firms engaged in technological production. Policymakers should also consider investing in education and training programs to build a skilled workforce capable of supporting high-tech industries (Acemoglu & Restrepo, 2019).

However, the study also points to several factors that may hinder the progress of technology-producing firms. These include high research and development costs, regulatory challenges, and a shortage of skilled labor. Addressing these barriers is essential for sustaining innovation. Firms must find ways to manage R&D expenses, perhaps through collaborations and partnerships that share the financial burden and risks associated with innovation. Furthermore,

engaging with regulatory bodies to streamline processes and reduce bureaucratic delays can facilitate faster commercialization of new technologies (Cordes et al., 2022).

Moreover, this study highlights the importance of technology suppliers and their role in innovation diffusion. Their ability to share information about innovation through the social system influences the adoption of that technology (Hall & Khan, 2003). With such a slight prevalence of technology suppliers observed in the US market today, it is rational to conclude that this might be one of the reasons the adoption rate is still very low. Similarly, Dar et al. (2024) found that information intervention directed at suppliers increased the adoption of farming modernization in agriculture. Hence, supporting and investing in technology suppliers might help facilitate and expedite user adoption.

While this study provides valuable insights, it is not without limitations. One significant limitation is the scope and timeframe of the data, which may not fully capture the diverse landscape of technology production across different industries and regions today. Future research could benefit from more refined measures and questions to capture the technology categories and motivations more clearly (Zolas et al., 2020). Additionally, more recent longitudinal studies could provide deeper insights into technology production trends and impacts on firm performance and market dynamics. Studies assessing the reach and level of information propagation by advanced technology suppliers could reveal opportunities for intervention and further support. Finally, comparative studies involving firms from different countries could offer a more comprehensive understanding of global trends in technology production and how the US compares to other innovators like Europe (e.g., Eurostat, 2024).

In conclusion, this study sheds light on the strategic importance of technology production for firms and the factors influencing their ability to innovate. By addressing the identified challenges and leveraging the motivating factors, firms can better navigate the complex landscape of technological advancement and secure a competitive edge in the market. Policymakers, in turn, must create an enabling environment that supports sustained innovation and technological growth.

6. REFERENCES

Acemoglu, D., Anderson, G. W., Beede, D. N., Buffington, C., Childress, E. E., Dinlersoz, E., ... & Zolas, N. (2024). Automation and the Workforce: A Firm-Level View from the 2019 Annual Business Survey∗. National Bureau of Economic Research

Acemoglu, D., Anderson, G. W., Beede, D. N., Buffington, C., Childress, E. E., Dinlersoz, E., ... & Zolas, N. (2022). Automation and the workforce: A firm-level view from the 2019 Annual Business Survey (No. w30659). National Bureau of Economic Research.

Acemoglu, D. & Restrepo, P. (2019). Automation and New Tasks: How Technology Displaces and Reinstates Labor Journal of Economic Perspectives—Volume 33, Number 2—Spring 2019—Pages 3–30.

Bell, G. (2008). Bell's law for the birth and death of computer classes. Communications of the ACM, 51(1), 86-94.

Bhalerao, K., Kumar, A., Kumar, A., & Pujari, P. (2022). A study of barriers and benefits of artificial intelligence adoption in small and medium enterprise. Academy of Marketing Studies Journal, 26, 1-6.

Bunte, A., Richter, F., & Diovisalvi, R. (2021). Why It is Hard to Find AI in SMEs: A Survey from the Practice and How to Promote It. In Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) (pp. 614-620). SCITEPRESS Science and Technology Publications, Lda.

Cordes, J., Dudley, S., & Washington, L. (2022) Regulatory Compliance Burdens: Literature Review and Synthesis, The George Washington University Regulatory Studies Center. Retrieved from: https://regulatorystudies.columbian.gwu.edu/sites/g/files/zaxdzs4751/files/2022-10/regulatory_compliance_burdens_litreview_synthesis_finalweb.pdf

Cubric, M. (2020). Drivers, barriers and social considerations for AI adoption in business and management: A tertiary study. Technology in Society, 62, 101257.

Dar, M. H., De Janvry, A., Emerick, K., Sadoulet, E., & Wiseman, E. (2024). Private input suppliers as information agents for technology adoption in agriculture. American Economic Journal: Applied Economics, 16(2), 219-248.

Eurostat (2024). Community innovation survey. Retrieved from: https://ec.europa.eu/eurostat/web/microdat

a/community-innovation-survey

Hall, B. H., & Khan, B. (2003). Adoption of new technology, https://www.nber.org/system/files/working_papers/w9730/w9730.pdf

Hameed, M. A., Counsell, S., & Swift, S. (2012). A conceptual model for the process of IT innovation adoption in organizations. *Journal of Engineering and Technology Management*, *29*(3), 358-390.

Lai, P. C. (2017). The literature review of technology adoption models and theories for the novelty technology. JISTEM-Journal of Information Systems and Technology Management, 14(1), 21-38.

McElheran, K., Li, J. F., Brynjolfsson, E., Kroff, Z., Dinlersoz, E., Foster, L., & Zolas, N. (2024). AI adoption in America: Who, what, and where. Journal of Economics & Management Strategy, 33(2), 375-415.

McKinsey (2019). How Artificial Intelligence will transform Nordic businesses, https://www.mckinsey.com/featured-insights/artificial-intelligence/how-artificial-intelligence-will-transform-nordicbusinesses.

Rogers, E.M. (1995). Diffusion of Innovations. 4th ed., New York: The Free Press

Rosenberg, Nathan (1972). "Factors Affecting the Diffusion of Technology." Explorations in Economic History, Vol. 10(1), pp. 3-33. Reprinted in Rosenberg, N. (1976), Perspectives on Technology, Cambridge: Cambridge University Press, pp. 189-212.

United States Census Bureau (2019), Annual Business Survey, https://data.census.gov/cedsci/table?q=technology

Zolas, N., Kroff, Z., Brynjolfsson, E., McElheran, K., Beede, D., Buffington, C., Goldschlag, N., Foster, L. & Dinlersoz, E. (2020). Advanced Technologies Adoption and Use by U.S. Firms: Evidence from the Annual Business Survey, NBER Working Paper No. 28290, JEL No. M15,O3,O47,O51

Appendix A.
Additional Table

Meaning of NAICS Code	Artificial Intelligence	Cloud-Based	Robotics	Specialized Equipment	Specialized Software
Professional, scientific, and technical services	8148	60691	2786	18265	66944
Health care and social assistance	1573	14298	1192	13635	18362
Retail trade	1823	10556	1771	10839	13437
Wholesale trade	1377	7960	2356	12629	11401
Manufacturing	833	3630	3050	14116	9400
Construction	1198	7892	752	10242	9501
Other services (except public administration)(663)	649	5672	N/A	9604	8672
Administrative and support and waste management and remediation services	984	7020	837	6688	8747
Information	1328	10436	347	2145	9366
Finance and insurance(662)	803	8437	186	1304	9709
Accommodation and food services	749	5876	756	4258	6871
Real estate and rental and leasing	309	5155	189	1513	6646
Transportation and warehousing(661)	194	2345	107	2172	2620
Educational services	94	2255	179	731	2649
Arts, entertainment, and recreation	107	1188	181	1317	2171
Management of companies and enterprises	184	818	157	804	1131
Mining, quarrying, and oil and gas extraction	5	49	4	359	293
Agriculture, forestry, fishing and hunting(660)	N/A	210	N/A	132	172
Utilities	7	16	8	58	49
Industries not classified	N/A	N/A	N/A	N/A	104
Grand Total	**20365**	**154504**	**14858**	**110811**	**188245**

Note: Rows with N/A indicate that the number was unavailable in the data set for those records.

Table 6: Number of Firms Producing Technology by Industry Sector and Technology Group