Journal of Big Data and Artificial Intelligence

Volume 3 2025 Number 1

Journal of Big Data and Artificial Intelligence

Volume 3 **2025** **Number 1**

EDITORIAL

ARTICLES

Journal of Big Data and Artificial Intelligence

Journal of Big Data and Artificial Intelligence

Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 3, No. 1, 2025 DOI: 10.54116/jbdai.v3i1.55

EDITORIAL

When Machines Create! Envisioning Our Future as Shaped by the Transformative Power of Generative AI

Abhishek Tripathi
The College of New Jersey
tripatha@tcnj.edu

Jim Samuel
Rutgers University; AIXOsphere
jim.samuel@rutgers.edu

Margaret Brennan-Tonetta
MBT Consulting; Rutgers University (Rtd.)
mbrennan@scarletmail.rutgers.edu

Hieu Nguyen
Rowan University
Nguyen@rowan.edu

Ensela Mema
Kean University
emema@kean.edu

"The risk climate of modernity is thus unsettling for everyone: no one escapes."

—Anthony Giddens, 2013

Typically, models are designed to represent reality and to produce output. In this sense, artificial intelligence (AI) can be viewed as a model of human intelligence capabilities to learn, analyze, and generate new configurations of information. In this sense, AI "machines" have been generative since the inception of the AI discipline in the 1950s and we should not be surprised by what we are now seeing in the form of "generative AI" (gen AI) applications, but we are! The recent widespread appreciation of the generative aspects of AI applications is due to the ready availability (all that is needed is a connected browser on any device!) of such applications to the masses, ease of use, the increased speeds at which gen AI outputs can be produced, and the impressive usefulness of its novel output. Gen AI has achieved fast-food status on a consumer level and is rapidly being commoditized and woven into the socioeconomic fabric of human society. As we look to the future, strategic human enhancive AI architectures, for example, adaptive cognitive fit (ACF), have the potential to help unleash iterations of rapid and complex advancements that will be treated as hyper-value creation opportunities, and emerging latent risks could be underestimated (Samuel et al. 2022; Kasirzadeh 2025). We have a solemn responsibility to ensure the development of ACF and similar human-centered AI architectures, which will help nurture a society that supports mass-human ascendancy over AIs, as opposed to the converse (Kashyap et al. 2024).

The field of AI and machine learning (ML) is rapidly evolving, with new and improved techniques emerging each year. The recent development of gen AI is a huge breakthrough in AI and ML. Gen AI can create, imitate, and produce content that is very similar to human-created content. Through algorithms and ML techniques, gen AI can be

queried to generate text, music, and images. Although gen AI has been around for quite some time, the introduction of ChatGPT by OpenAI (2025) on November 30, 2022, has made a dramatic impact on many fields, including health care and academic circles. ChatGPT is a specific application of the Generative Pretrained Transformer (GPT) series that is finetuned for human-like conversation. GPT was trained on hundreds of billions of words and can learn from any text without additional training (Rudolph et al. 2023). It has 175 billion parameters (Rudolph et al. 2023). ChatGPT acquired 1 million users within 5 days of its launch. Recently, OpenAI released ChatGTP-4, a chatbot that can interact with human beings through written language as well as images. ChatGPT has a fast-growing user base and now plays a vital role in numerous sectors, such as health care, finance, sports, information systems, and education. Some researchers argue that the ability of ChatGPT to process huge amounts of data will save time and potentially create summaries of academic research with less bias (Dergaa et al. 2023). Others see a positive impact of gen AI for academia, given the ability of tools such as ChatGPT to optimize time and effort for writing and editing (Irigaray and Stocker 2023).

To illustrate the power of gen AIs, let us consider the domain of higher education as an example. Despite the benefits offered by gen AI, the massive increase in both the amount of unverified data available and the frequency at which these technologies are used has raised many concerns. Currently, the impact of gen AI on ethics and integrity is widely debated in higher education (Moya et al. 2023). The fear that gen AI can be used as a tool for misconduct (Susnjak 2022) and discussions of how it can change the landscape of higher education in academic circles are frequent. Plagiarism is considered inevitable in many circles because gen AI can write in a way that is indistinguishable from a student's work (Cotton et al. 2023). It is therefore critically important to reflect on different ways to navigate changes and mitigate the risks driven by gen AI, especially in higher education (Moya et al. 2023). To navigate these challenges, it is essential that the higher education community proactively establish policies to mitigate the risks posed by gen AI. This includes strategies to combat cheating, revising assessment plans, and incorporating plagiarism detection tools, for example, Turnitin (2025). Furthermore, it is necessary to understand the ethical challenges we face: future research can focus on exploring the factors that influence students' ethical behavior and identify the "right" and "good" usage of such technology. This understanding will guide the creation of ethical standards for faculty and students when using gen AI.

This volume of the *Journal of Big Data and Artificial Intelligence* contains a very interesting and diverse set of articles that cover AI, data analytics and domain specific research articulations. When investigating the application of AI to agriculture, "Accurate Crop Yield Estimation of Blueberries using Deep Learning and Smart Drones" presents an innovative AI pipeline that leverages smart drones equipped with computer vision to improve the accuracy of blueberry yield estimation in a field (Nguyen et al. 2025). The authors use two YOLO (You Only Look Once) based object detection models: the Bush model, which identifies blueberry bushes from aerial and angled images, and the Berry model, which detects individual berries on a bush to improve crop yield estimation. The study also discusses deployment strategies, annotation challenges for small objects, and complexities in model evaluation.

With applications to the financial sector, the article "Unveiling Industry Pressures: A Data-Driven Analysis of SEC Filings, Amendments, and Reclassifications" reveals new insights into pressures faced by companies that drive their behavior through an analysis of their SEC regulatory filings (Olsen 2025). The author uses four key indicators based on the frequency and type of filing to introduce new metrics reflective of external forces that influence a company's business strategy and describes patterns extracted from these filings by using the open-source query language Malloy to answer key questions that help to inform investors, regulators, and policymakers about underlying trends that are distinctive of certain industries. The article "Multinational Investment under Uncertainty" builds a real options model to quantify multinational investment timing decisions under foreign market demand and exchange rate dynamics (Nagaraj et al. 2025). The article "Applications of Analytics in Disease Prediction Types" explores the identification of genetic and clinical variables that can serve as predictors for disease classification (Tsai et al. 2025). When using data from The Cancer Genome Atlas (TCGA) and the Genomic Data Commons (GDC), a comprehensive collection of patient genetic and clinical information compiled by the National Institutes of Health, the study examines whether there are significant differences in the characteristics of landmark and non-landmark genes.

With providing valuable insights for educators and policymakers, in "A Holistic Approach to Identifying Subject Correlation Analysis in Secondary Education," the authors used advanced data mining techniques, including correlation, regression, factor analysis, and hierarchical clustering, which indicated that holistic approaches can lead to more effective educational strategies and improved student outcomes (Ayesha et al. 2025). Gene AI has demonstrated transformative potential, and this is well reflected in the paper "Improving Large Language Model (LLM) Performance with Retrieval Augmented Generation (RAG)," in which the authors develop a RAG-based system for university support and educational purposes. This research emphasizes transparency, visibility, and control of key processes in the gen AI workflow, which is an excellent approach for ethical AI, and, even though it may present some immediate risks, it is much more rewarding in the long run (Chidipothu et al. 2025).

In conclusion, it is necessary to emphasize the importance of developing AI innovation, AI education, and AI policy and ethics dimensions simultaneously. This is critical due to the complexity of the evolving AI ecosystem. As has been aptly stated: *With AI technologies, given the scope, speed and scale at which damage can occur, it is compellingly necessary to implement forward-thinking policies now to ensure the future safety and sustainability of human rights and the human way of life* (Samuel 2021). AI education for all is essential and misnomers must be addressed, for example, gen AI can be developed apart from LLMs by using innovative designs (Garvey et al. 2020). AI is advancing rapidly, and, as we create more AI technologies, we should anticipate a structuration-like process to occur, which would facilitate the technologies we create to shape human society (Giddens 1984). ACF and similar architectures will continue to be developed, along with arguments for open-source AI. Open-source AI can help increase transparency and accountability, and help reduce both operational and ethical risks (Samuel 2023). It is expected that a dream scenario of a utopian future is more likely than a nightmarish dystopian future; however, this will require hard work, and deliberate and educated efforts from society as a system of people and resources.

References

Ayesha, B., A. Jayasooriya, W. Mendis, B. Mahanama, M. Dayasiri, U. Thayasivam, et al. 2025. "A Holistic Approach to Subject Correlation Analysis in Secondary Education." *Journal of Big Data and Artificial Intelligence* **3**, no. 1: 85–101. https://doi.org/10.54116/jbdai.v3i1.45

Chidipothu, N., R. Anderson, J. Samuel, A. Pelaez, J. Esguerra, and M. N. Hoque. 2025. "Improving Large Language Model (LLM) Performance with Retrieval Augmented Generation (RAG): Development of a Transparent Generative Artificial Intelligence (Gen AI) University Support System for Educational Purposes." *Journal of Big Data and Artificial Intelligence* **3**, no. 1: 102–122. https://doi.org/10.54116/jbdai.v3i1.50

Cotton, D. R. E., P. A. Cotton, and J. R. Shipway. 2023. "Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT." *Innovations in Education and Teaching International* **61**, no. 2: 228–239. doi: 10.1080/14703297.2023.2190148

Dergaa, I., K. Chamari, P. Zmijewski, and H. Ben Saad. 2023. "From Human Writing to Artificial Intelligence Generated Text: Examining the Prospects and Potential Threats of ChatGPT in Academic Writing." *Biology of Sport* **40**, no. 2: 615–622. doi:10.5114/biolsport.2023.125623

Garvey, M. D., J. Samuel, and A. Pelaez. 2021. Would You Please Like My Tweet?! An Artificially Intelligent, Generative Probabilistic, and Econometric Based System Design for Popularity-Driven Tweet Content Generation. Decision Support Systems 144: 113497.

Giddens, A. 1984. "The Constitution of Society: Outline of the Theory of Structuration." University of California Press.

Giddens, A. 2013. "Modernity and Self-Identity: Self and Society in the Late Modern Age", p. 128, John Wiley & Sons.

Irigaray, H., and F. Stocker. 2023. "ChatGPT: A Museum of Great Novelties." *Cadernos EBAPE.BR* **21**, no. 1: 1–5. doi: 10.1590/1679-395188776x

Kashyap, R., Y. Samuel, L. W. Friedman, and J. Samuel. 2024. "Artificial Intelligence Education and Governance-Human Enhancive, Culturally Sensitive and Personally Adaptive HAI." *Frontiers in Artificial Intelligence* **7:** 1443386.

Kasirzadeh, A. 2025. "Two Types of AI Existential Risk: Decisive and Accumulative." Philosophical Studies, 1–29.

Nagaraj, P., M. Nishihara, and C. Zhang. 2025. "Multinational Investment under Uncertainty." *Journal of Big Data and Artificial Intelligence* **3**, no. 1: 41–63. https://doi.org/10.54116/jbdai.v3i1.29

Nguyen, H. D., B. McHenry, T. Nguyen, H. Zappone, A. Thompson, C. Tran, et al. 2025. "Accurate Crop Yield Estimation of Blueberries using Deep Learning and Smart Drones." *Journal of Big Data and Artificial Intelligence* **3**, no. 1: 5–24. https://doi.org/10.54116/jbdai.v3i1.44

OpenAI. 2025. "ChatGPT Large Language Model." Accessed February 14, 2025. https://openai.com/chatgpt/overview/

Olsen, T. 2025. "Unveiling Industry Pressures: A Data-Driven Analysis of SEC Filings, Amendments, and Reclassifications." *Journal of Big Data and Artificial Intelligence* **3**, no. 1: 25–40. https://doi.org/10.54116/jbdai.v3i1.53

Rudolph, J., S. Tan, and S. Tan 2023. "ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education?" *Journal of Applied Learning & Teaching* **6**, no. 1: 342–363. doi: 10.37074/jalt.2023.6.1.9

Samuel, J. 2021. "*A Call for Proactive Policies for Informatics and Artificial Intelligence Technologies*." *Scholars Strategy Network.*

Samuel, J. 2023. "The Critical Need for Transparency and Regulation Amidst the Rise of Powerful Artificial Intelligence Models." *Scholars Strategy Network (SSN) Key Findings*. Accessed November *27*, 2023.

Samuel, J., R. Kashyap, Y. Samuel, and A. Pelaez. 2022. "Adaptive Cognitive Fit: Artificial Intelligence Augmented Management of Information Facets and Representations." *International journal of information management* **65**: 102505.

Tsai, C.-Y., S. M. Srinivasan, and A. Tripathi. 2025. "Applications of Analytics in Disease Prediction Types." *Journal of Big Data and Artificial Intelligence* **3**, no. 1: 64–84. https://doi.org/10.54116/jbdai.v3i1.49

Turnitin. 2025. "Education and Academic Research Support Company." Accessed February 14, 2025. https://www.turnitin.com/resources/topics/ai-writing

Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 3, No. 1, 2025 DOI: 10.54116/jbdai.v3i1.44

ACCURATE CROP YIELD ESTIMATION OF BLUEBERRIES USING DEEP LEARNING AND SMART DRONES

Hieu D. Nguyen
Rowan University
nguyen@rowan.edu

Brandon McHenry
Rowan University
mchenr49@students.rowan.edu

Thanh Nguyen
Rowan University
nguyent@rowan.edu

Harper Zappone
Rowan University
zappon37@students.rowan.edu

Anthony Thompson
Rowan University
thomps79@students.rowan.edu

Chau Tran
Rowan University
tranch29@students.rowan.edu

Anthony Segrest
Rowan University
segres62@students.rowan.edu

Luke Tonon
Rowan University
tonon153@students.rowan.edu

ABSTRACT

We present an AI pipeline that involves the use of smart drones equipped with computer vision to obtain a more accurate fruit count and yield estimation of the number of blueberries in a field. The core components are two object-detection models based on the YOLO deep learning architecture: a Bush Model, which is able to detect blueberry bushes from images captured at low altitudes and at different angles, and a Berry Model, which can detect individual berries that are visible on a bush. Together, both models allow for more accurate crop yield estimation by allowing intelligent control of the drone's position and camera to safely capture side-view images of bushes up close. In addition to providing experimental results for our models, which show good accuracy in terms of precision and recall when captured images are cropped around the foreground center bush, we also describe how to deploy our models to map out blueberry fields by using different sampling strategies and discuss the challenges of annotating very small objects (blueberries) and difficulties in evaluating the effectiveness of our models.

Keywords *blueberry crop yield, precision agriculture, smart drones.*

Figure 1: Image of a blueberry bush.

1. Introduction

Precision agriculture by using AI and autonomous drones has been shown to be an effective approach in not only estimating yield for many different crops but also in detecting and managing weeds and diseases. One popular specialty crop for which an accurate estimation of crop yield is important is the blueberry; prediction early in the growing season is important in helping farmers make pricing decisions, hire a sufficient number of pickers, and inform their distributors of the available supply before harvest time.

One AI-based approach to estimating crop yield involves using deep learning models trained on images to detect fruits and vegetables; many such models have been developed for many of these, including blueberries, grapes, apples, and tomatoes. However, early models were trained on either simulated data (Wang et al. 2023) or close-up images of clusters of fruit (from hand-held cameras or mobile devices) and not the entire bush or vine (Hani et al. 2020; Osman et al. 2021; Egi et al. 2022; Hofinger et al. 2023; Melnychenko et al. 2024); thus, these models are inappropriate for industrial use in which crop count and disease detection must be performed over a large field many acres in size. For example, a one-acre blueberry field can contain up to 1,000 bushes, which would make walking through such a field to capture a good sample of close-up images quite time-consuming; thus, a "boots-on-the-ground" approach is impractical.

However, by using unmanned aerial vehicles (UAV), in our case, drones, to capture images of blueberry bushes from a farther distance presents a much more efficient approach to mapping large fields.[1] In particular, we consider smart drones programmed with computer vision, that is, drones that have an onboard mini computer to process images captured by an onboard camera, to identify a blueberry bush and position itself (drone) to capture an optimal view of the bush that maximizes the number of visible berries on the bush. In addition, we make a distinction between a smart drone versus an autonomous drone, in which the former is capable of making its own decisions during flight, for example, by altering its path to avoid collision, as opposed to an autonomous drone in which its actions, including its flight path, are pre-programmed before takeoff. For example, commercial drones that are equipped with collision-avoidance systems (based on technologies such as infrared, stereo vision, and LiDAR) and provide object-tracking capabilities would fall under our classification of a smart drone.

In this work, we present a pipeline of object-detection models based on the YOLO (You Only Look Once [Redmon et al. June 2016]) deep learning architecture to estimate the crop yield of a blueberry field by using *smart* drones programmed with these models to accurately capture images of blueberry bushes and detect the number of harvestable berries on each bush (an example of a blueberry bush is shown in Figure 1). Our work is novel in its approach of detecting not only individual berries that are visible on an entire blueberry bush but also in detecting the bush itself to guarantee an accurate drone position and image capture of the bush. We distinguish such a smart mission from an autonomous mission as follows: in the latter, a drone is programmed before the start of the mission to fly to predetermined points of a blueberry field, say, along a row of bushes but keeping itself some fixed distance away

[1]We note that ground autonomous vehicles (or robots) offer an alternative solution; however, we shall not discuss them and their trade-offs in this paper because this topic has been well addressed in the literature and doing so will set us too far adrift from the focus of our work.

from the row (and high enough to clear neighboring rows); moreover, its onboard camera can be programmed to capture images of these bushes but from a fixed angle of view. However, this does not always guarantee that the bushes will be fully captured by the drone's camera. For example, this situation may occur if the drone's position is blown off course due to wind or temporary loss of GPS. Thus, it is not possible for the drone to adjust its distance from the bushes or its camera's angle of view in real time to compensate for this.

In contrast, drones programmed with our Bush Model can fly a more intelligent mission in which it is able to adjust its position and fly as close as possible alongside a bush (while maintaining a safe distance) to capture an optimal view of the bush so that berries on it appear as large as possible (for more accurate detection). This also allows for the implementation of various sampling strategies, for example, random stratified sampling of bushes, without having to know the exact GPS location of each bush, something that would be required for autonomous missions. However, coupled with real-time kinematic positioning, our pipeline makes precision mapping of blueberry fields possible where the location of each bush can be geotagged by using our Bush Model. The berry count can then be calculated for each bush after the flight by using our Berry Model. Then, given appropriate field data, which we describe in Section 3, crop yield can be estimated more accurately than from current methods.

The rest of the paper is divided as follows. In Section 2, we discuss related works. In Section 3 we present our proposed pipeline for mapping a blueberry field to obtain a more precise estimation of crop yield. In Section 4, we present experimental results for our models, which we hope provides baseline results for future researchers to compare their work against, and discuss challenges we faced in annotating tiny objects (blueberries) and how this impacted the effectiveness of our models.

2. Related Works

Recent advances in computer vision, in particular object-detection models based on deep learning, such as the YOLO architecture, have led to a proliferation of works in precision agriculture. More specifically, many of these works present highly accurate models to perform fruit detection and yield estimation. Because many articles have recently been published in this field (a 2020 survey article [van Klompenburg et al. 2020] reviewed 30 articles that used deep learning models), we limit our discussion to works that involve either deep learning models for detecting blueberries or drone-based methods.

Works similar to ours include Yildirim and Ulu. 2023, in which a pipeline for estimating crop yield of apples by using drones and objection-detection models based on the Faster R-CNN and SSD-Mobilenet architectures is described; Wang et al. 2023 and Melnychenko et al. 2024, in which improved YOLOv5 models are used to detect apples from aerial drone images; Hani et al. 2020, in which a semantic segmentation approach is used to detect and count apples but trained on ground-based images; Egi et al. 2022, in which a YOLOv5 is trained to detect tomatoes from drone-based images; Osman et al. 2021, in which YOLO-based models are used to detect apples, oranges, and pumpkins; and Hofinger et al. 2023, in which a YOLOv5 model is trained to detect black pine tree tops from UAV images.

We note that the aforementioned fruits and trees in the articles cited previously are somewhat large in comparison with blueberries, which, due to their size, are much more difficult to accurately annotate, especially in aerial drone images in which the spatial resolution of the berries is poor due to their small size and distance from the drone (Figure 1). Works related to fruits that are comparable in size include Akiva et al. 2020, in which a deep learning model based on U-Net and trained on aerial images is described to segment and count cranberries for yield estimation and sun exposure; Shen et al. 2023, in which YOLO-based models are used to perform real-time tracking and counting of grape clusters; and Pinheiro et al. 2023 in which YOLO-based models are used to not only detect grape bunches but also assess their quality in terms of damage from lesions.

Works that involve blueberries include MacEachern et al. 2023 and Yang et al. 2023, in which YOLOv3-v4 models are used to detect and estimate different stages of ripeness in blueberries (the latter in wild blueberries) but trained on images captured by hand-held cameras; Ni et al. 2020, in which a Mask R-CNN model is used to segment individual blueberries to estimate fruit maturity; Stefanović et al. 2022, in which a row detection segmentation model based on the U-Net architecture and trained on UAV images is described; and Filipović et al. 2023, in which a bush-detection model is described but trained on annotations of only the trunks of blueberry bushes and not the entire bush; however, our Bush Model is trained to detect the entire bush, which is necessary for accurate crop yield estimation.

There are very few works that are similar to ours in which their computer vision models are validated on images and the results are compared against the actual fruit count per plant or tree. In particular, for our validation dataset, berries on 15 blueberry bushes were all hand picked to obtain actual fruit counts (what we call "picked ground

truth"). These works are few in number because actual or harvested crop yield is typically recorded by weight and not by the number of fruit. Among such works, most involve large fruit such as apples (Bargoti and Underwood, 2017) and mangos (Stein et al. 2016; Payne et al. 2013) but also for almonds (Underwood et al. 2016) and grapes (Palacios et al. 2023). However, none of these works explicitly discuss the ratio of visual fruit count to actual fruit count (per bush) as we do in Section 6.4 to better understand the amount of occlusion.

3. Pipeline

Let F be a two-dimensional rectangular blueberry field of size A (in acres) that contains C bushes. We assume that GPS coordinates of the corners of F and the direction D of the rows of bushes in F are known; moreover, we assume that all rows run along the same direction (Figure 2). Let Y denote the crop yield of F (berries/acre).

Our pipeline for estimating Y can be summarized as follows: we first use stratified sampling to fly a smart drone over random points of F and capture images of bushes at these points by using our Bush Detection model. We then calculate the number of berries on these bushes by using our Berry Detection model to estimate Y by using formula 1. Here are the main steps of our pipeline, depending on the sampling method:

1. Stratified Sampling: We partition F into a grid of $M \times N$ non-overlapping square cells, denoted by $\{C_{mn}\}$. We present two stratified sampling strategies to sample bushes within each cell: point sampling and row sampling. In point sampling, we select a single bush closest to a randomly chosen point inside each C_{mn}. In row sampling, we sample a row of bushes inside C_{mn} (Figure 3).

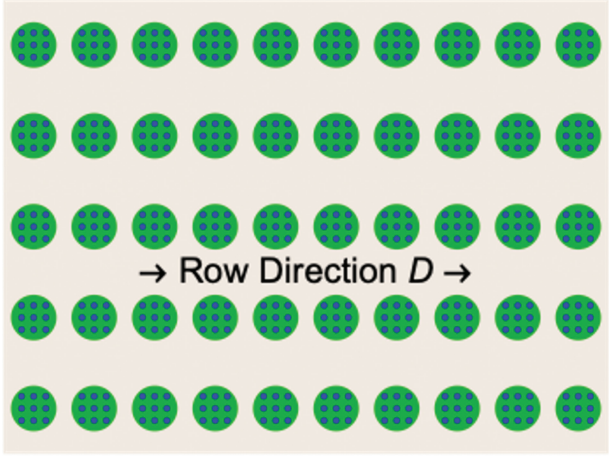


Figure 2: Row direction of a field.

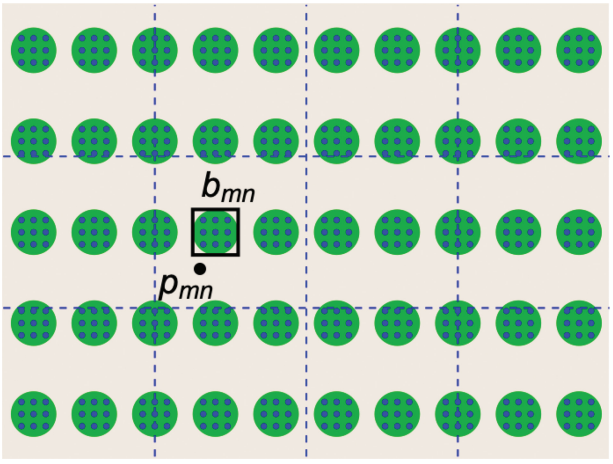


Figure 3: Stratified sampling of a field.

(a) Point Sampling: Select a random point, denoted by p_{mn}, inside each cell C_{mn}.

(b) Row Sampling: Select a random point p_{mn} along an edge of C_{mn} whose direction is perpendicular to D.

(c) Fly the drone to each position p_{mn} (at given altitude h).

2. Single Bush Detection

(a) At position p_{mn}, use a drone camera (pointed down) to capture bushes in its angle of view (see Figure 4 for a birds-eye view) and use our Bush Model to identify a bush closest to p_{mn}. Denote by b_{mn} the position of this closest bush, that is, the center of the bounding box enclosing the bush as illustrated in Figure 3. Then, apply object tracking by using DeepSort to begin tracking the bush.

(b) Program the drone to fly horizontally to position b_{mn} (while maintaining altitude h) so that the drone is directly over the bush.

3. Bush Image Capture (Angled-Side View)

(a) Fly the drone horizontally to one side of the bush (side chosen randomly to account for factors such as the sun's angle) in a direction perpendicular to D (Figure 2) so that it is distance d away from b_{mn} while simultaneously adjusting the camera angle to keep the bush in view by using object tracking. Further adjustments to either the drone position or camera angle (or combination of both) can be made so that the entire bush is in view of the camera (Figure 5).

Figure 4: Blueberry field from a birds-eye view.

Figure 5: Row of blueberry bushes from a angled-side view.

(b) Point Sampling: Use the drone camera to capture the image of the bush (at position b_{mn}) and record the coordinates of the bounding box predicted by the Bush Model, denoted by c_{mn}.

(c) Row Sampling: Fly the drone along D and use the camera to capture photos of bushes that appear in the center of its view and record the bounding box coordinates c_{mn} of each captured bush until the drone reaches the opposite edge of the cell C_{mn}. Apply object tracking to distinguish different bushes and adjust the drone's position to ensure a photo capture of the entire bush.

4. Berry Counting (Post-Mission)

(a) Use bounding box coordinates c_{mn} to determine the number of visible berries on the corresponding bush (from one side) captured by images in the previous step. We describe two approaches (we discuss their trade-offs in Section 4):

i. Image Cropping: Crop each image to contain only the bush (with bounding box coordinates c_{mn}) and apply the Berry Model on cropped image to obtain the number of visible berries on each bush.

ii. Bounding Box Filtering: Apply the Berry Model on the entire image and count only detections of berries that are inside the bush's bounding box to obtain the number of visible berries on each bush.

(b) Calculate the mean number of berries obtained in the previous step (averaged over all bushes) and double this answer to obtain the mean number of berries per bush, denoted by B.

5. Crop Yield Estimation: We calculate crop yield Y as follows:

$$Y = \alpha \cdot \frac{B \cdot C}{A} \tag{1}$$

where B is determined from the previous step, C is the bush count, A is the size of F, and α is a fixed proportionality constant, called the picked-visual ratio (PVR), which describes the ratio of what we refer to as the picked ground truth (GT) to the visual GT:

$$\alpha = \frac{\text{Picked GT}}{\text{Visual GT}} \tag{2}$$

The picked GT is defined to be the number of berries that are actually on a bush (or on a group of bushes) and verified by manually picking and counting all the berries on that bush. However, the visual GT is defined to be the number of berries on the same bush that is visible from a side view and ideally given by B if our Berry Model was perfect. We assume α to be given and that it would be determined from historical data because α would be highly dependent on factors such as climate, soil, variety of blueberry, and height and angle of the drone camera. In Section 6, we provide the first estimates of α that were obtained from two validation datasets in which all the berries on the foreground center bush of each image were picked by hand by our team to obtain the picked GT.

4. Datasets

4.1. Data Collection

Our data consist of images (still photos and video frames) of blueberry bushes (highbush blueberry, *Vaccinium corymbosum*) from two different varieties, Duke and Draper. These images were collected at outdoor blueberry farms in southern counties of New Jersey and captured by using a combination of hand-held (including cellphone) cameras and drone cameras to create a more diverse dataset. Although our total collection consists of more than a thousand such images, only a fraction of them were used to train our models due to limited resources and the time-consuming process of annotating these images, which we further discuss below.

The following datasets were created to train and validate our Berry and Bush Models.

4.2. Berry Datasets

The Berry datasets in total consist of 95 annotated images that were either captured by using drone cameras or hand-held cameras:

1. 35 aerial photos (drone)
2. 60 ground photos (hand-held)

These 95 images represent the total number that we have been able to completely annotate to date. Although relatively few in number, these images contain well over 1,00,000 annotations of berries, which we discuss later, in

Section 4.4. Thus, we believe that the size of our datasets at this point is sufficient to obtain reasonably accurate models, as we demonstrate later.

A train/validation split of 84/16 was used to divide our data into a Training Set (80 images) and three Validation Sets A, B, C, each consisting of 5 images (Table 1). We divided the Training Set into two parts (Table 2): a Drone subset consists of 20 images captured by DJI Phantom 3 and Autel Evo II Pro drones, and a hand-held subset that consists of 60 images captured by various hand-held cameras (Iphone, Android, Canon EOS Rebel). All in the Training Set were taken in the summer of 2021 and 2022 (Figures 6(a), 6(b) and 7(a), 7(b) for sample images). The validation datasets are different as follows: Validation Set A consists of five drone images taken in summer 2022 with a DJI Phantom 3 drone; Validation Set B consists of five drone images of the same foreground bushes as in Set A, but of their opposite side; Validation Set C consists of five drone images taken in Summer 2023 by using a DJI Mini 3 drone.

The reason for merging higher-quality hand-held images with drone images to create our (Merged) Training Set when we began this project 3 years ago is that we anticipated future improvements in drone cameras to match the spatial resolution of current hand-held cameras (different cameras can have the same pixel resolution but different spatial resolution, even after accounting for ground sampling distance). In particular, images captured with the DJI Mini 3 from 2023 (Validation Set C) seems to have higher spatial resolution in comparison with those captured with the DJI Phantom 3 in 2022 (Validation Set A). We present validation metrics in Section 6 that provide evidence to support this, although it should be noted that Validation Sets A and C are images of different blueberry

Table 1: Berry datasets.

Berry Datasets	Images	Green Annotations	Blue Annotations	Total Annotations
Training Set	80	61,680	17,471	79,151
Validation Set A	5	11,328	1,059	12,387
Validation Set B	5	8,810	906	9,716
Validation Set C	5	13,093	7,060	20,153
Total	95	94,911	26,496	1,21,407

Table 2: Berry training set.

Berry Training set	Images	Green Annotations	Blue Annotations	Total Annotations
Drone	20	31,372	4,694	36,066
Hand-held	60	30,308	12,777	43,085
Total (Merged)	80	61,680	17,471	79,151

(a) DJI PHantom 3 (b) Autel Evo II

Figure 6: Sample drone images from the Training Set. (a) DJI PHantom 3; (b) Autel Evo II.

(a) Canon EOS Rebel (b) iPhone

Figure 7: Sample hand-held images from the Training Set. (a) Canon EOS Rebel; (b) iPhone.

varieties, Duke and Draper, respectively, and that the ratio of green-to-blue berries differs significantly for these two datasets, which we address in the next section.

4.3. Bush Datasets

The Bush datasets consist of 256 drone images of blueberry bushes captured at various different altitudes and camera angles (e.g., birds-eye versus side views). We used a train/validation split of 90/10 to define our Training and Validation Sets (Table 3).

4.4. Annotation

Images in our dataset were manually annotated by using computer vision platforms Roboflow and CVAT, and involved a team of more than 10 people (mostly undergraduate research students in our research laboratory).

Berry Model: Images used for training our object detection Berry Model were annotated by drawing a rectangular bounding box tightly around each individual blueberry that is visible in the image and labeling it according to one of two color classes: Green or Blue (Figure 8). We did not have an objective criteria for separating the berries into these two classes, except by providing the annotators with examples of berry images that were labeled by consensus (Figures 9 and 10). We considered creating additional color classes to distinguish the berries but decided on two classes to allow for quicker annotation given the total number of berry annotations in our dataset (>120,000 annotations), which we believe to be the largest of its kind. The number of annotations (Green, Blue, total) are given in Table 1. Observe that the number of Green annotations is significantly higher than the number of Blue annotations, with the ratio of Green to Blue highest for Validation Set A (10.7) and smallest for Validation Set C (1.85). This is because the images in Set C were captured much closer to the first harvest date compared with the images in Set A.

Occluded berries were annotated if the annotator was convinced that the object was a berry and only its visible portion was annotated. Shadowy and/or blurry berries were also annotated according to the same criterion. We acknowledge that this criterion is dependent on the visual acuity of the annotator and exposes the difficulty of annotating tiny objects such as blueberries. One can use majority voting by using a group of annotators; however, given the large number of berries in a single image (>1,000 berries were annotated on average), we considered this approach to not be feasible given our limited resources. In addition, we annotated only those berries that were large enough to be *harvestable*. Berries that fruited late in the season are too small in size to be harvestable by commercial farms. Examples of berries that were not annotated are shown in Figure 11.

Table 3: Bush datasets.

Bush Dataset	Images	Bush Annotations
Training Set	473	2,684
Validation Set	26	314
Total	256	2,998

Figure 8: Example of blueberry annotations (Green and Blue classes).

Figure 9: Examples of berries labeled as Green.

Figure 10: Examples of berries labeled as Blue.

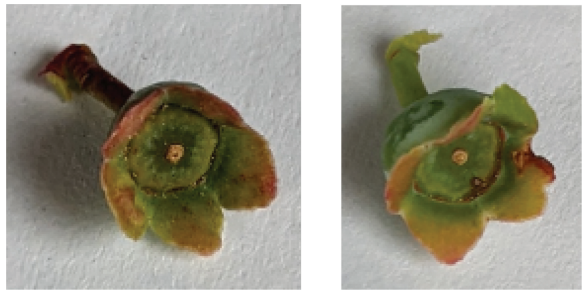


Figure 11: Examples of berries considered too small (not annotated).

Bush Model: For our Bush Detection Model, images were annotated by drawing a rectangular bounding box around each blueberry bush in the foreground of the image, including its trunk (if visible) and all branches (Figures 12 and 13). Neighboring bushes with long branches will unavoidably cause their bounding boxes to overlap, as seen in Figure 13.

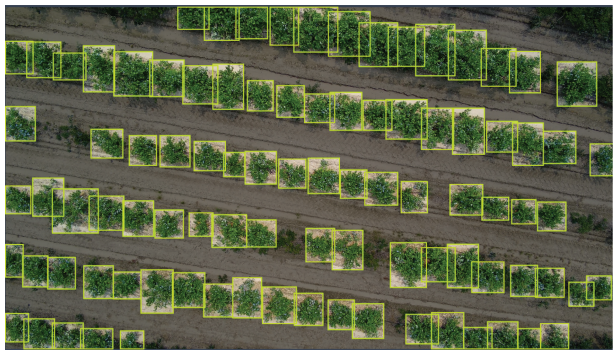


Figure 12: Example of bush annotations (birds-eye view).

Figure 13: Example of foreground bush annotations (side view).

4.5. Data Augmentation

We perform data augmentation as follows to create a diverse training dataset for more effective training. For our Bush Model, we applied variations in Hue (± 0.015), Saturation (± 0.7), Value (brightness) (± 0.4), Translation (± 0.1), Scale (± 0.5), and Flip Left-Right (0.5 probability).

However, for our berry counting model, data augmentation consisted of dividing each original full-size image into tiles. This is because the YOLO model works better with smaller, square images; thus, we divide each image into 640×640 tiles. If the image cannot be perfectly cut into 640×640 tiles (starting at top left), then the leftover pieces are discarded. See Figure 14 for a depiction of the tiling process. Any berry annotation with its bounding box that appears in two neighboring tiles is decided by choosing the tile that contains the center of the box.

Figure 14: Depiction of the tiling process.

5. YOLO Objection Detection

5.1. YOLOv5

YOLOv5 is the fifth version of the YOLO family of compound-scaled object detection models, first introduced by Redmon et al. (2016). We chose YOLOv5 instead of other object-detection models, for example, Faster R-CNN, because of its state-of-the-art performance (back when we began this project more than 3 years ago) in terms of inference speed relative to accuracy. There are many higher versions of YOLO now available; however, experiments show negligible improvement in training metrics for both the Berry and Bush Models when using say version YOLOv8 in comparison with YOLOv5. The focus of this paper is not to present a comparison of all the different versions of YOLO but to establish baseline results for YOLOv5 that others can compare their results against.

Both Berry and Bush Models were trained by using YOLOv5s (small) because there was little improvement in the training metrics when resorting to larger-sized models such as YOLOv5m (medium) and YOLOv5l (large). In addition, especially for the Bush Model, in which inference needs to be performed in real time by the drone on-board computer during flight, the small-size model makes this possible. We were able to achieve 6-9 frames per second performing detection the Bush model on a Jetson Nano with Deepstream running YOLOv5s.

5.2. Training

Both the Bush and Berry Models were trained on Bush and Berry Datasets, respectively, by using YOLOv5s' default hyperparameter settings found in its hyperparameter yaml file "hyp.scratch-low.yaml." We used a batch size of 32 and trained for up to 400 epochs for the Bush Model and 300 epochs for the Berry Model; the default early stopping criterion was used to stop training when cross-validation loss diverged from the training loss. Training metrics were calculated by using Ultralytics YOLOv5 utils Python library through the script "metric.py." We found little difference in accuracy by changing other default parameters and believe that optimizing these parameters would not change the conclusions of our paper; we believe that it is more important to increase the size of our dataset and add more higher-resolution images to significantly improve our model.

5.3. Validation

The best fold from fivefold cross-validation of each model (base on the highest mAP:0.5) was used to perform validation on the datasets in Tables 1 and 3. For the Berry Model, each image was first divided into overlapping 700 $\times$ 700 tiles, or as close as possible to these dimensions, so that they overlap by 60 pixels in each dimension; this avoids double-counting berries that may be split if non-overlapping tiles were used. Each tile was then passed through the Berry Model and post-processing was used to remove duplicate bounding boxes that appear in two overlapping tiles. To compute precision and recall for each class, a confidence threshold of 0.1 was applied to generate detections, and an Intersection over Union (IOU) threshold of 0.3 was applied to count true positives when comparing them against GTs; if more than one detection matched a GT in terms of both IOU threshold and class, then the detection with the highest confidence is selected. A low IOU threshold was used (in comparison with YOLOv5's default threshold of 0.6) to avoid eliminating correct predictions that did not overlap sufficiently with the GT. This is because bounding boxes of blueberries are small in dimension; thus, detection errors in the position of these boxes by just a few pixels can significantly impact their IOU.

5.4. DeepSORT Tracking

To test the accuracy of our Bush Model in tracking bushes as discussed in our pipeline in Section 3, we used DeepSORT to calculate multiple object tracking accuracy (MOTA). DeepSORT is a computer vision tracking algorithm for tracking objects from a video stream by assigning an ID to each detected object (Wojke et al. 2017). It is an extension of the Simple Online and Realtime Tracking (SORT) because it integrates appearance information based on a deep appearance descriptor. We applied DeepSORT to two short video clips: one of a drone performing Point Sampling and the other performing Row Sampling. Results are presented in Section 6.

6. Experimental Results

In this section, we present training and validation results for our Berry and Bush Models, both separately and also when combined to perform bush cropping to obtain a total berry count for only the foreground center bush. We also present tracking results (MOTA) for the Bush Model by using the DeepSORT algorithm.

6.1. Berry Model

Three different Berry Models were trained by using fivefold cross-validation: Drone, Hand-Held, and Merged. The Drone and Hand-Held Berry Models were trained on the 20 drone images and 60 hand-held images, respectively (Table 2). The Merged Berry Model (or simply Berry Model) was trained on the merged dataset of 80 images (drone and hand-held); we refer to this merged dataset as the Berry Training Set.

Training Results: Training metrics for the three Berry Models (Drone, Hand-Held, Merged) are given in Tables 4-6, respectively. The Hand-Held Model performed best across all metrics as expected (precision, recall, mAP, and F1) because the bushes in the hand-held images were shot at a closer distance, with berries appearing larger than those in drone images, thus making detection easier. The Drone Model performed almost as well as the Hand-Held Model in terms of precision, but recall was significantly worse. These two models are only as good as the data that they are trained on, and, so. when we present validation results below, we will see that their performance is reversed, thus supporting the need for a Merged Model that is robust to a variety of different types of images.

Validation Results: Tables 7, 8, and 9 show precision and recall for the three Berry Models (Drone, Hand-Held, Merged) validated on Validation Sets A, B, and C, respectively. For Sets A and B (Tables 7 and 8), the Drone Berry Model had the highest overall precision among all the models; however, the Merged Model had significantly higher overall recall and slightly better Blue precision. However, the Hand-Held Berry Model performed the worst in all categories. In particular, Green recall was extremely low, which indicated that the model failed to detect many

Table 4: Training metrics for the Drone Berry Model.

Training	Precision	Recall	mAP 0.5	mAP 0.5:0.95	F1
Fold 1	0.8271	0.6472	0.7281	0.3773	0.7262
Fold 2	0.8039	0.6378	0.7143	0.3724	0.7113
Fold 3	0.8184	0.6692	0.7350	0.3848	0.7363
Fold 4	0.8347	0.6583	0.7446	0.3945	0.7361
Fold 5	0.8032	0.6707	0.7295	0.3753	0.7310
Mean	0.8175	0.6566	0.7303	0.3809	0.7282
SD	0.0139	0.0142	0.0110	0.0089	0.0103

Table 5: Training metrics for the Hand-Held Berry Model.

Training	Precision	Recall	mAP 0.5	mAP 0.5:0.95	F1
Fold 1	0.8230	0.7518	0.8159	0.5119	0.7858
Fold 2	0.8527	0.7623	0.8339	0.5290	0.8050
Fold 3	0.8553	0.7450	0.8287	0.5176	0.7963
Fold 4	0.8606	0.8009	0.8639	0.5418	0.8297
Fold 5	0.8259	0.7640	0.8269	0.5320	0.7938
Mean	0.8435	0.7648	0.8338	0.5264	0.8021
SD	0.0177	0.0216	0.0180	0.0119	0.0169

Table 6: Training metrics for the Merged Berry Model.

Training	Precision	Recall	mAP 0.5	mAP 0.5:0.95	F1
Fold 1	0.8434	0.7267	0.7965	0.4822	0.7807
Fold 2	0.8303	0.7160	0.7852	0.4643	0.7690
Fold 3	0.8477	0.7253	0.7941	0.4805	0.7817
Fold 4	0.8323	0.7260	0.7962	0.4806	0.7755
Fold 5	0.8319	0.7085	0.7820	0.4636	0.7653
Mean	0.8371	0.7205	0.7908	0.4742	0.7744
SD	0.0079	0.0080	0.0067	0.0094	0.0072

Table 7: Drone, Hand-Held, Merged Berry Models: Validation metrics for Validation Set A.

Berry Model	Precision (Green Class)	Recall (Green Class)	Precision (Blue Class)	Recall (Blue Class)	Precision (Both Classes)	Recall (Both Classes)
Drone	0.775	0.708	0.807	0.245	0.776	0.67
Hand-Held	0.749	0.069	0.092	0.283	0.255	0.087
Merged	0.755	0.745	0.804	0.361	0.757	0.713

Table 8: Drone, Hand-Held, Merged Berry Models: Validation metrics for Validation Set B.

Berry Model	Precision (Green Class)	Recall (Green Class)	Precision (Blue Class)	Recall (Blue Class)	Precision (Both Classes)	Recall (Both Classes)
Drone	0.657	0.675	0.621	0.245	0.655	0.637
Hand-Held	0.608	0.205	0.111	0.461	0.337	0.228
Merged	0.601	0.734	0.69	0.317	0.605	0.697

Table 9: Drone, Hand-Held, Merged Berry Models: Validation metrics for Validation Set C.

Berry Model	Precision (Green Class)	Recall (Green Class)	Precision (Blue Class)	Recall (Blue Class)	Precision (Both Classes)	Recall (Both Classes)
Drone	0.609	0.478	0.866	0.433	0.68	0.461
Hand-Held	0.793	0.459	0.818	0.476	0.802	0.465
Merged	0.815	0.462	0.872	0.434	0.835	0.451

Table 10: Merged Berry Model validation metrics for Validation Sets A, B, C.

Validation Dataset	Precision (Green Class)	Recall (Green Class)	Precision (Blue Class)	Recall (Blue Class)	Precision (Both Classes)	Recall (Both Classes)
Set A	0.755	0.745	0.804	0.361	0.757	0.713
Set B	0.601	0.734	0.69	0.317	0.605	0.697
Set C	0.815	0.462	0.872	0.434	0.835	0.451

green berries, especially those in background bushes where they appear much smaller, which makes it more difficult for the model to detect them. Also, precision for class Blue was also quite low, which shows that the Hand-Held Berry Model did a poor job of correctly detecting blue berries.

As for Validation Set C (Table 9), performance reversed with the Merged Model now having the highest overall precision and only slightly worse overall recall compared with the Drone and Hand-Held Models. Observe that overall recall for the Hand-Held Model significantly improved compared with results in Tables 7 and 8. However, all three models performed poorly on overall recall. An inspection of the false negatives were of berries on background bushes that were annotated but either too small for any of the models to detect or too shaded for the model to distinguish as a berry. The results on overall precision provide evidence that training on combined drone and hand-held images helped to improve the (Merged) Berry Model, with only a slight decrease in overall recall (but best Green recall), in comparison with training on drone and hand-held images separately.

Precision and recall for only the Merged Berry Model to help better compare results for the three validation sets (A, B, and C) are isolated in Table 10. Results for Set C yielded the highest overall precision, but, unfortunately, also yielded the lowest overall recall, which we previously described as due to berries on background bushes that are too small in terms of pixel resolution for the model to detect. This is supported by results that we present later on when we consider detecting berries only on the foreground center bush.

Conversely, results for Set B had the lowest overall precision but highest overall recall. Figures 15, 16, and 17 show a sample image (B2) from Set B, but with different types of bounding boxes drawn: GTs (Figure 15),

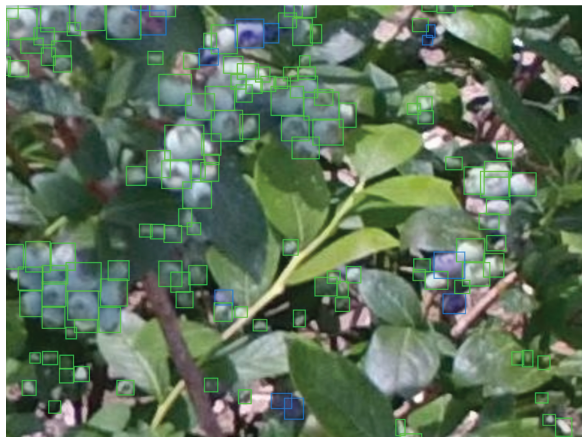


Figure 15: Close-up view of ground truth annotations in sample image from Validation Set B (green and blue boxes denote Green and Blue classes, respectively).

Figure 16: Close-up view of Berry Model predictions in sample image from Validation Set B (blue and red boxes denote Green and Blue classes, respectively).

Figure 17: Close-up view of Berry Model false positives (green) and false negatives (red) for sample image in Validation Set B.

predictions (Figure 16), and false positives and false negatives (Figure 17). A close inspection of the false positives shows that some of them could possibly be berries, but were difficult to discern clearly, which explains why they were not annotated.

6.2. Bush Model

Training/Validation Results: Training metrics for the Bush Model trained on the Bush Training Set are given in Table 11. Validation metrics for the Bush Model validated on the Bush Validation Set are given in Table 12. Both tables show high precision, at approximately 90%, and good recall, ranging from high 70% for training to low 80% for validation. A review of the true negatives (bushes that were not detected) indicates that the Bush Model struggled to detect those bushes at the edge of the image. Fortunately, this issue is not a concern because the goal of the Bush Model is to detect and track foreground bushes.

Sample predictions of the Bush Model from angled-side, birds-eye, and slanted views are shown in Figures 18, 19, and 20, respectively, including a false positive and false negative in the latter figure.

Bush Tracking: We calculated MOTA for two video clips: Bush Video 1 and Bush Video 2. Bush Video 1 (24 seconds) was captured by a DJI drone flying overhead to a bush and simultaneously adjusting its position and camera angle from birds-eye to angled-side view. Bush Video 2 (5 seconds) was captured by a DJI drone flying sideways along a row of blueberry bushes. MOTA results for both video clips are given in Table 13. Although MOTA is lower than what we hoped for, a review of the detections shows that the Bush Model does a very good job of tracking the foreground center bush, which is the primary goal of the model and performs worse for bushes at the edges of the image, something that we previously mentioned.

Table 11: Training metrics for Bush Model trained on Bush Validation Set.

Training Fold	Precision	Recall	mAP 0.5	mAP 0.5:0.95
Fold 1	0.899	0.76	0.867	0.508
Fold 2	0.842	0.723	0.807	0.433
Fold 3	0.881	0.774	0.869	0.493
Fold 4	0.939	0.893	0.946	0.592
Fold 5	0.877	0.733	0.829	0.492
Mean	0.888	0.777	0.864	0.504
SD	0.035	0.068	0.053	0.057

Table 12: Validation metrics for Bush Model validated on Bush Validation Set (using best fold).

Dataset	Precision	Recall	mAP 0.5	mAP 0.5:0.95
Bush Validation Set	0.916	0.834	0.916	0.538

Figure 18: Bush Model prediction on a sample validation image (angled-side view of bush).

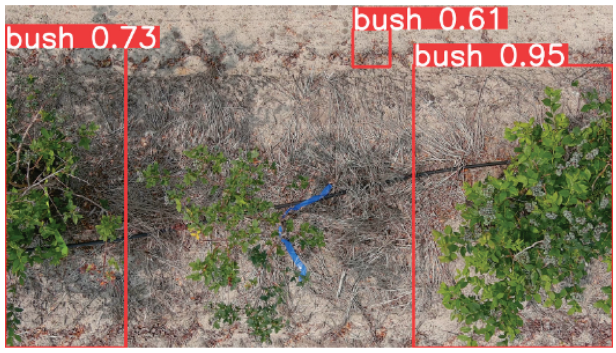


Figure 19: Bush Model prediction on a sample validation image (birds-eye view), showing a false positive (bush marked with confident 0.61) and false negative (bush tagged with blue ribbon).

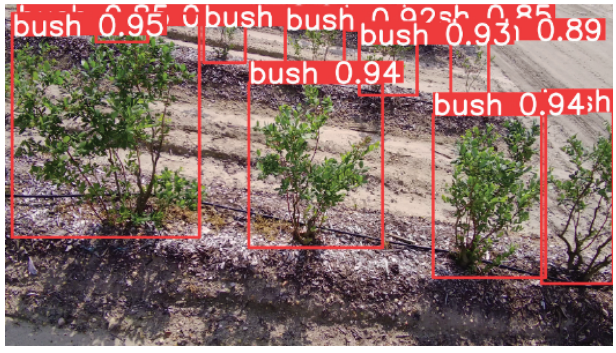


Figure 20: Bush Model prediction on an example validation image (slanted view).

Table 13: MOTA for Bush Model of Videos 1 and 2.

Parameter	Bush Video 1	Bush Video 2
Number of frames	365	75
Bush annotations	6464	324
Predictions	7104	352
Mismatch errors	149	7
False positives	2315	96
False negatives	1745	66
IOU threshold	0.5	0.5
MOTA	**0.3489**	**0.4786**

6.3. Bush-Cropped Berry Model

Together, the Berry and Bush Models can be combined into a pipeline to detect only those berries that appear in a single bush. We call this pipeline the Bush-Cropped Berry Model. In particular, we first pass a full-size image through the Bush Model to obtain an array of detected bushes and their corresponding bounding boxes. From these bounding boxes, we select one called the *central bounding box* (corresponding to the foreground center bush) (Figures 21(a) and 21(b)), whose center is closest (in terms of radial distance) to the center of the image and then crop the image (using OpenCV2) around the central bounding box. The cropped image is then passed through the Berry Model to detect berries. Detections are compared against those GTs contained within the central bounding box.

Shown in Tables 14, 15, and 16 are validation results by using the Bush-Cropped Berry Model (Drone, Hand-Held, and Merged) on Validation Sets A, B, and C, respectively. This time we see the Merged Model outperforming the Drone and Hand-Held Models in overall recall for all validation sets. The Hand-Held Berry Model performed the

(a) Full image (b) Cropped image

Figure 21: Example of image cropped around foreground center bush by Bush Model and then fed into Berry Model. (a) Full image; (b) Cropped image.

Table 14: Bush-Cropped Berry Model (Drone, Hand-Held, Merged): Validation metrics for Validation Set A.

Bush-Cropped Berry Model	Precision (Green Class)	Recall (Green Class)	Precision (Blue Class)	Recall (Blue Class)	Precision (Both Classes)	Recall (Both Classes)
Drone	0.778	0.675	0.733	0.168	0.776	0.631
Hand-Held	0.483	0.007	0.069	0.019	0.222	0.008
Merged	0.758	0.712	0.811	0.265	0.76	0.673

Table 15: Bush-Cropped Berry Model (Drone, Hand-Held, Merged): Validation metrics for Validation Set B.

Bush-Cropped Berry Model	Precision (Green Class)	Recall (Green Class)	Precision (Blue Class)	Recall (Blue Class)	Precision (Both Classes)	Recall (Both Classes)
Drone	0.683	0.688	0.727	0.251	0.684	0.655
Hand-Held	0.753	0.154	0.171	0.328	0.504	0.167
Merged	0.601	0.747	0.746	0.35	0.605	0.718

Table 16: Bush-Cropped Berry Model (Drone, Hand-Held, Merged): Validation metrics for Validation Set C.

Bush-Cropped Berry Model	Precision (Green Class)	Recall (Green Class)	Precision (Blue Class)	Recall (Blue Class)	Precision (Both Classes)	Recall (Both Classes)
Drone	0.708	0.583	0.848	0.618	0.752	0.595
Hand-Held	0.799	0.589	0.837	0.615	0.812	0.598
Merged	0.809	0.617	0.847	0.605	0.821	0.613

worst in both overall precision and recall for Sets A and B (Tables 14 and 15), as already observed when validated on entire images (Table 10) but performed surprising well on Green precision for Set B (Table 15).

For Set C, the Merged Model achieved the best overall precision and recall, with the Hand-Held Model having only slightly worse results. This demonstrates that the drone images in Set C have almost the same spatial resolution as the hand-held images. We believe going forward that the Merged Model will performed best (in terms of precision and recall) on images captured by future drones, whose camera resolution will only continue to improve.

For comparison, for Validation Sets A, B, and C, all validated by using the same Bush-Cropped Merged Berry Model, are isolated in Table 17. Here, the results are similar to those when validated on entire images (Table 10)

Table 17: Bush-Cropped Merged Berry Model: Validation metrics for Validation Sets A, B, and C.

Validation Dataset	Precision (Green Class)	Recall (Green Class)	Precision (Blue Class)	Recall (Blue Class)	Precision (Both Classes)	Recall (Both Classes)
Set A	0.758	0.712	0.811	0.265	0.76	0.673
Set B	0.601	0.747	0.746	0.35	0.605	0.718
Set C	0.809	0.617	0.847	0.605	0.821	0.613

Table 18: Calculation of α (predicted vs experimental) for the Berry Validation Set A by using the Bush-Cropped (Merged) Berry Model.

Image	Detections	Visual GT	Picked GT	α_p (Predicted)	α (Experimental)
A1	882	1,010	3,312	3.755	3.279
A2	1,451	1,230	3,996	2.754	3.249
A3	511	493	2,888	5.652	5.858
A4	711	847	2,920	4.107	3.447
A5	420	708	1,404	3.343	1.983
Mean	795	858	2,904	3.92	3.56
SD	408	282	950	1.09	1.41
Total	3,975	4,288	14,520	3.65	3.39

Table 19: Calculation of α (predicted vs experimental) for Berry Validation Set B by using the Bush-Cropped (Merged) Berry Model.

Image	Detections	Visual GT	Picked GT	α (Predicted)	α (Experimental)
B1	891	785	1,404	1.576	1.789
B2	885	806	2,920	3.299	3.623
B3	1,012	972	2,888	2.854	2.971
B4	1,856	1,842	3,996	2.153	2.169
B5	1,071	1,043	3,312	3.092	3.175
Mean	1,143	1,090	2,904	2.59	2.75
SD	406	435	950	0.71	0.75
Total	5,715	5,448	14,520	2.54	2.67

but observe that recall significantly improved for Set C because the model no longer needs to detect tiny berries on background bushes, which it had difficulty with when validating on entire images.

6.4. Estimation of PVR

The Bush-Cropped (Merged) Berry Model allows us to estimate the PVR α (see Part 5 of Section 3) by using its detections as an estimate for the picked GT. We denote by α_p (predicted α) to be any approximation of α calculated based on this estimate:

$$\alpha_p = \frac{\text{Picked GT}}{\text{Detections}} \approx \frac{\text{Picked GT}}{\text{Visual GT}} = \alpha \tag{3}$$

However, we distinguish α_p from experimental values of α calculated by using the annotated visual GT of the cropped image, that is, the number of berry annotations within the central bounding box), which we assume to be a very accurate estimate of the true value of α.

Tables 18, 19, and 20 give both predicted and experimental values for α for Validation Sets A, B, and C, respectively, including the total number of detections, visual GT, and picked GT are given for each image (cropped around the foreground center bush). Although the total number of detections seem to be good approximations of the visual GT for Validation Sets A and B (Tables 18 and 19, respectively), this is misleading because these detections

Table 20: Calculation of α (predicted vs experimental) for Berry Validation Set C by using the Bush-Cropped (Merged) Berry Model.

Image	Detections	Visual GT	Picked GT	α (Predicted)	α (Experimental)
C1	1,109	1,507	2,407	2.170	1.597
C2	831	924	3,215	3.869	3.479
C3	1,261	1,491	1,963	1.557	1.316
C4	713	618	2,307	3.236	3.733
C5	1,210	1,457	1,963	1.622	1.347
Mean	955	1,199	2,371	2.67	2.29
SD	225	406	513	1.10	1.21
Total	4,774	5,997	11,855	2.48	1.98

contain many false positives of berries (see overall precision and recall for the Bush-Cropped (Merge) Berry Model in Table 17), which cancel out the many false negatives of berries that were not detected. Thus, an accurate estimation of α_p will depend on an accurate Berry Model.

Experimental values of α differ widely for all three sets (Tables 18, 19, and 20), with mean experimental values highest for Set *A* and lowest for Set C. This shows that estimating α will be challenging because it seems to depend not only on the blueberry variety (recall that Sets A and C correspond to Duke and Draper varieties, respectively) but also which side of the bush is captured (recall that Sets A and B correspond to two sides of the same five bushes).

7. Discussion

Results of our Berry Model highlight challenges with annotating berries and training our models. Berries that are difficult to discern due to their small size, especially those on background bushes, can lead to subjective annotations and thus an ambiguous GT. Many false-positive detections could be argued as true detections of berries, depending on one's visual acuity, but difficult to confirm with certainty because of their low resolution. Moreover, occlusion of partially hidden berries, camouflage of green berries by leaves, and shaded berries make for training an accurate Berry Model quite challenging.

Results of the Bush Model clearly show that detecting bushes is not necessarily an easier task than detecting berries. Obviously, a bush is considerably larger than a berry; however, the complicated branch structure of a bush, in addition its branches possibly overlapping with a neighboring bush, creates challenges in training an accurate bush model.

Results of the Bush-Cropped Berry Model show the effectiveness of cropping around the foreground center bush to eliminate background berries and thus improved the model's precision and recall, which, in turn, provided a more accurate estimation of crop yield. Estimates of the PVR α based on the Bush-Cropped Berry Model show that it can vary significantly and depends on many factors such as the particular side of the bush that is captured and the blueberry variety, and other factors that we did not take into account: bush size, bush foliage density, environmental and soil conditions.

8. Conclusion

In this paper, we presented a pipeline of object detection models based on deep learning for detecting blueberry bushes and individual berries on them. These models allow a smart drone programmed with them to fly intelligent missions, namely to precisely locate bushes and capture their side views, thus obtaining a more accurate estimate of crop yield. We have already begun to test our pipeline by using a custom-build programmable drone to capture data and hope to report on our experimental results in the near future. We hope our work will spur interest in others to address the challenges raised in this paper and improve on our baseline results. All datasets, models, and source code will be made available on Github.

Acknowledgments: *The authors would like to acknowledge partial financial support from the New Jersey Council of County Colleges through their NJ Pathways to Career Opportunities Program, Department of Mathematics and College of Science and Mathematics at Rowan University. The following blueberry farms in South Jersey kindly provided us access to their fields to collect data: Macrie Brothers Farm, Moore's Meadow Farm, and Vacarrella Farm. We also thank other former and current team members who contributed to annotating our datasets: Robert Czarnota, Jacob Green, Lori Green, Felix Hakimi, Lance Ilagan, Nicholas Kaegi, Jamie Kahle, Brian Kim, Tuan Le, Ik Jae Lee, Duy Nguyen, Jonah Rodriguez, and Iosefa Sunia.*

References

Akiva, P., K. Dana, P. Oudemans, and M. Mars. 2020. "Finding Berries: Segmentation and Counting of Cranberries Using Point Supervision and Shape Priors." in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 219–228.

Bargoti, S., and J. Underwood. 2017. "Image Segmentation for Fruit Detection and Yield Estimation in Apple Orchards." *Journal of Field Robotics* **34**, no. 6: 1039–1060.

Egi, Y., M. Hajyzadeh, and E. Eyceyurt. 2022. "Drone-Computer Communication Based Tomato Generative Organ Counting Model Using Yolo v5 and Deep-Sort." *Agriculture* **12**, no. 9: 1290–17.

Filipović, V., D. Stefanović, N. Pajević, Z. Grbović, N. Djuric, and M. Panić. 2023. "Bush Detection for Vision-Based Ugv Guidance in Blueberry Orchards: Data Set and Methods." in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3646–3655.

Hani, N., P. Roy, and V. Isler. 2020. "A Comparative Study of Fruit Detection and Counting Methods for Yield Mapping in Apple Orchards." *Journal of Field Robotics* **37**, no. 2: 263–282.

Hofinger, P., H.-J. Klemmt, S. Ecke, S. Rogg, and J. Dempewolf. 2023. "Application of yolov5 for Point Label Based Object Detection of Black Pine Trees with Vitality Losses in Uav Data." *Remote Sens* **15**: 1–13.

MacEachern, C. B., T. J. Esau, A. W. Schumann, P. J. Hennessy, and Q. U. Zaman. 2023. "Detection of Fruit Maturity Stage and Yield Estimation in Wild Blueberry Using Deep Learning Convolutional Neural Networks." *Smart Agricultural Technology* **3**: 100099–11.

Melnychenko, O., L. Scislo, O. Savenko, A. Sachenko, and P. Radiuk. 2024. "Intelligent Integrated System for Fruit Detection Using Multi-Uav Imaging and Deep Learning." *Sensors* **24**, no. 6: 1913–1936.

Ni, X., C. Li, H. Jiang, and F. Takeda. 2020. "Deep Learning Image Segmentation and Extraction of Blueberry Fruit Traits Associated with Harvestability and Yield." *Horticulture Research* **7**, no. 1: 1–14.

Osman, Y., R. Dennis, and K. Elgazzar. 2021. "Yield Estimation and Visualization Solution for Precision Agriculture." *Sensors* **21**, no. 19: 6657.

Palacios, F., M. P. Diago, P. Melo-Pinto, and J. Tardaguila. 2023. "Early Yield Prediction in Different Grapevine Varieties Using Computer Vision and Machine Learning." *Precision Agriculture* **24**, no. 2: 407–435.

Payne, A. B., K. B. Walsh, P. Subedi, and D. Jarvis. 2013. "Estimation of Mango Crop Yield Using Image Analysis – Segmentation Methodn, Localisation and Yield Estimation Using Multiple View Geometry." *Computers and Electronics in Agriculture* **91**: 57–64.

Pinheiro, I., G. Moreira, D. Queirós da Silva, S. Magalhães, A. Valente, P. M. Oliveira, et al. 2023. "Deep Learning Yolo-Based Solution for Grape Bunch Detection and Assessment of Biophysical Lesions." *Computers and Electronics in Agriculture* **206**: 1–14.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. 2016. "You Only Look Once: Unified, Real-Time Object Detection." in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),

Shen, L., J. Su, R. He, L. Song, R. Huang, Y. Fang, et al. 2023. "Real-Time Tracking and Counting of Grape Clusters in the Field Based on Channel Pruning with yolov5s." *Computers and Electronics in Agriculture* **206**: 107662.

Stefanović, D., A. Antić, M. Otlokan, B. Ivošević, O. Marko, V. Crnojević, et al. 2022. "Blueberry Row Detection Based on Uav Images for Inferring the Allowed Ugv Path in the Field." in ROBOT2022: Fifth Iberian Robotics Conference.

Stein, M., S. Bargoti, and J. Underwood. 2016. "Image Based Mango Fruit Detection, Localisation and Yield Estimation Using Multiple View Geometry." *Sensors* **16**, no. 11: 1915.

Underwood, J., C. Hung, B. Whelan, and S. Sukkarieh. 2016. "Mapping Almond Orchard Canopy Volume, Flowers, Fruit and Yield Using Lidar and Vision Sensors." *Computers and Electronics in Agriculture* **130**: 83–96.

van Klompenburg, T., A. Kassahun, and C. Catal. 2020. "Crop Yield Prediction Using Machine Learning: A Systematic Literature Review." *Computers and Electronics in Agriculture* **177**: 105709–105718.

Wang, H., J. Feng, and H. Yin. 2023. "Improved Method for Apple Fruit Target Detection Based on yolov5s." *Agriculture* **13**, no. 11: 2167.

Wojke, N., A. Bewley, and D. Paulus. 2017. "Simple Online and Realtime Tracking with a Deep Association Metric." in 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649.

Yang, W., X. Ma, and H. An. 2023. "Blueberry Ripeness Detection Model Based on Enhanced Detail Feature and Content-Aware Reassembly." *Agronomy* **13**, no. 6: 1613–1619.

Yildirim, S., and B. Ulu. 2023. "Deep Learning Based Apples Counting for Yield Forecast Using Proposed Flying Robotic System." *Sensors* **23**, no. 13: 6171.

Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 3, No. 1, 2025 DOI: 10.54116/jbdai.v3i1.53

Unveiling Industry Pressures: A Data-Driven Analysis of SEC Filings, Amendments, and Reclassifications

Timothy Olsen
Gonzaga University
olsent@gonzaga.edu

Abstract

This study explores the dynamics of industry behavior through an analysis of regulatory filings, providing insights into the pressures and strategies shaping various sectors. Leveraging detailed financial data from the U.S. Securities and Exchange Commission (SEC), including over 83,000 10-K filings, we examine four key indicators: the frequency of 8-K filings, amendments to 10-Ks, address changes, and industry reclassifications. These indicators serve as proxies to measure external forces such as regulatory scrutiny, competitive pressure, and economic volatility.

Our findings reveal distinct patterns of filing behavior, with certain industries, such as Pharmaceutical Preparations, Metal Mining, and Programming Services, showing high levels of address changes, amended 10-Ks, and shifts in industry size—signals of underlying industry pressures. Additionally, industries like Patent Owners, Medicinal Products, and Retail experienced significant company transitions and frequent address changes, further reflecting industry realignments.

By introducing new metrics for assessing industry pressures and demonstrating the use of the open-source data language Malloy for replicable analysis, this research contributes both to the academic understanding of industry dynamics and to the practical tools available for data exploration.

JEL: *O51, G14, G11, G32.*

Keywords *financial statement analysis, industry studies, industry pressures*.

1. Introduction

Understanding industry dynamics is critical for investors, regulators, and policymakers seeking to evaluate business environments and the pressures that companies face. While existing research on industry analysis has explored a

wide range of metrics, such as profitability, regulation, and market structure, there is a growing need to examine industry-specific behaviors reflected in regulatory filings. The U.S. Securities and Exchange Commission (SEC) filings, such as Forms 10-K and 8-K, contain valuable information that can shed light on the operational and strategic decisions of firms within various industries. However, the potential of these filings to serve as proxies for industry pressures remains underexplored.

This study addresses key gaps in the literature by analyzing patterns of filings, amendments, address changes, and industry reclassifications in SEC submissions. By focusing on these often-overlooked indicators, we aim to provide new insights into the external forces acting on different sectors. Specifically, we propose that the frequency of 8-K filings, 10-K amendments, address changes, and industry transitions may serve as proxy measures for identifying industries under greater regulatory, competitive, or financial pressure.

Through a detailed analysis of SEC financial datasets, this study explores five key research questions that seek to reveal underlying trends across industries. First, we provide descriptive statistics of the SEC dataset to establish a baseline understanding of the data structure and content. Next, we investigate which industries file the most 8-K forms relative to their 10-K filings, hypothesizing that frequent disclosures may correlate with heightened regulatory scrutiny or volatility. We also examine which industries amend their 10-Ks most often, suggesting potential governance or reporting challenges. Additionally, this study looks at which industries change their reported mailing or business address most frequently, a factor potentially indicative of strategic repositioning or regulatory arbitrage. Finally, we analyze industry transitions, focusing on which industries companies transition into or out of most often, revealing underlying strategic shifts or market realignments.

By addressing these questions, this research aims to contribute to both the academic understanding of industry dynamics and the practical challenges of regulatory compliance. Our findings offer new perspectives on how industry-level pressures manifest in corporate filings, while also introducing a novel methodological approach using open-source tools to analyze large-scale regulatory data.

2. Literature Review

The study of industries is an interdisciplinary endeavor. For a few decades now the Industry Studies Association (https://www.industrystudies.org/) has brought together scholars from dozens of disciplines to discuss and present academic research about industries from around the globe. Comparative analysis of industries is crucial for understanding economic dynamics, informing policy decisions, and guiding business strategies. Scholars have identified various factors and measures to compare industries, such as the degree of regulation, competition levels, consumer demand, investor interest, technological innovation, market structure, and barriers to entry and exit (Vanneste 2017).

Regulatory frameworks vary by industry, affecting how industries evolve and compete globally. High regulatory environments can create barriers to entry, affect competition, and influence profitability. The level of competition within an industry shapes market dynamics, pricing strategies, and innovation. Investor interest reflects an industry's attractiveness and potential for returns. High investor interest can lead to increased capital flow, facilitating growth and innovation. Industries that demonstrate strong financial performance and growth prospects typically garner more investor attention (Gompers 1997).

Profitability metrics, such as return on investment and net profit margins, are essential for comparing financial performance across industries. (Fama and French 1997) analyzed industry-specific costs of equity, revealing variations in expected returns and risks. Industries with higher profitability are often more attractive to investors and may experience greater competition as firms seek to capitalize on lucrative opportunities.

We suggest that four factors from the SEC dataset may be useful for comparing industries as proxy measures for pressures they may experience. The factors we propose to examine are the number of 8-K filings, amended returns (which includes earnings restatements), address changes, and industry transitions. Measures of these factors may be indicative of competitive or regulatory pressures facing these industries. Our reasons for choosing these factors are that they are objective and can be ascertained automatically through a simple structured query language (SQL) or Malloy query on data updated each month from the SEC. Indicators such as risk factor changes, or litigation disclosures cannot be objectively identified in the way the SEC currently structures their data.

2.1. 8-K Filings

Firms submit Form 8-K forms to comply with SEC regulations, specifically Rule 13a-11 of the Securities Exchange Act of 1934, which mandates the timely disclosure of material corporate events within four business days. These

events include earnings announcements, mergers, management changes, and legal proceedings. This requirement ensures transparency by providing all investors with access to critical information, reducing information asymmetry and preventing unfair advantages (Lerman and Livnat 2010).

Industries that submit a higher volume of Form 8-Ks often face increased scrutiny from regulators, investors, and the media. This may indicate that these sectors operate in more volatile or fast-paced environments, with frequent significant events requiring disclosure (Watkins 2022). High-profile sectors like technology and finance, marked by rapid innovation and tighter regulatory oversight, commonly issue frequent updates to maintain investor confidence (McMullin et al. 2019). In addition, industries with intense competition or higher regulatory risks often see more demand for transparency, leading to frequent filings aimed at addressing stakeholder concerns and reducing uncertainty (Ben-Rephael et al. 2022).

Starting in 2019, submissions of form 8-K were required to use an XBRL format.Given this context, our second research question arises: Which industries file the most Form 8-Ks as a percentage of total 10-K submissions? This question aims to investigate the relationship between industry characteristics and the frequency of required disclosures, providing insight into which sectors face the most regulatory or stakeholder pressure.

2.2. 10-K Amendments

Academic literature identifies several reasons why companies amend their 10-K filings, often due to errors or omissions such as financial misstatements or incomplete disclosures. These errors may result from miscalculations, incorrect estimates, or the improper application of accounting standards (Thompson 2023). Amendments can also arise to incorporate new information like post-reporting events or in response to SEC comments requesting clarification (Krishnan and Zhang 2014). Frequent amendments may indicate governance issues or weak internal controls, necessitating corrections after initial filings (Cassell et al. 2019).

Amendments may include financial restatements or changes to the wording in the 10-K report. Amendments resulting from errors or missing information may indicate less competitive markets, where incumbents are less meticulous in their reporting. In contrast, amendments prompted by the SEC could signal strong regulatory oversight or a commitment to thorough auditing. In the future it may be possible to distinguish objectively the nature of amendment from our data source.

Industries with high rates of 10-K amendments may face unique pressures from stakeholders such as regulators, auditors, or investors (Curling 2006). Frequent amendments may suggest that firms in certain sectors struggle to produce accurate, timely disclosures due to the complexity and volatility of their operating environments (Cassell et al. 2019). Identifying which industries make the most amendments is significant because it highlights sectors where enhanced due diligence may be necessary for investors and where regulatory bodies might focus their oversight efforts.

This leads us to our third research question: Which industries have the highest frequency of amendments to their 10-K filings?

2.3. Changing Addresses

Academic research has analyzed the strategic motivations behind organizations changing addresses on their 10-K filings, often interpreting such moves as indicators of broader corporate strategy (Birkinshaw et al. 2006). Companies may relocate to take advantage of more favorable tax regimes, regulatory environments, or market access. In some cases, address changes occur without physical relocation, prompted by mergers, adjustments in corporate governance, or legal obligations (Baaij et al. 2015; Gregory et al. 2005). Industries that see frequent address changes often face external pressures from stakeholders, reflecting volatile or dynamic environments (Klier and Testa 2002).

This brings us to our fourth research question: Which industries change their reported mailing or business address most frequently? Understanding these patterns will help illuminate the external pressures driving such strategic decisions.

2.4. Changing Industries

Academic literature identifies several reasons why companies change the industry listed on their 10-K filings. Organizational restructuring—such as mergers, acquisitions, or divestitures—can shift a company's core operations into a different industry category (Tosun and Moon 2024; Vanneste 2017). Companies may also adjust their industry classification to better reflect evolving business models, especially in rapidly advancing sectors like Pharmaceuticals and Computer Services. Firms may change classifications to enhance investor perceptions or align their public image with their most profitable segments (Bhojraj et al. 2003; Wang and Coff 2022).

The U.S. SEC continues to utilize Standard Industrial Classification (SIC) codes for public filings within its Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system. The SEC's version of SIC codes is slightly modified to suit its regulatory needs, focusing on categorizing companies based on their primary lines of business for disclosure and oversight purposes (Phillips and Ormsby 2016). Because the SEC uses a slightly modified version of SIC codes, it is difficult to cross-reference them with the standard SIC or North American Industry Classification System (NAICS) codes. Furthermore, assigning a company to a specific NAICS code from a more general SIC code is not an objective task.

Reclassification might indicate responses to market demand shifts, regulatory changes, or competitive pressures that require repositioning. Industries undergoing significant disruption, firms realign themselves to maintain strategic flexibility—a response to shareholder demands for higher returns or new regulatory standards (Li et al. 2013). Identifying which industries are most transitioned into or out of is significant because it highlights sectors experiencing dynamic changes, offering insights for investors and regulators into market trends and potential risks (Gaspar et al. 2024).

This leads us to our fifth research question: Which industries experience the highest rates of industry reclassification in consecutive 10-K filings?

3. Methodology

This study employs exploratory data analysis (EDA) as a foundational approach. EDA is a flexible and intuitive method that allows for the exploration and understanding of large and complex datasets without the constraints of predefined hypotheses. Its primary goals are to summarize the main characteristics of the data, uncover patterns, identify outliers, and visualize relationships between variables. By using EDA, we aim to derive insights from the dataset without imposing rigid assumptions. This open-ended approach is particularly well-suited to large datasets, where traditional statistical methods may overlook emerging patterns or subtle relationships that become more apparent through visual exploration (Komorowski et al. 2016).

EDA is an appropriate method for analyzing the 10-K filings due to the high dimensionality of the dataset, which includes a variety of attributes ranging from basic descriptive information to filing periods, filing dates, industry codes, and further amendments. By allowing the data to guide the discovery process, EDA serves as a valuable tool in preparing the dataset for more detailed, hypothesis-driven analysis in future research (Nielsen 2022). This method has recently been employed to analyze 10-K data (Chakri et al. 2023) credit card usage and customer churn (Chakri et al. 2023) and a comprehensive analysis looking for anomalies and trends (Schroeder and Posch 2023).

The dataset for this study was sourced from the U.S. SEC, which provides publicly available financial data on various industries and companies. The dataset includes detailed financial reports such as balance sheets, income statements, and other key performance indicators critical for analyzing industry trends and performance.

As of June 15, 2011, all SEC filers are required to submit XBRL-tagged financial statements. Later in 2019, 8-K forms were required to be submitted in XBRL format. As the SEC dataset only provides information for XBRL tagged documents, no 8-Ks appear in the data prior to 2019. These filings are available for download, and many researchers have utilized this dataset in their work. Given the dataset's size and complexity, some studies use proprietary software to format and summarize the data. However, we demonstrate how open-source software, Malloy, can be effectively used to clean, prepare, and summarize this data.

The specific dataset used in this study was drawn from the SEC's Financial Statement Notes Data Sets. This dataset includes detailed information on corporate filings, such as form types, submission and acceptance dates, and SIC codes. The data consists of multiple zip files containing tab-separated value (.tsv) files.

3.1. Data Collection and Processing

The dataset was processed in several stages to prepare it for analysis:

3.1.1. Download and extraction

Approximately 22 GB of zip files were downloaded from the SEC's Financial Statement Notes Data Sets repository. After unzipping, the dataset expanded to 221 GB. Each zip file contained eight separate .tsv files, each providing different financial data aspects. We downloaded all of the .zip files up to August 2024.

3.1.2. Data selection

For this study, we focused on the sub.tsv files within each zip archive. These files contain key submission data, such as company name, address, form type (e.g., 10-K, 10-Q), filing and acceptance dates, and SIC codes, which classify the company's industry. Other .tsv files were excluded as they did not align with our research objectives focused on industry trends and performance.

3.1.3. Data conversion

We used the duckdb python module to combine and convert all the sub.tsv files to parquet format, an open-source columnar storage format optimized for data compression and performance during querying. This conversion reduced the dataset size back to approximately 22 GB, allowing for more efficient data access and analysis.

3.2. Querying and Analysis

After converting the dataset to a more efficient format, we used Malloy, an open-source query language that compiles to SQL, to explore various aspects of the data. Malloy simplifies reading and writing complex SQL logic, making it easier to perform iterative and exploratory queries. The code and data for this paper are available on GitHub (https://github.dev/mrtimo/IndustryStress).

Our exploratory, a-theoretic approach allowed flexibility in querying, adapting the analysis as patterns emerged. Key analyses involved refining Malloy queries based on initial results, especially for the more complex queries addressing Research Questions 4 and 5. These queries employed a lag function to identify changes in company address or industry in the first phase. Once identified, the results were passed to a second phase, this is also known as a subquery or nested query. In our analysis, we limited our results to industries that had at least 30 unique companies for the duration of our sample years (2012–2023).

4. Findings

In this section, we present the findings from our analysis addressing each of the research questions. The code used for these analyses is provided in the Appendix A and can also be accessed on GitHub, where it can be re-run to replicate the results. GitHub features an in-browser version of Visual Studio Code that can be launched by authenticated users by pressing the "." key, and the code can be executed after installing the Malloy extension.

RQ 1. What are some basic descriptors of the SEC dataset?

Our first research question aims to provide a general descriptive overview of the SEC dataset. We focus on 10-K filings, as every public company is required to submit one 10-K annually. If a company needs to resubmit a corrected version, it is filed as a 10-K/A (amended 10-K).

Figure 1 illustrates the number of 10-K submissions per year. The data reveal a general decline in submissions over time, with a noticeable increase in 2021. This spike is likely attributable to the rise of Special Purpose Acquisition Companies (SPACs), commonly known as "blank check" companies, which became particularly popular during the pandemic. On average, approximately 6,000 10-Ks are submitted each year.

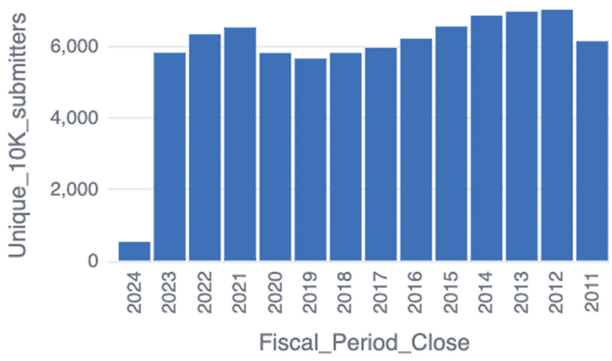


Figure 1: Unique 10-K submitters by year.

Many readers will have a good feel of what the market has been like recently, so by presenting this information first will hope to reduce their processing time. This is a reverse storytelling technique where starting with the present can aid in tracing patterns.

Figure 2 illustrates the number of unique industries reported in annual 10-K filings. Over time, there has been a slight but consistent decline in the number of unique industries, with around 400 different industries reported each year.

Figure 3 highlights the most common forms submitted to the SEC, including 11 of the most frequently filed forms in the dataset. Among them, Form 8-K, which can be submitted multiple times a year, reports significant events such as bankruptcies, acquisitions, changes in directors or officers, and shifts in financial condition. Meanwhile, 10-Qs are submitted quarterly when 10-Ks are not required for that period. As expected, the dataset reflects roughly three times as many 10-Q filings as 10-Ks.

The data also shows that 10-K/A filings (amended 10-Ks) are more common than amended 10-Q filings (10-Q/As). Specifically, 6,983 of 83,336 10-Ks were amended (8.3%), while only 11,604 of 256,934 10-Qs were amended (4.5%). This discrepancy could be attributed to the fact that 10-Ks are audited, while 10-Qs typically are not.

The breadth of information within the dataset is notable, capturing forms beyond the commonly analyzed 10-Ks and 10-Qs. For instance, data about Initial Public Offerings (Form S-1), governance structures (DEF 14 A), and international reports (Form 20-F) are also included. Interestingly, there are more amended S-1 forms (S-1/A) than

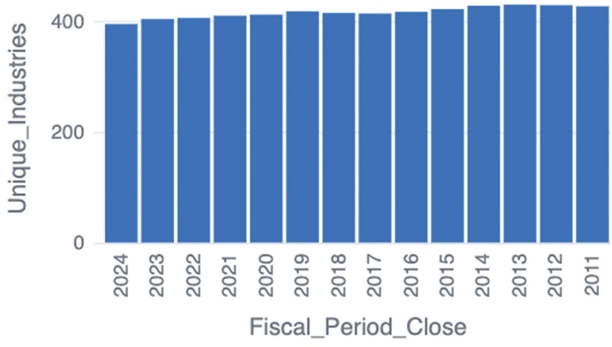


Figure 2: Unique industries by year.

form	num_forms	by_year
8-K	277,136	
10-Q	261,728	
10-K	83,518	
10-Q/A	11,737	
20-F	7,107	
10-K/A	7,005	
8-K/A	6,804	
DEF 14A	4,947	
S-1/A	4,298	
6-K	3,308	
S-1	3,255	

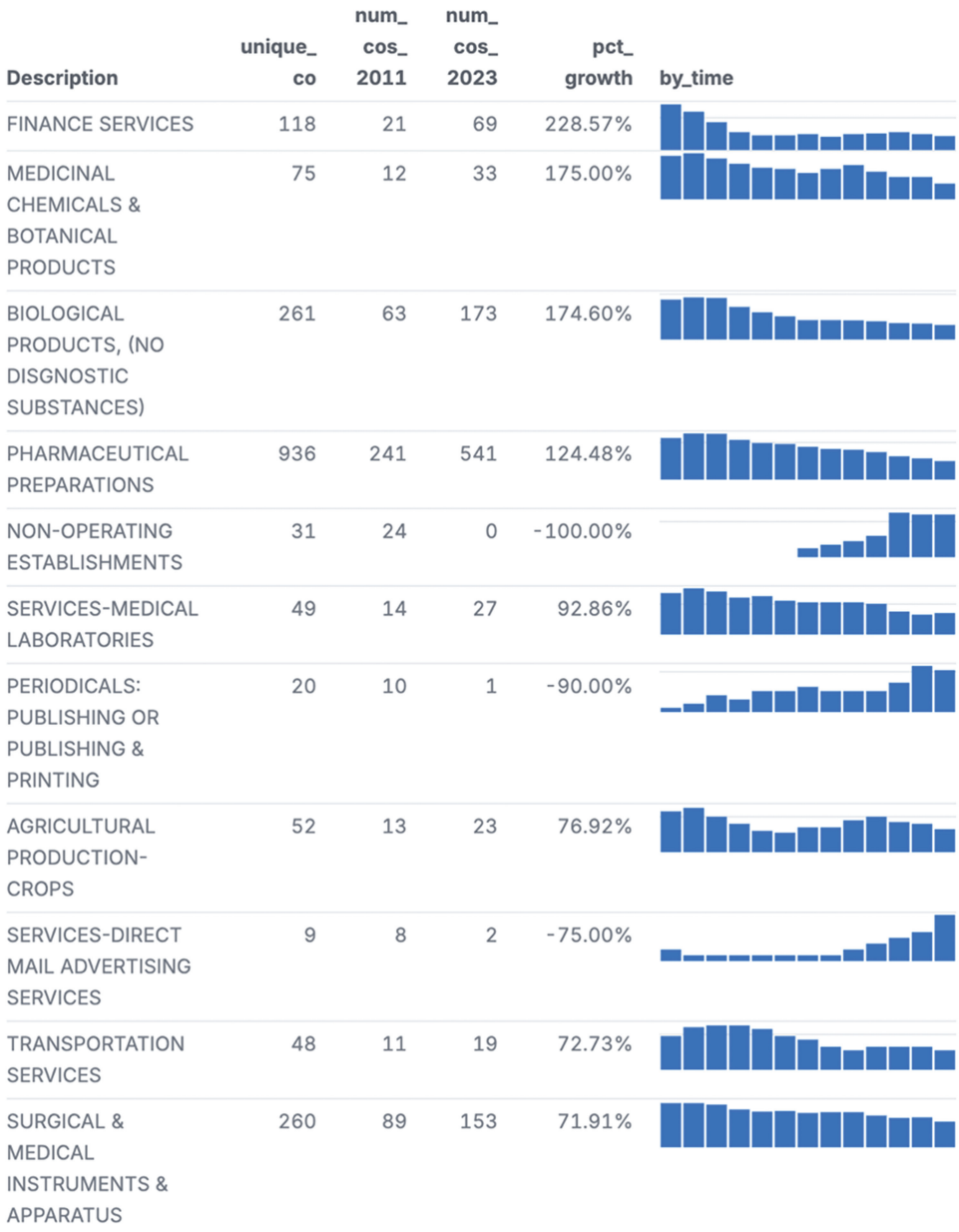
Figure 3: Most common forms in SEC financial statement dataset).

initial submissions, suggesting a topic worth further investigation. Additionally, the higher rate of amended 10-Ks compared to 10-Qs may indicate that audited forms require more revisions.

Figure 4 displays the percentage growth of industries based on the number of unique companies in the dataset (as reflected in the *unique_co* column). Each company is counted only once, regardless of whether they have submitted for a single year or for multiple years over the period.

The trend column, represented by a sparkline, visualizes the growth pattern for each industry from 2011 to 2023, with the most recent years presented first. The data highlights a sharp spike in Blank Check companies in 2021, a phenomenon tied to the global pandemic and the rise of SPACs.

Conversely, the data shows steady growth in the Pharmaceutical Preparations industry, reflecting its increasing prominence. In contrast, several industries, including Crude Petroleum and Natural Gas, State and National Commercial Banks, Real Estate Investment Trusts, Metal Mining, and Savings Institutions, have experienced steady



Description	unique_co	num_cos_2011	num_cos_2023	pct_growth	by_time
FINANCE SERVICES	118	21	69	228.57%	
MEDICINAL CHEMICALS & BOTANICAL PRODUCTS	75	12	33	175.00%	
BIOLOGICAL PRODUCTS, (NO DISGNOSTIC SUBSTANCES)	261	63	173	174.60%	
PHARMACEUTICAL PREPARATIONS	936	241	541	124.48%	
NON-OPERATING ESTABLISHMENTS	31	24	0	-100.00%	
SERVICES-MEDICAL LABORATORIES	49	14	27	92.86%	
PERIODICALS: PUBLISHING OR PUBLISHING & PRINTING	20	10	1	-90.00%	
AGRICULTURAL PRODUCTION-CROPS	52	13	23	76.92%	
SERVICES-DIRECT MAIL ADVERTISING SERVICES	9	8	2	-75.00%	
TRANSPORTATION SERVICES	48	11	19	72.73%	
SURGICAL & MEDICAL INSTRUMENTS & APPARATUS	260	89	153	71.91%	

Figure 4: Industries by absolute percentage growth.

declines. Additionally, the last line in the figure points to a recent trend of companies submitting 10-K filings with a missing standard industry code, which warrants further exploration.

RQ 2: Which industries file the most 8-Ks?

Figure 5 presents the industries with the highest ratio of 8-K filings per 10-K submission. Companies are required to file a Form 8-K within four business days of any event deemed material to investors. The data reveal a significant spike in 8-K filings for the Air Transportation industry during the pandemic, likely due to the volatile nature of the sector at that time. The number of unique companies group in the industry is shown (num_orgs), followed by the total number of 8-Ks submitted by members of the industry (num_8ks), followed by the number of 10-Ks for that industry between 2012 and 2023 (num_10ks), and the average number of 8-Ks submitted per company (per_co_8k), followed by a recent trendline of the number of 8-Ks for the past five years (by_8k_per_). For context, we found that 7,987 companies submitted 277,136 8-K forms, averaging almost 35 submissions between 2019 and 2024. We found that 5,625 companies submitted no 8-K forms.

Description	num_ orgs	num_ 8ks	num_ 10ks	per_ co_8k	trend_2024_2019
AIR TRANSPORTATION, SCHEDULED	25	1,022	176	42.583	
ELECTRIC & OTHER SERVICES COMBINED	56	1,597	398	36.295	
MOTOR VEHICLES & PASSENGER CAR BODIES	42	1,291	236	31.488	
WATER TRANSPORTATION	20	626	147	31.3	
TRUCKING (NO LOCAL)	25	763	227	30.52	
PIPE LINES (NO NATURAL GAS)	28	810	226	28.929	
RETAIL-AUTO DEALERS & GASOLINE STATIONS	41	1,184	267	28.878	
HOSPITAL & MEDICAL SERVICE PLANS	22	626	160	28.455	
INVESTMENT ADVICE	74	2,052	574	28.11	
NATURAL GAS DISTRIBUTION	30	815	201	28.103	

Figure 5: Industries with most 8-Ks per 10-K.

Additionally, the SEC began mandating 8-K filings in XBRL format for large filers starting in 2019, with all filers required to comply by 2021. This shift in reporting requirements explains why earlier data reflects relatively few filings before 2019.

RQ 3. Which industries make the most amendments to their 10-Ks?

Figure 6 illustrates the trend in amended 10-K filings over time. The surge in amended returns after 2019 may be attributed to both the challenges of operating during the pandemic and increased regulatory scrutiny. Figure 7 ranks industries by the percentage of amended 10-K filings in 2023, listing both the number of companies in each industry (unique_co), the total number of 10-Ks filed in that industry (total_filed_10Ks), the number of companies in the industry in 2023 (ind_size_2023), the percentage of companies that amended their return in 2023 (pct_a_2023) and the percentage of firms that have amended at least one return (pct_a_all).

Notably, Gold, Silver, and Mining companies exhibit a consistent level of amendments across the years, while the Pharmaceutical Preparations industry shows a more recent and sustained increase in amendments. This pattern may suggest heightened pressures in these industries, such as regulatory challenges or complex reporting environments.

RQ 4. Which industries change their reported mailing or business address most frequently?

We identified a company as having moved if they changed the street address, city, state, zip code, or country listed under either their "business address" or "mailing address" on their 10-K filings from one year to the next. Table 1 provides details on these address changes, revealing 6,223 address changes across 4,529 unique public companies.

Interestingly, 24.8% of these changes occurred within the same zip code, likely reflecting minor adjustments such as standardizing street abbreviations (e.g., "Blvd" to "Boulevard"). Overall, 66% of address changes remained within the same state, while 20% involved moves to a different state. Additionally, 2.4% of changes were from a U.S. address to an ex-U.S. location, while 3.4% reflected the reverse—ex-U.S. to U.S. Furthermore, 8% of address changes involved moves between two non-U.S. locations.

Next, we examined address changes by industry. Table 2 shows the industries with the highest percentage of cross-state address changes (out_state_moves). It also shows other metrics like in-state address changes, address change to the U.S. (to_usa), address changes out of the U.S. (out_of_usa, and ex_US_to_ex_US) Notably, industries such as Miscellaneous Retail, Agricultural Production, and Gold and Silver Ores show a high percentage of moves between non-U.S. locations, reflecting potential global mobility in these sectors.

We also assessed which industries change their business or mailing address most frequently. To do so, we calculated the percentage of 10-K filings that reported a change in business or mailing address by dividing the total number of address changes in each industry by the total number of 10-K filings in that industry. These results, limited to industries with at least 20 companies, are presented in Table 3 and Table 4 showing industries with both frequent and infrequent address changes. Both figures show the total number of unique companies in the industry, the number of address changes by industry, and the total number of submitted 10-Ks by industry. The last column shows the percent of 10-Ks that showed a change of address for the industry. A number of 14% would mean that 14 percent of 10-Ks submitted in that industry showed a change in address. We can see that Medicinal Chemicals and Retail-Misc are the industries that have recorded the highest percentages of address changes as reported on their 10-Ks.

The data reveal that banks, real estate, and insurance companies tend to report fewer address changes, while industries such as medicinal chemicals, pharmaceutical preparations, mining, retail, and various service sectors exhibit

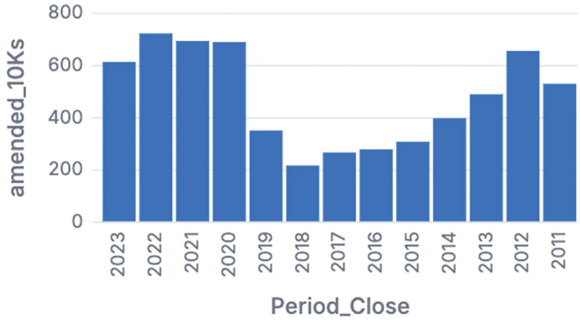
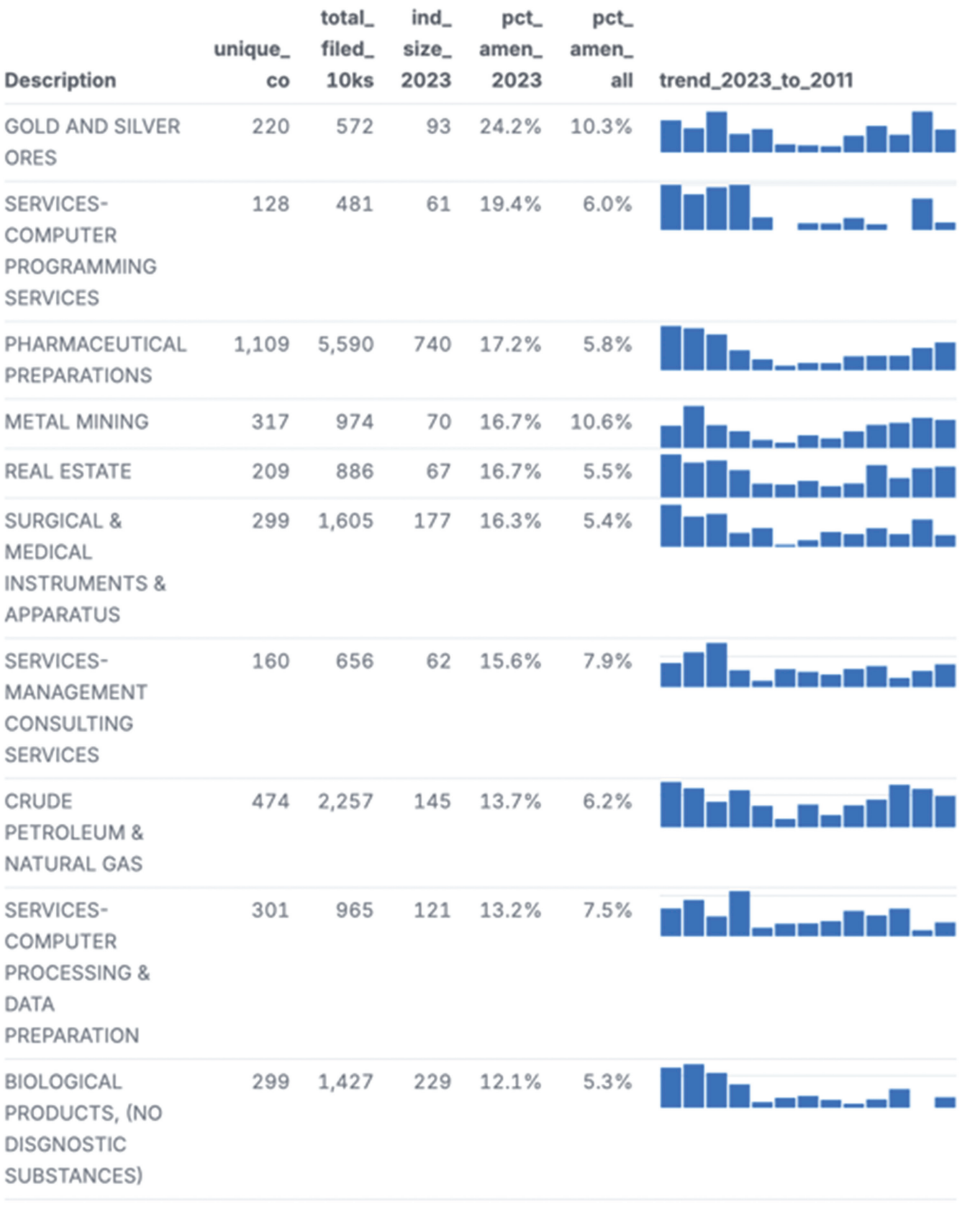


Figure 6: Amended 10-Ks by fiscal year.



Description	unique_ co	total_ filed_ 10ks	ind_ size_ 2023	pct_ amen_ 2023	pct_ amen_ all	trend_2023_to_2011
GOLD AND SILVER ORES	220	572	93	24.2%	10.3%	
SERVICES-COMPUTER PROGRAMMING SERVICES	128	481	61	19.4%	6.0%	
PHARMACEUTICAL PREPARATIONS	1,109	5,590	740	17.2%	5.8%	
METAL MINING	317	974	70	16.7%	10.6%	
REAL ESTATE	209	886	67	16.7%	5.5%	
SURGICAL & MEDICAL INSTRUMENTS & APPARATUS	299	1,605	177	16.3%	5.4%	
SERVICES-MANAGEMENT CONSULTING SERVICES	160	656	62	15.6%	7.9%	
CRUDE PETROLEUM & NATURAL GAS	474	2,257	145	13.7%	6.2%	
SERVICES-COMPUTER PROCESSING & DATA PREPARATION	301	965	121	13.2%	7.5%	
BIOLOGICAL PRODUCTS, (NO DISGNOSTIC SUBSTANCES)	299	1,427	229	12.1%	5.3%	

Figure 7: Industries highest number of amended 10-Ks.

Table 1: Details on address changes on 10-Ks.

total_moves	unique_orgs_ moving	pct_within_ zip_us	pct_in_ state_moves	pct_out_ state_moves	to_usa	out_of_ usa	ex_US_to_ ex_US
6,230	4,530	24.8%	66.2%	20.0%	3.4%	2.4%	8.0%

more frequent changes. This discrepancy may reflect the different strategic needs or regulatory environments across industries, with sectors like banking and insurance being more stable in terms of location compared to more dynamic sectors like pharmaceuticals and retail.

RQ 5: Which industries are transitioned into most often? Which industries are transitioned out of most often?

We identified 1,154 instances where companies changed their industry classification (SIC code) from one year to the next, representing 1.38% of the 83,518 total 10-K filings in our sample (Table 5). This provides insight into the frequency of industry transitions, allowing us to examine which industries experience the most shifts.

Table 2: Industries with highest percentage of cross-state address changes.

Industry	Total mvs	unique_ cos	in_state_ mvs	out_sta te_mvs	to_ usa	out_of_ usa	ex_US_to_ ex_US
Communications Equipment. NEC	35	21	48.6%	48.6%	2.9%	0.0%	0.0%
Services—Educational Services	39	29	53.8%	35.9%	2.6%	2.6%	5.1%
Patent Owners & Lessors	37	24	51.4%	35.1%	5.4%	5.4%	2.7%
Electromedical & Electrotherapeutic Apparatus	40	28	65.0%	35.0%	0.0%	0.0%	0.0%
Telephone Communications (No Radiotelephone)	34	27	52.9%	32.4%	2.9%	5.9%	5.9%
Retail—Miscellaneous Retail	45	26	42.2%	31.1%	6.7%	2.2%	17.8%
Services—Advertising	37	27	51.4%	29.7%	2.7%	5.4%	10.8%
Electric Services	48	34	56.3%	29.2%	0.0%	4.2%	10.4%
Blank Checks	159	133	50.9%	28.9%	3.1%	4.4%	12.6%
Agricultural Production—Crops	28	21	42.9%	28.6%	0.0%	0.0%	28.6%
Services—Computer Processing & Data Preparation	101	84	61.4%	26.7%	4.0%	2.0%	5.9%
Services—Computer Integrated Systems Design	54	39	66.7%	25.9%	1.9%	1.9%	3.7%
Gold and Silver Ores	51	39	33.3%	25.5%	9.8%	7.8%	23.5%
Commodity Contracts Brokers & Dealers	99	77	73.7%	25.3%	0.0%	1.0%	0.0%
Beverages	28	21	67.9%	25.0%	0.0%	7.1%	0.0%

Table 3: Industries with frequent address changes.

Industry	unique_cos	num_mvs	num_10ks	pct_10k_w_mv
Medicinal Chemicals & Botanical Products	34	46	317	14.5%
Retail—Miscellaneous Retail	26	45	315	14.3%
Agricultural Production—Crops	21	28	220	12.7%
Communications Services. NEC	31	42	337	12.5%
Patent Owners & Lessors	24	37	305	12.1%
Services—Help Supply Services	22	38	321	11.8%
Metal Mining	74	115	988	11.6%
Pharmaceutical Preparations	429	617	5651	10.9%
Services—Management Consulting Services	56	72	663	10.9%
Biological Products. (No Disgnostic Substances)	110	149	1447	10.3%
Services—Computer Processing & Data Preparation	84	101	984	10.3%
Services—Computer Programming. Data Processing. Etc.	49	63	619	10.2%
Communications Equipment. NEC	21	35	349	10.0%
Perfumes. Cosmetics & Other Toilet Preparations	22	27	278	9.7%
Services—Business Services. NEC	135	189	1967	9.6%
Services—Advertising	27	37	386	9.6%
Oil & Gas Field Services. NEC	27	34	357	9.5%

When submitting a 10-K, companies can change their SIC code annually. Table 6 highlights the most common industry transitions, presenting both the prior (Old Industry) and new industry classifications (New Industry) for companies that reclassified. The data shows that approximately 10% of all SPACs transitioned into the Pharmaceutical Preparations industry. Often, SPACs change their name and receive a new unique identifier, which

Table 4: Industries with infrequent address changes.

Industry	unique_cos	num_mvs	num_10ks	pct_10k_w_mv
Savings Institution. Federally Chartered	25	27	1,126	2.4%
State Commercial Banks	81	91	2,983	3.1%
National Commercial Banks	50	58	1,605	3.6%
Investment Advice	25	28	574	4.9%
Life Insurance	24	26	518	5.0%
Motor Vehicle Parts & Accessories	24	30	573	5.2%
Blank Checks	133	159	2,882	5.5%
Real Estate	36	51	889	5.7%
Electric Services	34	48	784	6.1%
Retail—Eating Places	36	47	758	6.2%
Fire. Marine & Casualty Insurance	45	59	893	6.6%
Telephone Communications (No Radiotelephone)	27	34	509	6.7%
Semiconductors & Related Devices	70	92	1,348	6.8%
Industrial Organic Chemicals	27	39	569	6.9%
Real Estate Investment Trusts	206	264	3,820	6.9%
Electromedical & Electrotherapeutic Apparatus	28	40	542	7.4%
Natural Gas Transmission	26	32	420	7.6%

Table 5: Number of industry transitions.

total_10Ks	num_sic_changes	pct_10ks_w_change
83,518	1,154	1.4%

Table 6: Common industry transitions from left description to right description.

Old Industry	num_industry_change	New Industry	unique_co
Blank Checks	212	Pharmaceutical Preparations	20
		Services—Prepackaged Software	14
		Services—Business Services NEC	12
Services—Business Services. NEC	52	Pharmaceutical Preparations	5
		Services—Computer Processing & Data Preparation	4
		Personal Credit Institutions	3
Services—Prepackaged Software	32	Services—Management Consulting Services	3
		Finance Services	3
		Surgical & Medical Instruments & Apparatus	3
Services—Computer Processing & Data Preparation	29	Finance Services	11
		Retail—Miscellaneous Retail	3
		Medicinal Chemicals & Botanical Products	2
Metal Mining	29	Pharmaceutical Preparations	4
		Services—Prepackaged Software	2
		Crude Petroleum & Natural Gas	2

would be recorded in the data as a new company rather than an industry transition. Additionally, 38% (11 out of 29) of companies in the Computer Processing & Data Preparation industry reclassified into the Finance Services industry, suggesting that some firms may transition to industries where they are perceived as more attractive investments.

We further examined which industries companies transitioned out of most frequently (Table 7) and which industries they transitioned into most often (Table 8). Unsurprisingly, Non-Operating Establishments frequently transitioned into operational industries, changing their SIC code accordingly. Similarly, Investors Not Elsewhere Classified (NEC) tended to reclassify their industry to better align with their business activities. SPACs often transitioned into a new industry after acquiring a company, although many changed their name and unique identifier in the process, which in some cases might be recorded as the end of reporting under their previous identity.

Companies may reclassify for a variety of reasons, including becoming more attractive to investors. The Patent Owners, Medicinal Chemicals, and Finance Services industry is a common destination, where 25% percent of the industry reclassified there from a previous industry. Likewise, the Patent Owners and Medicinal Chemicals industries saw a high percentage of companies transitioning into them, reflecting broader trends in these sectors.

Table 7: Industries transitioned away from (sorted by percent of companies that switched out of that industry).

old_desc	num_10ks	unique_cos	num_switching_out	pct_switched_out	totalcos_2011	totalcos_2023
Non-Operating Establishments	111	50	11	22.0%	8	2
Investors NEC	180	69	11	15.9%	12	9
Short-Term Business Credit Institutions	148	43	6	14.0%	8	8
Services—Motion Picture & Video Tape Production	193	95	13	13.7%	25	7
Services—Commercial Physical & Biological Research	335	107	14	13.1%	15	15
Agricultural Services	132	49	6	12.2%	6	12
Blank Checks	2,882	1,946	212	10.9%	324	240
Communications Services NEC	337	129	14	15.9%	22	24
Construction—Special Trade Contractors	180	65	7	10.8%	11	9
Computer Peripheral Equipment NEC	212	67	6	15.9%	13	15
Oil & Gas Field Exploration Services	314	95	8	8.4%	31	14
Services—Miscellaneous Business Services	196	72	6	8.3%	15	12
Services—Amusement & Recreation Services	259	110	9	8.2%	18	17
Retail—Drug Stores and Proprietary Stores	132	49	4	8.2%	14	7
Arrangement of Transportation of Freight & Cargo	142	37	3	8.1%	13	9
Plastics Products NEC	195	51	4	7.8%	14	12
Real Estate Agents & Managers (For Others)	193	78	6	7.7%	12	14
Retail—Eating & Drinking Places	144	53	4	7.5%	9	9
Instruments for Meas & Testing of Electricity & Elec Signals	190	41	3	7.3%	10	10
Refuse Systems	150	41	3	7.3%	7	11

Table 8: Industries transitioned into (sorted by percent of companies that switched into the industry from another industry).

Industry	num_ 10ks	unique_ cos	num_ switching_in	pct_ switched_in	totalcos_ 2011	totalcos_ 2023
Patent Owners & Lessors	305	50	14	28.0%	24	16
Medicinal Chemicals & Botanical Products	317	76	19	25.0%	12	33
Finance Services	412	118	29	24.6%	21	69
Mining & Quarrying of Nonmetallic Minerals (No Fuels)	295	49	10	20.4%	21	19
Retail—Miscellaneous Retail	315	63	12	19.0%	21	19
Construction—Special Trade Contractors	180	37	7	18.9%	10	12
Transportation Services	228	48	8	16.7%	11	19
Services—Miscellaneous Amusement & Recreation	268	63	10	15.9%	15	25
Beverages	296	66	10	15.2%	20	24
Services—Medical Laboratories	285	49	7	14.3%	14	27
Services—Advertising	386	78	11	14.1%	24	21
Communications Services NEC	337	72	10	13.9%	36	20
Cable & Other Pay Television Services	318	51	7	13.7%	27	15
Oil & Gas Field Services NEC	357	60	8	13.3%	23	21
Insurance Agents Brokers & Service	249	39	5	12.8%	19	21
Services—Management Consulting Services	663	135	17	12.6%	34	45
Petroleum Refining	288	40	5	12.5%	23	15

5. Discussion

5.1. Strengths and Weaknesses

This study leverages a large dataset comprising over 83,000 10-K filings from a publicly available source, providing a strong foundation for analysis. The use of both the public dataset and the open-source data language Malloy enhances the transparency and replicability of the research, enabling future studies to build on or validate the findings.Our research offers a novel perspective by examining address changes and industry reclassifications as proxies for industry pressure or turbulence—areas that are underexplored in current literature. By focusing on trends over time, we uncover patterns in industry behavior and highlight emerging topics of interest. For practitioners, such as investors, regulators, and policymakers, these findings offer critical insights into industry dynamics and the evaluation of business environments.

However, the data used in this study only covers a 12-year period (2012–2023), which may limit the generalizability of the findings, as industries and national economies follow growth cycles. Another limitation is the focus on basic descriptive information about the 10-K submitters, without getting into more detailed financial concepts, such as Net Income or Assets. Additionally, SIC codes, which are self-reported, may not always accurately reflect a company's business activities.

6. Future Research

One potential avenue for future research is the development of an "Industry Stress Index." This index could aggregate multiple factors—such as address changes, industry reclassifications, 8-K filing frequency, and 10-K amendment rates—into a comprehensive measure of industry pressure and volatility. Further studies could explore how to weight these factors in the index and assess its predictive power for industry performance and trends. Additionally, since SIC codes are hierarchical, future research could extend the analysis by investigating trends at higher levels of classification, rather than focusing solely on the lowest (four-digit) level.

Future studies may examine these concepts in markets outside of the U.S., as the concept of industry pressures should exist in any capitalist market. If other markets provide well-structured data, as the SEC does, the same analysis should be possible.

Future research could also explore the broader capabilities of Malloy in data analysis, either with this dataset or others. This might include comparative studies of Malloy's performance and usability against other tools, investigations into its applications across various research domains, and the establishment of best practices for its effective use.

We observed that certain industries contain a high proportion of companies that have reclassified themselves into those sectors. Further research could examine the drivers of such transitions, including mergers and acquisitions, technological innovation, regulatory shifts, and changes in consumer demand. Understanding these dynamics would provide valuable insights for investors, policymakers, and businesses in these evolving industries.While we do not attempt to uncover the nature, severity, or genesis of the 10-K amendment, this is a ripe area for future research. Advanced SQL or Malloy queries should be able to identify precise differences between the original and amended submissions.

Similarly, some industries show a greater frequency of address changes. Future research could explore the strategic factors behind these changes, such as tax optimization, regulatory arbitrage, market expansion, and talent acquisition. Further research is necessary on the mechanisms causing address changes and how this differs across industries. In some industries address changes may signal growth, while in other it may signal times of stress. Further research is necessary to tease out these differences. This could offer deeper insights into the decision-making processes behind corporate relocations and their broader implications.

7. Conclusion

Several key trends emerge from the analysis, particularly the overlap between industries that frequently transition into new classifications and those that often change their addresses. Industries such as Patent Owners, Medicinal Products, and Miscellaneous Retail are commonly transitioned into and exhibit frequent address changes, which may suggest a high level of mergers and acquisitions in these sectors. Similarly, industries like Metal Mining, Gold and Silver Ores, Pharmaceutical Preparations, and Services – Programming also report frequent address changes alongside a high number of amended returns, further indicating potential industry pressures or volatility.

While industries have been compared using a variety of metrics in previous research, this study introduces two new indicators—address changes and industry reclassifications—that have not been widely explored in the literature. These factors provide valuable insights into industry-level pressures and may serve as proxies for understanding external pressures. Future research could build on this work by integrating multiple factors to create an industry stress index.

Additionally, this study showcases a novel methodological tool, Malloy, which has yet to be widely adopted in academic research. Malloy offers significant advantages for replicating analyses over traditional software like Tableau or PowerBI, as its code can be easily shared and re-executed. Future research is needed to further explore and demonstrate the full capabilities and utility of Malloy for large-scale data analysis.

References

Baaij, M. G., T. J. M. Mom, F. A. J. Van den Bosch, and H. W. Volberda. 2015. "Why Do Multinational Corporations Relocate Core Parts of Their Corporate Headquarters Abroad?" *Long Range Planning* **48**, no. 1: 46–58.

Ben-Rephael, A., Z. Da, P. D. Easton, and R. D. Israelsen. 2022. "Who Pays Attention to SEC Form 8-K?" *The Accounting Review* **97**, no. 5: 59–88.

Bhojraj, S., C. M. C. Lee, and D. K. Oler. 2003. "What's My Line? A Comparison of Industry Classification Schemes for Capital Market Research." *Journal of Accounting Research* **41**, no. 5: 745–774.

Birkinshaw, J., P. Braunerhjelm, U. Holm, and S. Terjesen. 2006. "Why Do Some Multinational Corporations Relocate Their Headquarters Overseas?" *Strategic Management Journal* **27**, no. 7: 681–700.

Cassell, C. A., L. M. Cunningham, and L. L. Lisic. 2019. "The Readability of Company Responses to SEC Comment Letters and SEC 10-K Filing Review Outcomes." *Review of Accounting Studies* **24**, no. 4: 1252–1276.

Chakri, P., Pratap, S. Lakshay, Gouda. and S. K. 2023. "An Exploratory Data Analysis Approach for Analyzing Financial Accounting Data Using Machine Learning." *Decision Analytics Journal* **7**: 100212.

Curling, M. L. 2006. Mandatory 10-K amendments and strategic disclosure: An examination of firms under review by the Securities and Exchange Commission [Ph.D., The Pennsylvania State University]. https://www.proquest.com/docview/305247997/abstract/E66A6A17526C40A0PQ/1

Fama, E. F., and K. R. French. 1997. "Industry Costs of Equity." *Journal of Financial Economics* **43**, no. 2: 153–193.

Gaspar, J.-M., S. Wang, and L. Xu. 2024. "Digitalization and the Performance of Non-Technological Firms: Evidence from the COVID-19 and Natural Disaster Shocks." *Journal of Corporate Finance* **89**: 102670.

Gompers, P. A. 1997. "Optimal Investment, Monitoring, and the Staging of Venture Capital." In *Venture Capital*. Routledge: London, United Kingdom.

Gregory, R., J. R. Lombard, and B. Seifert. 2005. "Impact of Headquarters Relocation on the Operating Performance of the Firm." *Economic Development Quarterly* **19**, no. 3: 260–270.

Klier, T., and W. Testa. 2002. "Location Trends of Large Company Headquarters during the 1990s." *Economic Perspectives* **26**, no. 2: 12–27.

Komorowski, M., D. C. Marshall, J. D. Salciccioli, and Y. Crutain 2016. "Exploratory Data Analysis." In *Secondary Analysis of Electronic Health Records*, 185–203. Springer International Publishing: Cham, Switzerland. https://doi.org/10.1007/978-3-319-43742-2_15

Krishnan, G. V., and Y. Zhang. 2014. "Is There a Relation Between Audit Fee Cuts During the Global Financial Crisis and Banks' Financial Reporting Quality?" *Journal of Accounting and Public Policy* **33**, no. 3: 279–300.

Lerman, A., and J. Livnat. 2010. "The New Form 8-K Disclosures." *Review of Accounting Studies* **15**, no. 4: 752–778.

Li, F., R. Lundholm, and M. Minnis. 2013. "A Measure of Competition Based on 10-K Filings." *Journal of Accounting Research* **51**, no. 2: 399–436.

McMullin, J. L., B. P. Miller, and B. J. Twedt. 2019. "Increased Mandated Disclosure Frequency and Price Formation: Evidence from the 8-K Expansion Regulation." *Review of Accounting Studies* **24**, no. 1: 1–33.

Nielsen, S. 2022. "Management Accounting and the Concepts of Exploratory Data Analysis and Unsupervised Machine Learning: A Literature Study and Future Directions." *Journal of Accounting & Organizational Change* **18**, no. 5: 811–853.

Phillips, R. L., and R. Ormsby. 2016. "Industry Classification Schemes: An Analysis and Review." *Journal of Business & Finance Librarianship* **21**, no. 1: 1–25.

Schroeder, J., and P. N. Posch. 2023. "Anomalies, Trends and Patterns in Disclosure Activities: Understanding EDGAR." (SSRN Scholarly Paper No. 4664431). Social Science Research Network. https://doi.org/10.2139/ssrn.4664431

Thompson, R. A. 2023. "Reporting Misstatements as Revisions: An Evaluation of Managers' Use of Materiality Discretion." *Contemporary Accounting Research* **40**, no. 4: 2745–2784.

Tosun, O. K., and S. K. Moon. 2024. "Socially Responsible Investment Funds and Firm Performance Improvement." Review of Quantitative Finance and Accounting.

Vanneste, B. S. 2017. "How Much Do Industry, Corporation, and Business Matter, Really? A Meta-Analysis." *Strategy Science* **2**, no. 2: 121–139.

Wang, H., and R. Coff. 2022. "On the Matter of How Much Industry Matters." *Strategic Management Review* **3**, no. 2: 295–323.

Watkins, J. 2022. "Consequences of Prescribed Disclosure Timeliness: Evidence from Acceleration of the Form 8-K Filing Deadline." *The Accounting Review* **97**, no. 7: 429–463.

Appendix A

All Code is available here: https://github.dev/mrtimo/IndustryStress

Queries can be run by using this method: https://docs.malloydata.dev/documentation/user_guides/basic.html

1. Logging into GitHub
2. Go to the repository: https://github.dev/mrtimo/IndustryStress
3. Press the period key – "."
4. Install the Malloy Extension
5. Open the .malloynb file and press "Play" on the first cell

Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 3, No. 1, 2025 DOI: 10.54116/jbdai.v3i1.29

MULTINATIONAL INVESTMENT UNDER UNCERTAINTY

Priya Nagaraj
William Paterson University
NAGARAJP1@wpunj.edu

Michi Nishihara
Osaka University
nishihara@econ.osaka-u.ac.jp

Chuanqian Zhang
William Paterson University
zhangc4@wpunj.edu

ABSTRACT

This paper builds a real options model to quantify multinational investment timing decisions under both foreign market demand and exchange rate dynamics, which are largely overlooked in academia yet very common in the real world of international investments. We find that (1) a domestic firm may prefer to undertake foreign direct investment (FDI) under an exchange rate depreciation environment provided high foreign demand and domestically sourced investment costs. (2) Both exchange rate and demand uncertainties could have either positive or negative impacts on international investments, depending on their correlations and the relative dominance between "real option effect" and "revenue effect." A simple simulation exercise confirms model predictions and shows that generally the impact of demand uncertainty should be more prominent than that of exchange rate uncertainty.

JEL: *F21, G31.*

Keywords: *real options model, demand uncertainty, exchange rate uncertainty, FDI.*

1. Introduction

Multinational investments play an important role in economic growth and over the last several decades, the amount of global foreign investment has steadily increased. The decision to invest abroad however, is not an easy one for a multinational enterprise (MNE). According to the international economics and business literature (Conconi et al. 2016), the process of firm internationalization generally evolves from exporting, to building overseas distribution facilities, and eventually foreign production. An important factor that affects the decision to invest abroad is market demand uncertainty. Given the sunk cost of foreign investment, the minimum market demand required by MNEs to invest abroad, is significantly high. Additionally, the foreign investment decision is also influenced by the variation of foreign exchange rate as it affects the home currency denominated profits, both short-run and long-run. Although in prior literature, scholars have investigated the impact of these two factors—market demand uncertainty and exchange rate variations—on multinational investment, they have so far been studied separately. Few scholars have analyzed their joint impact, particularly in a dynamic environment. This paper tries to fill that gap.

The theoretical model we propose applies the canonical real options framework to multinational investment decisions. Real options theory (ROT) has been applied extensively in research in international economics and management studies.[1] ROT posits that optimal capital investment decision under uncertainty is equivalent to solving the optimal exercise timing of the American-type options, given that the output commodities can be traded across complete markets.

Our model is a two-country model, home (or source) and foreign (or host/destination), with one monopolistic firm headquartered in the home country, producing a homogeneous commodity to serve only the foreign market. The firm begins with exporting to the foreign market and eventually needs to make a decision on when to establish affiliates to produce in the foreign country to serve the local demand there. The timing of overseas investment is a function of exchange rate and foreign demand, both of which vary with time in a random fashion. The exchange rate is defined as the home currency value of foreign currency. The operating mode in both countries is such that the firm can freely adjust its production output (zero or full capacity) in response to the uncertain demand and currency evolution. The MNE manager makes decisions on the operating status, as well as the timing of when production is shifted to the offshore destination. In the benchmark case, we assume both countries have unlimited production capacities and then extend it to asymmetric capacity limits. The model delivers following results.

First, the relation between exchange rate movement and foreign investment depends on how the MNEs source the investment, the relative strength of its currency, and its production status before the foreign investment is undertaken. When the investment is sourced locally in the destination country, appreciating home currency will always accelerate the timing of investment outflow. This is intuitive because a higher value of home currency will make the foreign project more attractive and the investment cost cheaper. However, when the investment is sourced from the home country, the result is mixed. On the one hand, when the home currency is relatively weak, depreciating home currency will have a negative impact on foreign direct investment (FDI) because it increases the advantage of production in the home country. On the other hand, when the home currency is relatively strong, depreciating home currency will have a positive impact on FDI because it increases the cash flow value in the foreign country and the cost of foreign investment will be lower.

Second, foreign demand uncertainty and exchange rate uncertainty have either a positive or a negative impact on cross-border investment, depending on the correlation between the two shocks. When the correlation is positive, both uncertainties may either accelerate or delay the investment timing. On the one hand, a positive correlation indicates comovement of both variations. Since in our model higher exchange rate means lower home currency value, the comovement mitigates the variation of foreign revenues and thereby accelerates the investment. We call this the "revenue channel," which has been documented by Goldberg and Kolstad (1995) as well.[2] On the other hand, the existence of demand and exchange rate uncertainties will also give rise to the "real option" value of entering the foreign market, thereby delaying the entry, i.e., the "real option channel." The final outcome depends on the relative dominance of the two forces. However, when the correlation is negative, both uncertainties will deter FDI flow to the foreign country. This is expected since now the variation of foreign revenue will increase and the real option effect is operative. Finally, our model also shows that the relation between the two types of uncertainties and FDI depends on the operating status of the exporter. In particular, when the exporter is at suspension, both types of uncertainties will delay foreign investment, regardless of the correlation level. This is expected because at this stage the firm does not have cash-flow based assets, thus it will be impacted more by the "real options" channel.

In addition to the theoretical model, we also perform a simulation exercise. The simulation is necessary because it can straightforwardly present the possibility of production status and investments. Our simulation results are consistent the model's trigger-based predictions. More importantly, the simulation also illustrates that the impact of demand uncertainty should be more prominent than that of exchange rate uncertainty on the investment probability using our carefully selected parameters. This observation is in line with Choi and Jiang (2009) and Nguyen et al. (2018).

The paper is organized in the following manner. Section 2 presents the literature review. Section 3 lays out some empirical motivation for modelling work. Section 4 elucidates the assumptions and framework of the dynamic model. It also contains numerical solutions. Section 5 performs a simple simulation exercise based on the model-guided solutions. Section 6 concludes the paper.

[1]For a review of ROT applied in investment decision, see Dixit and Pindyck (1994). For the particular application of ROT in international business, see Chi et al. (2019) and Song et al. (2015) for a comprehensive review. Indeed, several recent empirical papers highlight the impact of real options on the multinationality, such as Aabo et al. (2016) and Belderbos et al. (2019).

[2]Note that Goldberg and Kolstad (1995) has different settings from ours. They define the correlation between two levels while we define it between two random shocks. However, despite the different expressions, the relation between covariance and correlation is generally similar.

2. Literature Review

The determinants of FDI decision have been investigated extensively in recent decades. A comprehensive review of the relation between various factors and FDI can be found in Blonigen (2005). Here we only discuss the literature that relates to the impact of exchange rate and demand dynamics on FDI.

First, some prior work shows that higher home currency value (relative to the destination country) will promote FDI outflow. This result was initially documented in Froot and Stein (1991). The reason is simple: higher home currency value will make the MNE wealthier and therefore it will be more likely to acquire foreign assets. Additionally, a depreciating currency will make foreign country's assets cheaper and more likely to be acquired by global firms. However, our results show that this result is mixed and usually it is insignificant due to the interaction between the stochastic environment and the sources of investment costs.

Second, the positive (or nonlinear) impact of currency volatility on FDI has been explored before. For example, Goldberg and Kolstad (1995) show that when the investors are risk-averse and there is negative correlation between exchange rate and demand shocks, the currency volatility could have a positive impact on the share of foreign production. However, when the correlation is positive, the firm will maintain 100% capacity overseas. Darby et al. (1999) theoretically find that exchange rate variation could either accelerate or delay foreign investments depending on the economic parameters and country types. Jeanneret (2016) explores the channel of firm heterogeneity and shows that MNE with high productivity may take the benefit of currency volatility to engage in FDI. Due to this, he finds a U-shaped relation between FDI and exchange rate volatility.

Third, for (country-level) demand uncertainty and foreign investment, Goldberg and Kolstad (1995) show mixed signs of demand uncertainty on FDI (see their equation A.2a) for risk-averse investors. Empirically, Conconi et al. (2016) study the influence of host countries' uncertainty on horizontal FDI decision using detailed Belgium firms' data. They show that the probability that a firm engages in foreign investment increases with its export experience. In more uncertain destinations, firms delay FDI entry, experimenting longer with exports before establishing foreign affiliates.

So far, Goldberg and Kolstad (1995) have similar setting as our analysis with regard to the joint effect of economic uncertainties (e.g., exchange rate and foreign demand). However, we have several critical differences. (1) We focus on modelling the nominal exchange rate instead of the real rate. Although real exchange rate takes into account the purchasing power and will impact the trade balance, nominal rate is more relevant for firm-level and short-term operations. (2) We build a dynamic model and derive its solutions in the context of timing of irreversible investment while they focus on the FDI production share. (3) We focus on the channel of interaction between economic uncertainties and operating flexibility instead of risk-aversion.

Our paper also contributes to the literature on the theory of foreign investment decisions. Rob and Vettas (2003) investigate optimal share of exports and FDI under growing demand and foreign capacity limits. Similar to our analysis, they consider both the investment irreversibility and capacity underutilization. They focus on the interior solution for the coexistence of exporting and foreign investment. Perhaps most distinct from ours is that they assume the demand either increases or stays put while we allow stochastic movement with uncertainties. Aray and Gardeazabal (2010) and Sung and Lapan (2000) investigate the impact of exchange rate on foreign investment under product market competition. Jeanneret (2016) examines the interaction between MNE's productivity heterogeneity and exchange rate uncertainty on FDI. However, unlike our analysis, these papers focus on either exchange rate uncertainty or demand uncertainty, and not both of them together.

3. Empirical Motivation

Before proceeding to the model part, we briefly present some empirical observations in this section. The goal of this exercise does not attempt to demonstrate any "stylized facts" to be proved by the model because there are significant gaps between the real data and model assumptions (listed in Section 4). For example, we only have country-level observations while the model is based on firm-level. The results from aggregating firms (e.g., country-level) could be different from single firms. Moreover, our data could not differentiate between horizontal and vertical investments while the model focuses the former type. Instead, our goal is to demonstrate the counterintuitive and ambiguous impacts of exchange rate movement and its volatility, and demand volatility on bilateral FDI, and therefore elucidate the importance of modelling firm decisions under multifactors. In what follows, we first discuss data background and then directly go to results summary. All empirical details are located in Appendix A.

3.1. Data Background

We sample bilateral FDI between the U.S. and its seven major FDI partners (United Kingdom, Japan, Canada, Germany, Netherlands, Switzerland, and France). These countries are either the recipients of the highest amount of FDI from the U.S. or have been the largest investors into U.S. We limit our analysis to the top seven partners for the years 1982 to 2018. Of the more than 200 countries investing in the U.S., these seven countries consistently remained as the top sources of FDI into the U.S. for the years considered and contributed about 70% of the total investment every year. Similarly, out of the many destination countries for U.S. FDI, these countries feature among the top 10 and receive about 50% of the total U.S. outward investment.

We focus on the FDI data in the manufacturing industries and exclude others such as financial and utility sectors.[3] Bureau of Economic Analysis (BEA) provides the stock of FDI for the year. We convert it into a flow-based measure by taking the difference in the stock of FDI between two years. We normalize the FDI variable by dividing it by gross domestic product (GDP).

3.2. Results Summary

(1) Appreciating home currency seldom promotes FDI outflow, the only two examples are U.S. to Japan and Germany to U.S.[4] Depreciating home currency sometimes increase FDI outflow. This scenario holds for five pairs: U.S. to Switzerland, U.K. to U.S., Japan to U.S., Switzerland to U.S., France to U.S. The rest of seven pairs show that exchange returns have insignificant impact on the FDI flows.

(2) The exchange rate volatility does not have significant impacts on the FDI decision in most cases. When the impact is significant, e.g., U.S. to Japan and U.S. to Switzerland, the signs are opposite.

(3) The foreign demand volatility, estimated indirectly by the GDP growth volatility, has either negative or insignificant impact. The negative impact holds for only three pairs: U.S. to Japan, U.S. to Switzerland, and Germany to U.S.

Although the empirical exercise is limited to certain country pairs, we can sense that the results are contrary to conventional wisdom in most times. Thus, it will be very necessary to develop a microlevel model to understand how these factors impact international investments.

4. Model Set Up

We assume a two-country (e.g., home and foreign), continuous-time and a partial equilibrium economy. A monopolistic firm whose headquarters is located in the home country starts to serve the foreign market through exporting. It possesses an option to establish affiliates abroad, to serve the local market. To simplify the analysis, we make several assumptions:

(1) The foreign investment is of horizontal nature. The firm at the home country makes foreign investment timing decision to maximize its pre-FDI value in terms of home currency.

(2) The home-based MNE only serves foreign market. It serves neither home country nor other foreign countries.

(3) Offshoring production is considered irreversible within the model.

(4) The firm chooses to produce by either exporting or building affiliates abroad but never adopts both strategies at the same time. This assumption enables us to focus on investment timing decision instead of solving for the optimal share of foreign production, which has been addressed in other papers such as Goldberg and Kolstad (1995) and Rob and Vettas (2003).

(5) The sunk cost of foreign investment is sourced from either home country and priced with home currency as in Jeanneret (2016), or foreign country and priced with foreign currency as in Aray and Gardeazabal (2010). In the latter case, the firm valuates the FDI cost in terms of the home currency. We will compare the results from these two scenarios as robustness check. In the following paragraphs, we introduce the model setup.

[3]We do this because our structural model in the later part is built on capacity investment, which is more consistent with manufactory sector. Such treatment has been widely applied in the past literatures. However, our data also shows that FDI of all industries also present similar results.

[4]Note that we define the exchange rate as the number of foreign currencies per USD for all cases. When U.S. is the "home country," the increase of the exchange return indicates dollar becomes stronger. When U.S. is the foreign destination, the increase of return suggests foreign currencies become weaker.

Table 1: Benchmark parameters calibration.

Symbol	Economic Determination	Value
rh	Risk-free rate of home country	0.04
rf	Risk-free rate of foreign country	0.07
γ	Demand elasticity	1
μ	Expected drift of demand	0
σS	Exchange rate volatility	0.10
$\sigma\theta$	Demand volatility	0.10
τ	Ice-berg transportation cost	0.7
I	Suck cost of foreign investment	1
vh	Production cost at home country	0.5
vf	Production cost at foreign country	0.5

The demand function for the foreign market is determined by the following linear equation:

$$P = \theta - \gamma q \tag{1}$$

here P is the local price in term of foreign currency, γ represents the (constant) elasticity of demand, θ is the foreign demand shock following Brownian motion.[5]

$$\frac{d\theta}{\theta} = \mu dt + \sigma_\theta dZ^Q \tag{2}$$

here μ is the expected growth of foreign demand and σ_θ captures the standard deviation.

The exchange rate,[6] defined by *number of home currency per unit of foreign currency*, evolves in a stochastic fashion under the risk-neutral probability measure ($\mathbb{Q}$ measure), for the domestic investor.

$$\frac{dS}{S} = (r_h - r_f)dt + \sigma_S dW^Q \tag{3}$$

where rh and rf are the domestic and foreign risk-free rate at which money can be borrowed or lent; σ_S is a positive constant representing standard derivation, and $(Wt)t \geq 0$ is a standard Brownian Motion process. Time is continuous and varies over [0, ∞]. Uncertainty is represented by the filtered probability space $(\Omega, F, (F_t)_{t\geq 0}, Q)$ over which all stochastic processes are defined. However, we essentially assume that the expected rate of return and standard deviation of exchange rate are constants. This assumption deviates from the real world, which might have a time-varying pattern instead of constants. We do so to facilitate the computation of the model. The derivation for Equation (3) is presented in Appendix B. The correlation coefficient between the variations of foreign demand and exchange rate is denoted as ρ, e.g., $dZdW = \rho dt$. As shown in Table 1, the correlation varies in a wide range among the selected countries, which necessitates its role in our model to explain the empirical facts.

Consider the constant marginal cost of production in home country as v_{h} and the production cost in foreign country is denoted as v_{f}, therefore the instantaneous revenue for home production to export (exporter "X" status) can be written as

$$\pi_X = (\tau SP - v_h)q = (\tau S\theta - \tau S\gamma q - v_h)q \tag{4}$$

here τ captures the revenue loss caused by the ice-berg type transportation cost and q is the production output. We assume there is no capacity limit, thus the optimal production can be derived by ${q_X}^* = \frac{\tau S\theta - v_h}{2\tau S\gamma}$. Since the optimal quantity cannot be negative, we get the profit flow as

$$\pi_X = \begin{cases} \dfrac{(\tau S_t\theta_t - v_h)^2}{4\tau S_t\gamma} & \theta_t > \dfrac{v_h}{\tau S_t} \\ 0 & \theta_t < \dfrac{v_h}{\tau S_t} \end{cases} \tag{5}$$

[5]The linear demand function has been used in Sung and Lapan (2000), Rob and Vettas (2003), Aray and Gardeazabal (2010), and Conconi et al. (2016) mostly for its modelling convenience. Another strand of literature that includes Helpman et al. (2004), and Jeanneret (2016) employs CES aggregate price because they focus on the impact of firm heterogeneity of productivities. The former setting is more suitable and tractable for our paper since our model is built on capacity investment.

[6]The exchange rate here is meant to nominal exchange rate because the model is not explicitly feed into price level or inflation discrepancy. As a matter of fact, there isn't clear boundary between nominal and real exchange rate when modeling the international business, and in most cases, when the commodity price is relatively sticky, both of the two rates are closely correlated and not much different in the econometric perspective (Clark et al. 2004, Section III)

It is noteworthy that the operating/suspension boundary is determined by a nonlinear boundary governed by two factors (θ_t, S_t). This indicates that the exporter will produce only when the demand for the product and the exchange rate are high enough $\theta_t S_t > \frac{v_h}{\tau}$. Notice that whenever the exchange rate is relatively small (i.e., the home currency is strong), domestic production will be more likely to suspend due to the high demand bar to restart. This is intuitive because the strong currency will hurt the export sector.

Similarly, profit flow for foreign affiliates (FDI "*F*" status) can be written as

$$\pi_F = S(P - v_f)q = S(\theta - \gamma q - v_f)q \tag{6}$$

and the optimal output quantity is ${q_F}^* = \frac{\theta - v_f}{2\gamma}$ since the optimal quantity cannot be negative, we get profit flow as

$$\pi_F = \begin{cases} \frac{S_t(\theta_t - v_f)^2}{4\gamma} & \theta > v_f \quad 0 \quad \theta < v_f \end{cases} \tag{7}$$

The foreign production process is only dependent on market demand because we assume all products will be sold locally although the MNE's consolidated revenue is still impacted by the exchange rate.

Our goal is to solve for the timing of foreign investment. Following standard protocol, we first present the valuation method for foreign affiliates and then we present the valuation for the exporter.

The value of foreign affiliate V_{F} is governed by the following partial differential equation (PDE) (the derivation is in Appendix B):

$$\frac{1}{2}{\sigma_S}^2 S^2 \frac{\partial^2 V_F}{\partial S^2} + \frac{1}{2}{\sigma_\theta}^2 \theta^2 \frac{\partial^2 V_F}{\partial \theta^2} + \rho\sigma_S\sigma_\theta S\theta \frac{\partial^2 V_F}{\partial S \partial \theta} + (r_h - r_f) S \frac{\partial V_F}{\partial S} + \mu\theta \frac{\partial V_F}{\partial \theta} + \pi_F = r_h V_F \tag{8}$$

Although the equation appears complex, it delivers clear economic sense. The first and second item on the left-hand side capture the instantaneous effect of volatilities of exchange rate and foreign demand on foreign assets' value; the third item evaluates the correlation effect between the two dynamic evolutions; the fourth and fifth term capture the instant effect of expected moving rate and the last term is the profit flow. The equation suggests that the entire instantaneous returns should be equal to the risk-free rate at home currency, which is expected as we employ a risk-neutral valuation. In the following, we introduce appropriate boundary conditions to solve the optimal stochastic control problem.

First, when the foreign demand declines to zero the value of the foreign affiliate in terms of home currency should also be zero, no matter what the exchange rate. This is intuitive because profit flow increases linearly with S_{t}:

$$V_F(\theta \downarrow 0, S) = 0 \tag{9}$$

Second, when the exchange rate is extremely low, or the foreign currency is very weak, the home currency value of the foreign counterpart also approaches zero, for the same reason as the previous one.

$$V_F(\theta, S \downarrow 0) = 0 \tag{10}$$

Third, when demand is very large, the affiliate will not suspend and the value can be conveniently obtained

$$V_F(\theta \uparrow \infty, S) = \Pi_F \tag{11}$$

here Π_F is the present value of all future profits of the affiliate in a risk-neutral measure and can be written as

$$\Pi_F = E^Q \int_t^\infty e^{-r_h(s-t)} \pi_F ds = \frac{S}{4\gamma}\left(\frac{\theta^2}{r_f - 2\mu - {\sigma_\theta}^2 - 2\rho\sigma_S\sigma_\theta} - \frac{2v_f\theta}{r_f - \mu - \rho\sigma_S\sigma_\theta} + \frac{{v_f}^2}{r_f}\right) \tag{12}$$

Finally, we have another boundary at the upper level of exchange rate, $S \uparrow \infty$. However, it is very hard to quantify this condition with any economic intuition. To mitigate this concern, we heuristically assume that the second-order impact is negligible, and that the profit flow is of first order importance to S_{t}. The condition can then be written as

$$\frac{\partial^2 V_F}{\partial S^2}(S \uparrow \infty) = 0 \tag{13}$$

The detailed implementation is discussed in Appendix C.

Applying dynamic programming, we evaluate the exporter value, V_{X}, as

$$V_X(\theta, R) = \{\pi_X(\theta_t, S_t)\Delta t + E[V_X(\theta, S)\Delta t]; V_F(\theta, S) - I\} \tag{14}$$

This equation indicates that the exporter value will be equal to the larger of (1) its own continuation value in the next instant and (2) its net present value upon foreign investment. Note that here we assume that the foreign investment expenditure I is sourced from home country. In the model extension section, we provide an example where the expenditure is sourced locally. In that case, the last item in the parentheses will be $V_F(\theta, S) - SI$. Following standard protocol, we can rewrite it as following PDEs

$$\frac{1}{2}{\sigma_S}^2 S^2 \frac{\partial^2 V_X}{\partial S^2} + \frac{1}{2}{\sigma_\theta}^2 \theta^2 \frac{\partial^2 V_X}{\partial \theta^2} + \rho\sigma_S\sigma_\theta S\theta \frac{\partial^2 V_X}{\partial S \partial \theta} + (r_h - r_f) S \frac{\partial V_X}{\partial S} + \mu\theta \frac{\partial V_X}{\partial \theta} + \pi_X = r_h V_X \tag{15}$$

To solve the equation, we need a series of boundary conditions:

First, similar to the valuation of foreign affiliates, when foreign demand shock is extremely low, the firm value approaches zero, irrespective of the exchange rate.

$$V_X(\theta \downarrow 0, S) = 0 \tag{16}$$

Second, when the exchange rate is extremely low, the firm will have no incentive to serve the foreign market because the demand bar for production will amount to infinity. This is intuitive since a strong home currency leads to weaker exports.

$$V_X(\theta, S \downarrow 0) = 0 \tag{17}$$

Finally, it is noted that the boundaries at upper levels of θ and S do not have clear economic intuition to characterize. We simply assume the third-order impact is negligible since the profit function is up to the magnitude of second-order. The detailed implementation of the conditions in the finite-difference scheme is discussed in Appendix B.

$$\frac{\partial^3 V_X}{\partial S^3}(S \uparrow \infty) \to 0 \tag{18}$$

$$\frac{\partial^3 V_X}{\partial \theta^3}(\theta \uparrow \infty) \to 0 \tag{19}$$

In the next part, we discuss the solution to the differential equation system.

4.1. Numerical Solution

We calibrate the parameters input to a mix of the ones being used extensively in past literature and from empirical data. For some of structural parameters, where the values are either new to the model or missing in literature, we simply take the best estimates.

4.1.1. Economic parameters

The risk-free rate of return of home country r_h: Jeanneret (2016) calibrated a dynamic model of sovereign debt and obtained a risk-free rate for emerging countries as 4.46%, in line with the average 10-year U.S. treasury rate; the 10-year yield of German government bond is 3.52%. Moreover, Jeanneret (2016) adopts a rate of 3.5% in a multinational investment model in a similar real options framework. Given the discrepancy in the model settings we adopt an average of r_h = 4% without loss of generality. The value loss of iceberg type exporting cost is set to τ = 0.7, consistent with Jeanneret (2016) and Fillat and Garetto (2015). We do not have an accurate measure of this cost and most literature treat it arbitrarily since it does not alter the main implication only the magnitude.

For the standard deviation of demand shock embedded in the Brownian motion σ_θ, Fillat and Garetto (2015) set the value as small as 0.022 to match the U.S. aggregate consumption uncertainty. However, since the direct measure of the standard deviation of real GDP is quite small to generate firm dynamics, Garetto et al. (2018) employ a new method. They calibrate a variety of countries and show that the standard deviation of real GDP ranges from 0.116 (U.S.) to 0.144 (Ireland). To accommodate those differences, we take a moderate level of 0.1. The standard deviation of foreign exchange rate, σ_S is set to 0.1, which is set close to the country average in Jeanneret (2016, online Appendix). In his sample, most developed countries are in the range of 4–7%. We take a relatively high exchange rate volatility to highlight the qualitative property. A lower value will not alter the results.

For the expected growth of demand shock μ: note that this value is constrained by a caveat of the dynamic model under risk-neutral measure. This is because the risk-adjusted discount rate should be positive or it leads to asset bubbles. A simple rationale can be found in Equation (12). To ensure the denominator is positive, we have to require $2r_h > r_f > 2\mu + {\sigma_\theta}^2 + 2\rho\sigma_S\sigma_\theta$, which means that the expected consumption growth should not be too large. For instance, given the value of other parameters, μ has to be less than 0.015. To facilitate computation, we set μ = 0.

However, it is still a reasonable level within [0, 1.5%] as it falls in this range in most dynamic corporate finance literature.

The sunk cost and the production cost of international investment are structural parameters, which lack empirical support. We simply set $I = 1$. Theoretically, it should not impact our results because it is a normalized value. The production cost at home and foreign countries are set to 0.5, i.e., $\nu_h = \nu_f = 0.5$. The value of production cost varies greatly in prior literature but such variation will mostly not impact our main results. Table 1 summarizes all parameters the model uses.

4.2. Results Discussion

In this section, we discuss solutions to our benchmark model. We will focus on the impact of exchange rate movement, exchange rate volatilities and foreign demand volatilities on foreign investment decision. We begin with the general shape of the investment threshold decision as a function of both demand and exchange rate movement. In Figure 1, the solid line represents the investment boundary, which can be characterized as a function $\theta(S)$, i.e., demand level varying with the exchange rate. It captures the lowest demand bar (as a function of exchange rate) across which the exporter shift production to foreign affiliates. Given the firm has not yet undertaken FDI, the dotted line represents the demand threshold, which varies with the exchange rate. The level of exchange rate separates the two operating stages, e.g., suspension and production. The figure clearly shows that exporters will conduct overseas investment if the foreign market demand is sufficiently large, irrespective of the exchange rate. This is consistent with our intuition.

For a more detailed understanding of the nonlinearity of the investment boundary, we study four different types of firms, labeled R1 $\sim$ R4. Firms in region R1 and R2, where the demand level is below $\sim$1.3, the exporter will never engage in foreign investment unless the foreign demand increases. It will only change between the suspension and production statuses as the exchange rate varies. In this case, the exchange rate will have no impact on the international investments because demand is too weak. It can also be observed that a weaker home currency will have a positive impact on the exporter so that the production status will be expected, while strong home currency will more likely leave the firm at idling status.

Notice that the foreign revenue will be converted to home currency while sunk cost is directly priced with home currency as we assumed. It means that the foreign revenue will fluctuate with exchange rate while sunk cost will

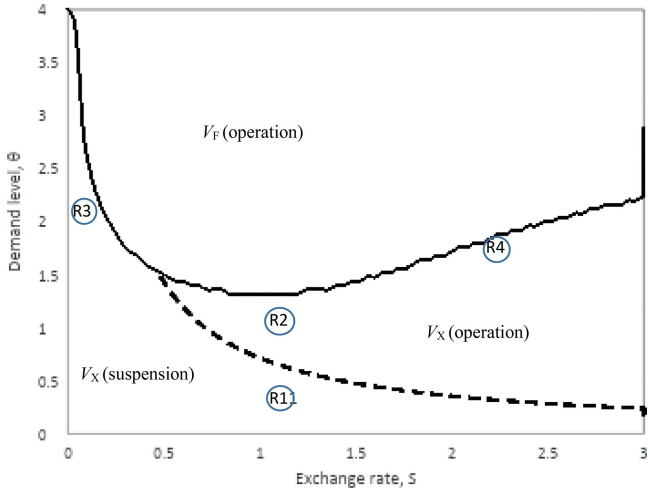


Figure 1: FDI threshold with demand and exchange rate variation. The solid line represents trigger of international investment as a function of (S_t, θ_t). The dotted line represents suspension/operation switching trigger for exporter-type firm, note that the switching threshold is also a function of both demand and exchange (S_t, θ_t). The suspension area lies to the left of switching trigger while operation area lies in the right side. The area above investment threshold belongs to foreign production. The parameters are: correlation between volatilities of demand and exchange $\rho = 0.4$, volatilities of demand and exchange rate are, respectively, $\sigma_\theta = 0.1$ and $\sigma_s = 0.1$, the expected growth rate of demand $\mu = 0$. The discount rate at home country and foreign country are $r_h = 0.04$ and $r_f = 0.07$, respectively. The FDI cost is $I = 1$ in terms of home currency.

not. In this case, if home currency is very expensive, then it will hurt the NPV of foreign investment. NPV = $SV_F - I - V_X$, where S is the home currency value of foreign currency and I is the investment cost directly measured in home currency. For example, for the firm type R3, the propensity to undertake FDI will increase as the home currency becomes weaker. This result might seem to be in conflict with conventional wisdom since weaker currency seems to promote exporting rather than FDI. This happens because firms like R3 have no cash flow-based assets, only real options to either go abroad or produce at home. As the home currency depreciates, the value of "investment option" to engage in the foreign investment dominates the "restart option" to export at full capacity. However, when the home currency is too weak, the firm is better off producing and exporting, since the benefits from FDI diminishes, making the investment curve convex.

Similarly, when the home currency is relatively weak and the demand is mildly high, the firm will be an exporter with production status, which is the firm R4. In this case, the firm is more likely to engage in international investment as the home currency become stronger (imagine that S_t moves left), consistent with conventional wisdom.

To better visualize the exporter value as a function of market demand θ and exchange rate S, we plot a three-dimensional figure as well as a contour map in Figure 2. It can be clearly seen that as both S and θ become smaller the firm value approaches zero. This is not surprising since the home currency is very strong and foreign demand is very low. Moving far from (0,0) away to the northeast direction, i.e., as demand becomes very strong and exchange rate very weak, the firm will immediately start foreign investment since it has been deep-in-the-money. It explains

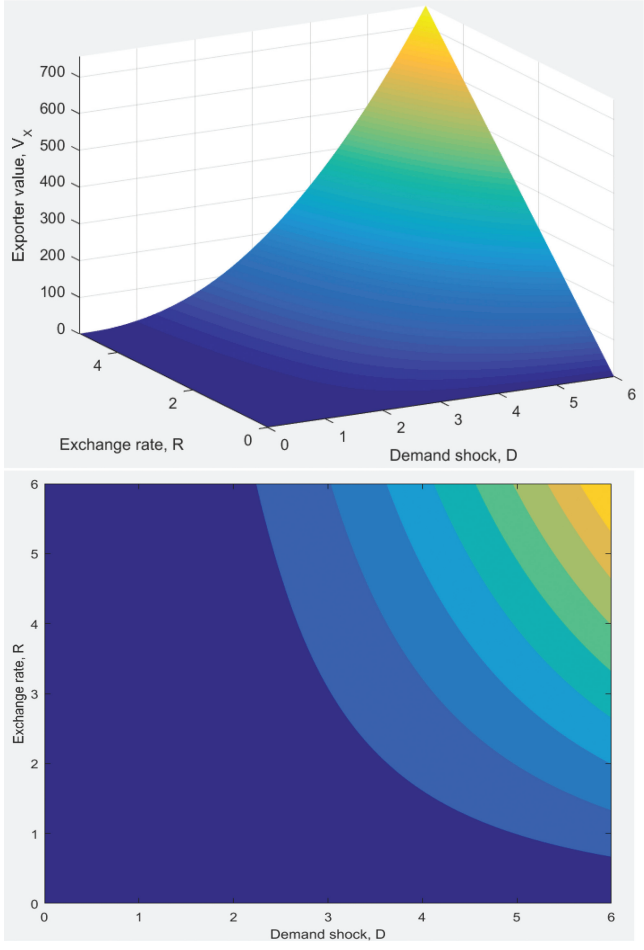


Figure 2: Firm value for exporter. The left panel plots market value of exporter status as a function of (S_t, θ_t) and the right panel plots from a two-dimensional contour view. The parameters are: correlation between volatilities of demand and exchange rate $\rho = 0.4$, volatilities of demand and exchange rate are, respectively, $\sigma_\theta = 0.1$ and $\sigma_s = 0.1$, the expected growth rate of demand $\mu = 0$. The discount rate at home country and foreign country are $r_h = 0.04$ and $r_f = 0.07$, respectively. The FDI cost is $I = 1$ in terms of home currency.

why a positive correlation can make demand or exchange rate uncertainties benefit investment decisions. We will leave the discussion of correlation to the next part.

In this section, we discuss the impact of demand and exchange rate volatilities (σ_θ and σ_S) and their correlations (ρ) on the trigger values. Figure 3 graphs the impact of demand volatilities on the investment trigger. When the correlation is positive, it can be observed that higher demand volatility may accelerate investment timing for a certain range of exchange S (for example, when S is relatively large). Notice that a larger V_F area represents a higher possibility to hit the foreign investment decision.

The result may look counterintuitive since it is conventional wisdom that higher demand volatility should deter investment timing. The intuition behind this is that positive correlation will make the post-investment foreign assets more valuable, as in Equation (12). This effect will be more prominent given the positive standard deviations of either demand or exchange rate. We call this the "revenue channel." Additionally, higher demand uncertainty will also deter investment due to the "real option" effect. When the revenue effect dominants the real option effect, the MNE will accelerate the investment. When the correlation is negative, demand uncertainty always delays entry, since the foreign asset value will be discounted more heavily. However, it needs to be noted that the investment trigger is defined by two-dimensions. Therefore, a simulation exercise (Section 5) would help to visualize the impact.

More interestingly, when the exporter is at suspension (exchange rate $S_t \sim <1$), higher demand uncertainty (σ_θ = 0.1) will delay foreign investment, irrespective of the sign of correlation. This is expected because at this stage the firm does not have any revenues but has an investment option (to enter foreign market) or a restart option (to restore production for export). In this case, the real option effect will play a more important role.

Figure 4 graphs the impact of exchange rate uncertainty on FDI decision. In general, we can observe that the results are similar to that of demand uncertainty. When the correlation is positive, the impact of exchange rate uncertainty is ambiguous, when the correlation is negative, higher volatility always delays foreign investments. The reasoning is similar to that of demand uncertainty so we will not repeat here. It will be valuable to compare our results with Goldberg and Kolstad (1995). In their paper, under positive correlation, the firm always allocates 100% capacity overseas since that lowers the profits variance. Although we apply the same logic, their model cannot predict the investment decision. Under negative correlation, they (proposition 3, page 864) show that exchange rate volatility will have a positive impact on the share of foreign production, however, their model cannot draw any conclusion on the absolute FDI level. In this sense, we are looking for different aspects of FDI and are not in conflict with their results.

In the next part, we introduce two important extensions that are relevant to the real economy. We repeat similar analysis as before.

4.3. Extension 1: Alternative Sunk Cost Source

In this section, we present solutions to the model with alternative sunk cost of foreign investment, that is, the investment cost is composed of materials and technologies bought locally. Consequently, the decision of firm owners can be rewritten in a dynamic programming fashion

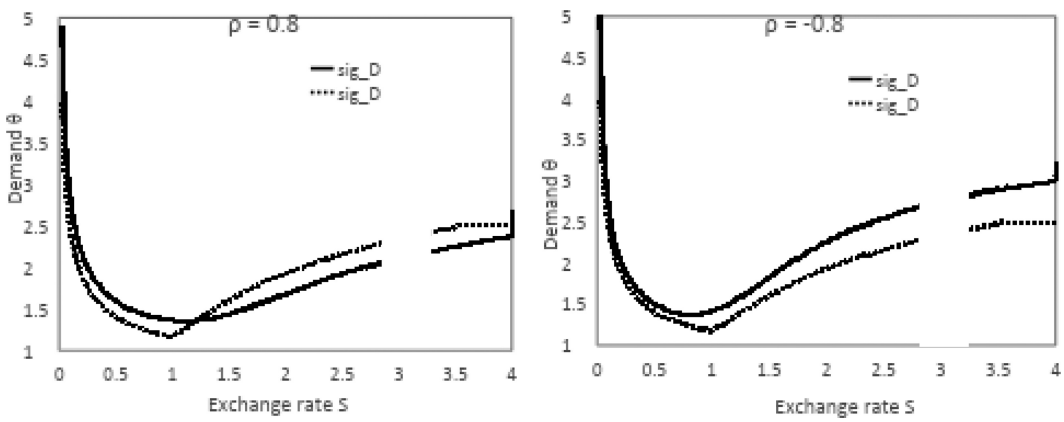


Figure 3: The impact of demand uncertainty on FDI. The figure depicts the impact of demand uncertainty on FDI for low demand volatility (dotted line, $\sigma_\theta = 0$) and high demand volatility (solid line, $\sigma_\theta = 0.1$). The exchange rate volatility is set as $\sigma_s = 0.2$, the expected growth rate of demand $\mu = 0$. The discount rate at home country and foreign country are $r_h = 0.04$ and $r_f = 0.07$, respectively. The FDI cost is $I = 1$ in terms of home currency. The demand elasticity $\gamma = 1$.

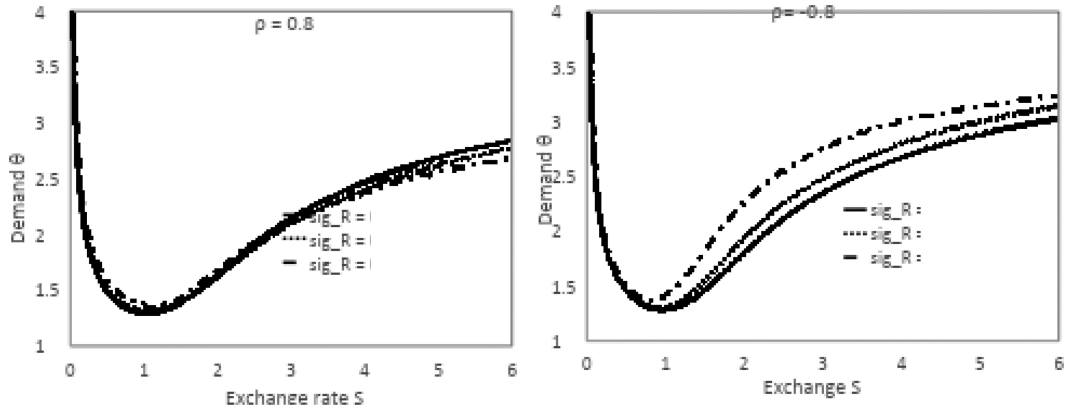


Figure 4: The impact of currency volatility on FDI. The figure depicts the impact of currency volatilities on FDI for positive correlation (left) and negative correlation (right). The solid lines correspond to the case of low currency volatility while the dotted line represents high volatility. The parameters are: volatilities of demand is $\sigma_\theta = 0.1$, the expected growth rate of demand $\mu = 0$. The discount rate at home country and foreign country are $r_h = 0.04$ and $r_f = 0.07$, respectively. The FDI cost is $I = 1$ in terms of home currency.

$$V_X(\theta, R) = \{\pi_X(\theta_t, S_t)\Delta t + E[V_X(\theta, S)\Delta t]; V_F(\theta, S) - SI\} \tag{20}$$

It is similar to Equation (14) and the only difference lies in the last item *SI*.

Figure 5 demonstrates the investment threshold as the function of demand, depending on the exchange rate levels. In contrast to the benchmark case, the foreign investment threshold presents monotonic relation with exchange rate. Foreign investment is postponed as exchange rate becomes weaker. At the suspension region, the demand trigger for foreign investment is independent of exchange rate. This is intuitive since now the sunk cost is priced in local currency and needs to be converted to home currency. In empirics, this figure suggests that appreciating home currency always has a positive impact on investment in the foreign countries and there will not be any nonlinearity.

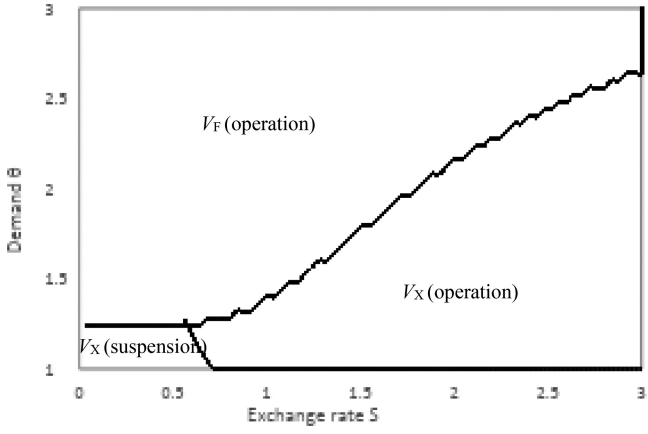


Figure 5: FDI threshold with demand and exchange rate variation. The solid line represents trigger of international investment as a function of (S_t, θ_t). The dotted line represents suspension/operation switching trigger for an exporter-type firm. The suspension area lies to the left of switching trigger while operation area lies in the right side. The area above investment threshold belongs to foreign production. The parameters are: correlation between volatilities of demand and exchange $\rho = 0.4$, volatilities of demand and exchange rate are, respectively, $\sigma_\theta = 0.1$ and $\sigma_s = 0.1$, the expected growth rate of demand $\mu = 0$. The discount rate at home country and foreign country are $r_h = 0.04$ and $r_f = 0.07$, respectively. The FDI cost is $I = 1$ in terms of home currency.

Figures 6 and 7 plot the impact of demand volatility and exchange rate volatility, respectively. It can be seen that the results pattern hold for an alternative sunk cost. Since the mechanisms are the same as the benchmark case, we will not repeat it here.

4.4. Extension 2: Asymmetric Capacity Limit

In this section, we present a solution to the case with asymmetric capacity at home and foreign. It is probably the most common scenario around the world. Without loss of generality, we assume the domestic production limit is Q_X and foreign is Q_F and $Q_F > Q_X$. This extension will be more appropriate for market seekers such as FDI outflow from other countries to U.S. We can rewrite the profit function for the exporter as follows.

$$\pi_X = \begin{cases} \left(\tau S\theta - \tau\gamma S Q_X - v_h\right)Q_X & \theta > 2\gamma Q_X + \dfrac{v_h}{\tau S} \\ \dfrac{(\tau S\theta - v_h)^2}{4\tau S\gamma} & \dfrac{v_h}{\tau S} < \theta < 2\gamma Q_X + \dfrac{v_h}{\tau S} \\ 0 & \dfrac{v_h}{\tau S} > \theta > 0 \end{cases} \tag{21}$$

And similarly for post-FDI profit

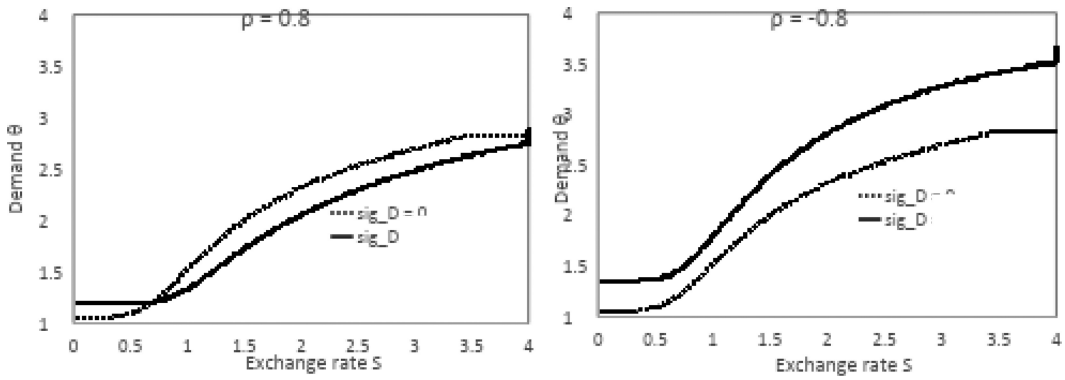


Figure 6: The impact of demand uncertainty on FDI. The figure depicts the impact of demand uncertainty on FDI for low demand volatility (dotted line, $\sigma_\theta = 0$) and high demand volatility (solid line, $\sigma_\theta = 0.1$). The exchange rate volatility is set as $\sigma_s = 0.2$, the expected growth rate of demand $\mu = 0$. The discount rate at home country and foreign country are $r_h = 0.04$ and $r_f = 0.07$, respectively. The FDI cost is $I = 1$ in terms of home currency. The demand elasticity $\gamma = 1$. In particular, the FDI expenditure will be sourced locally.

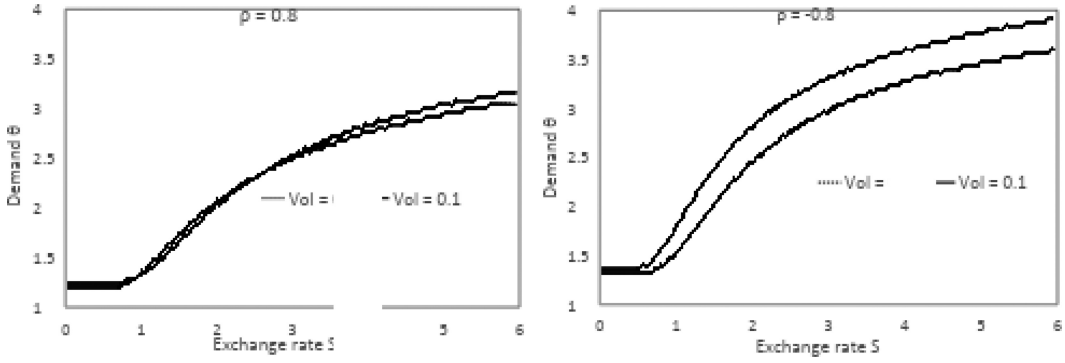


Figure 7: The impact of currency volatility on FDI. The figure depicts the joint impact of correlation and currency volatilities on FDI for positive correlation (left) and negative correlation (right). The solid lines correspond to the case of low currency volatility while the dotted line represents high volatility. The parameters are: volatilities of demand is $\sigma_\theta = 0.1$, the expected growth rate of demand $\mu = 0$. The discount rate at home country and foreign country are $r_h = 0.04$ and $r_f = 0.07$, respectively. The FDI cost is $I = 1$ in terms of home currency.

$$\pi_F = \begin{cases} S\left(\theta - \gamma \underline{Q}_F - v_f\right)\underline{Q}_F & \theta > 2\gamma \underline{Q}_F + v_f \\ \dfrac{S(\theta - v_f)^2}{4\gamma} & v_f < \theta < 2\gamma \underline{Q}_F + v_f \\ 0 & v_f > \theta > 0 \end{cases} \tag{22}$$

We define the capacity limit trigger at home country as $\underline{\theta}_h = 2\gamma \underline{Q}_X + \frac{v_h}{\tau S}$ and for the foreign country as $\underline{\theta}_f = 2\gamma \underline{Q}_F + v_f$. It can be observed that the order of the two depends on the exchange rate dynamics S_t as well as the relative magnitude of the capacity limit, and variable costs. The foreign capacity limit, however, is independent of both exchange rate and exporter cost.

Figure 8 presents the investment decisions as a function of demand and exchange rate. It can be observed that the general shape is similar to that of Figure 1, except for an additional exporter status, i.e., underutilized. There are two interesting observations here: (1) the chances of being in the underutilized status is seemingly higher than the other two; (2) the exporter will always begin to conduct foreign investment when the foreign capacity is at underutilized status. This is intuitive because investment in full foreign capacity may require higher demand threshold and appropriate exchange rate, causing longer waiting time. Meanwhile, investment at underutilized capacity will maximize firm value *ex ante*.

Figure 9 graphs a 2-D and 3-D view of exporter value. The general shape is similar to Figure 2. Larger foreign demand and exchange rate (weaker home currency) will make the home-based firm more valuable. However, it is noteworthy that the highest value occurs not necessarily at the highest demand level (D $\sim$ = 5).

Figure 10 and Figure 11 plot the impact of demand volatility and exchange rate volatility on the FDI decision, respectively. It can be seen that the results still hold for alternative capacity settings.

5. Simulation Exercise

So far, our model has delivered numerical results for foreign investment decision under both foreign demand and exchange rate uncertainties. However, the model only presents time-invariant (stationary) decisions while the real firms operate in a dynamic world. Second, the timing decisions are hard to interpret with the empirical data, especially given our two-dimensional background. We need to convert the trigger decision to an intensive margin. In the next paragraphs, we first attempt to address some parameter generation process then we proceed to quantitative analysis.

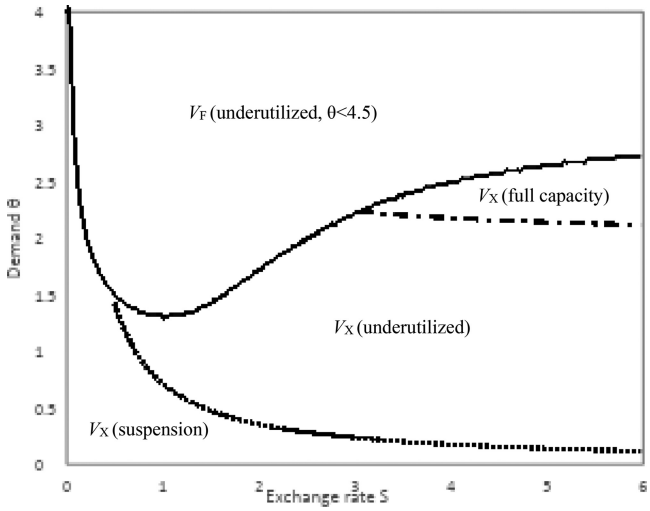


Figure 8: FDI threshold for limited capacity. The solid line represents trigger of international investment as a function of (S_t, θ_t). The dotted line represents suspension/underutilization switching trigger for exporter-type firm, the dot dashed line represents boundary between underutilization and full capacity production. The area above investment threshold belongs to foreign production below full capacity. The parameters are: correlation between volatilities of demand and exchange ρ = 0.4, volatilities of demand and exchange rate are, respectively, σ_θ = 0.1 and σ_s = 0.1, the expected growth rate of demand μ = 0. The discount rate at home country and foreign country are r_h = 0.04 and r_f = 0.07, respectively. The FDI cost is I = 1 in terms of home currency. The capacity limit of home country is Q_h = 1 while that for foreign country is Q_f = 2.

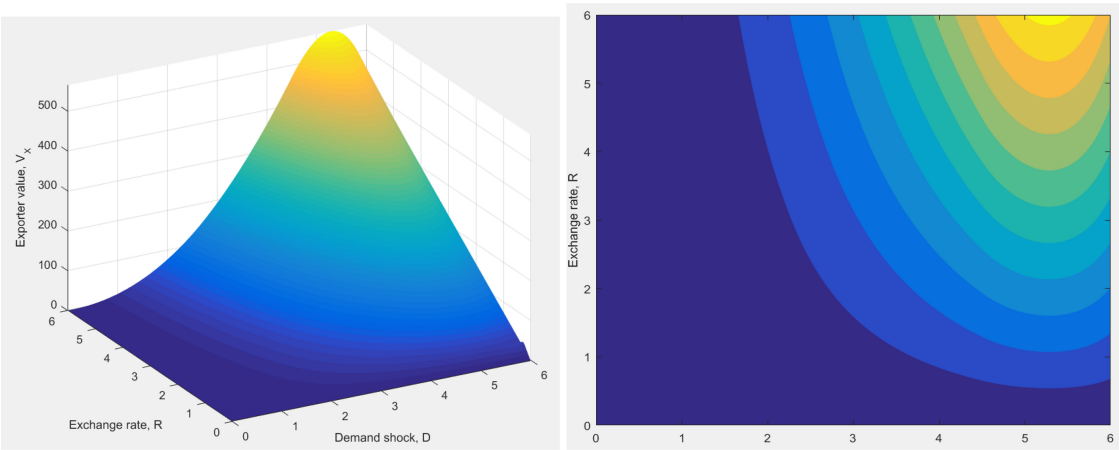


Figure 9: Exporter value. The left panel plots market value of exporter status as a function of (S_t, θ_t) and the right panel plots from a 2-D contour view. The parameters are: correlation between volatilities of demand and exchange $\rho = 0.4$, volatilities of demand and exchange rate are, respectively, $\sigma_\theta = 0.1$ and $\sigma_s = 0.1$, the expected growth rate of demand $\mu = 0$. The discount rate at home country and foreign country are $r_h = 0.04$ and $r_f = 0.07$, respectively. The FDI cost is $I = 1$ in terms of home currency. The capacity limit of home country is $Q_h = 1$ while that for foreign country is $Q_f = 2$.

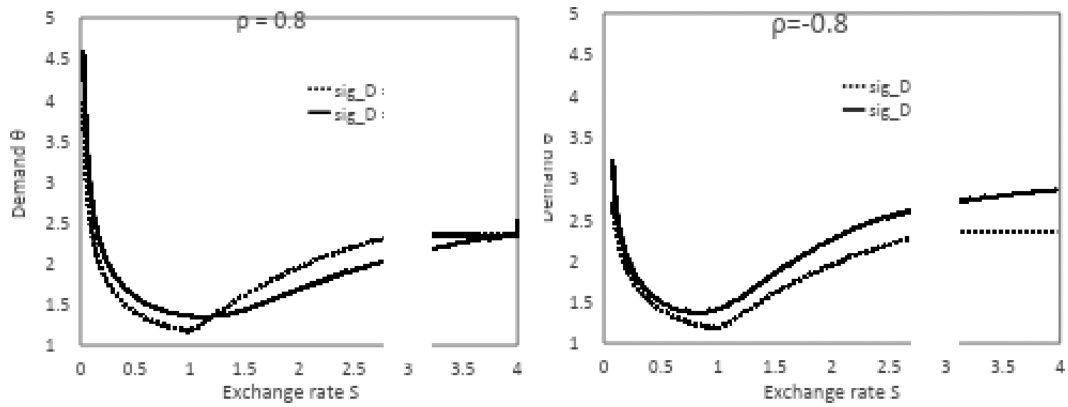


Figure 10: The impact of demand volatility on FDI.

5.1. Price Dynamics and Parameters Setting

The simulated economy is also composed of one home country and one foreign country. Consistent with the model, we simplify the economy so that there is only one firm in the home country and it will either export or make direct investment in the foreign market. In this regard, our simulation will neglect firm heterogeneity (e.g., firm level parameters). The economy is simulated in a quarterly frequency $\Delta t = 1/4$ over 50 years.

We first express the dynamics of foreign demand and exchange rate in an explicit form, i.e., obtain the solution to stochastic differential equations (SDEs) of Equation (2) and Equation (3), given that there is correlation between the two processes. For instance, in the special case of zero correlation between the demand and exchange rate shock, $\rho = 0$. We can obtain the exact solution to the equation since they are log-normal distribution, in particular, the price flow can be discretized as follows:

$$\theta_t = \theta_0 \left[exp\left(\mu - \frac{1}{2}\sigma_\theta^2 \right) t + \sigma_\theta \sqrt{t} \sum_{j=0}^{t} Z_j \right]$$

and the exchange rate can be discretized as follows:

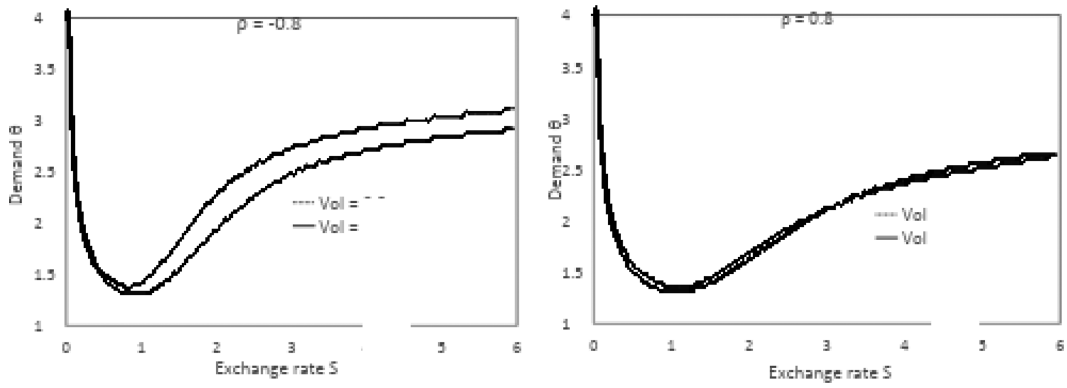


Figure 11: The impact of currency volatility on FDI. The figure depicts the joint impact of correlation and currency volatilities on FDI for positive correlation (left) and negative correlation (right). The solid lines correspond to the case of low currency volatility while the dotted line represents high volatility. The parameters are: correlation between volatilities of demand and exchange $\rho = 0.4$, volatilities of demand and exchange rate are, respectively, $\sigma_\theta = 0.1$ and $\sigma_s = 0.1$, the expected growth rate of demand $\mu = 0$. The discount rate at home country and foreign country are $r_h = 0.04$ and $r_f = 0.07$, respectively. The FDI cost is $I = 1$ in terms of home currency. The capacity limit of home country is $Q_h = 1$ while that for foreign country is $Q_f = 2$.

$$S_t = S_0\left[exp\left(r_h - r_f - \frac{1}{2}{\sigma_\theta}^2\right)t + \sigma_S\sqrt{t}\sum_{j=0}^{t} W_j\right]$$

However, whenever $\rho \neq 0$, there is no closed form solution because the two Weiner processes share a covariance matrix, $Corr(dZ, dW) = \rho$ therefore we need to generate

$$d\Omega = \left(\frac{dZ}{dW}\right) \sim N(0, \Lambda)$$

and the covariance matrix is

$$\Lambda = \begin{pmatrix} \Delta t & \rho\Delta t \\ \rho\Delta t & \Delta t \end{pmatrix}$$

Since there are no closed form solutions for the discretization, we opt for a numerical solution to the SDEs, in particular, under Ito-Tylor expansion and keeping only order-1 items we obtain *Euler* scheme

$$\theta_t = \theta_{t-1} + \mu\theta_{t-1}\Delta t + \sigma_\theta\Delta\tilde{Z}$$
$$S_t = S_{t-1} + \mu S_{t-1}\Delta t + \sigma_S\Delta\tilde{W}$$

The new unknown Weiner process $\Delta\tilde{Z}$ and $\Delta\tilde{W}$ have no analytic expression but they have to obey the above covariance matrix. In order to do this, we calculate L knowing that $\Lambda = LL^T$ and simulate $D\psi \sim N(0, I^2)$ to obtain $d\Omega = Ld\Psi$. We use the quadratic resampling method to generate $D\psi$, which will affect $d\Omega$ by the way we construct. Let us define $E\Psi$ and $\Lambda\Psi$ as the theoretical mean and covariance matrix of $D\psi$, respectively, namely,

$$E\Psi = (0\ 0\ ) \text{ and } \Lambda\Psi = (1\ 0\ 0\ 1\ )$$

In our model, the investment timing decision is captured by two underlying variables, that is, *trigger* = *trigger* (S, θ). We wish to transit it to investment intensity (e.g., probability measure) in a simulated panel. In particular, at $t = 0$ we assume the MNE is born in the home country. The combination (θ_0, S_0) is too low to allow this firm to engage in overseas investment. However, it will keep the firm as an exporter at production status. As the two variables evolve with time, the firm may serve the foreign market through either exporting or foreign direct investment, or suspension as an exporter. For example, according to the solution in Figure 1, the contour map shows that there are three regions neighboring the threshold line. The firm status controlled by (θ_t, S_t) in the simulated economy will be mapped to corresponding regions contained in the theoretical solution. Whenever the FDI event is triggered, the home-based MNE will engage in the investment in the foreign market and meanwhile, the existing MNE is "retired" and replaced by a newborn MNE in the home country. The new MNE is endowed with the same operating parameters.

There is another caveat in simulating the two-dimensional stochastic control problem. The PDE model only delivers a limited range of solutions with pre-determined stepwise in the finite-difference scheme for θ(S) thus its calculation is time-consuming. In this case, for any simulated exchange rate S_t outside the model range we interpolate these new points to fit the nonlinear curve. This method not only allows solutions for any possible simulated value of exchange rate S_t but also makes the solutions more continuous and thus smooths out the curve.

To conduct quantitative analysis of simulated foreign investments, we focus on simple graphic presentations instead of regression analysis. This is because regression analysis demands matching moments between simulated and empirical data. In our case, since we focus on a simple two-country economy, producing data moments in time series is more important than in the cross-section. Unfortunately, due to the curse of model dimensions, we have to limit the model capacity up to stationary (time-invariant) solutions and we exclude time dependent returns and volatilities. The lack of time-series accuracy lowers the validity of regression results. Second, in a random simulation, a relatively large batch of samples is required to generate a robust result. However, this protocol may even weaken the impact of economic uncertainties especially when we average the sample to get expected investment intensity, losing our main purpose. As a consequence, our simulation exercise is not intent to perform empirical specification, instead it tries to validate intuitions reflected from previous bi-dimensional figures. In the following, we introduce a series of simulated results.

5.2. Exhibition 1: The Impact of Exchange Rate (Returns) on International Investment Decisions

In Figure 12(a), we present the simulated diffusive process of exchange rate S_t (solid line) and foreign demand θ_t (dashed line). They are obtained by averaging over 50 simulated samples following previous construction protocol.

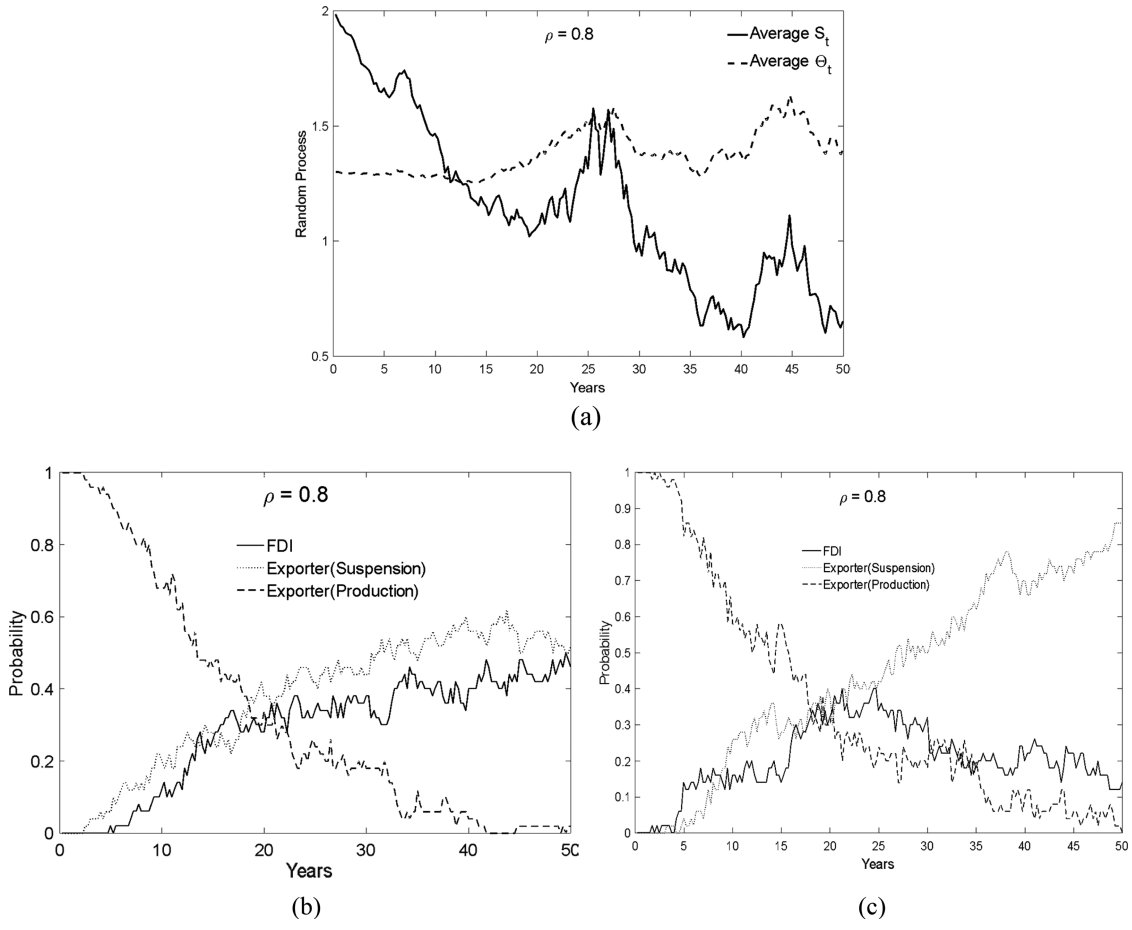


Figure 12: (a) A sample path of (S_t, θ_t) with the correlation of Winer process $\rho = 0.8$. The parameters are selected as $N = 100$, $\theta_0 = 1.3$, $S_0 = 2$, $\rho = 0.8$, $\sigma_S = 0.2$, $\sigma_\theta = 0.1$. The value of each randomness is calculated by averaging the 100 paths. (b) Depicts the production status for a MNE whose foreign investment cost is sourced locally. (c) Depicts the production status for a MNE whose foreign investment cost is sourced from home country.

We choose small samples to average to avoid eliminating volatilities as mentioned before. The starting point is set at $\theta_0 = 1.3$ and $S_0 = 2$ (R4 firm type in the Figure 1), so the exporter is at production status and has not engaged in the foreign investment. The correlation of shocks is $\rho = 0.8$. It can be observed that the volatilities of the two process comove together, consistent with the positive correlation. The exchange rate decreases amid ups and downs due to the negative drift ($4\% - 7\% = -3\%$). The demand process does not deviate much since the drift is set to zero.

Figure 12(b) and 12(c) present an MNE's production status as the exchange rate and foreign demand evolve with time. The difference between them lies in the assumption of cost structure of foreign investment expenditure. Panel (b) is for the case that the investment expenditure is expensed at the local currency. The probability associated with each production/investment status is computed by counting the frequency of each happenings scaled by total simulated paths. Recall that we assume the MNE is initially an exporter with active production. Consistent with the intuition from Figure 5, the chances of both engaging in FDI and suspending current home production increase as the home currency appreciates (as shown in Figure 12(a)). Panel (c) is for the case that the expenditure is sourced from home country, i.e., our main model. Recall in the Figure 1, the numerical solution shows that if the firm is initially at an active exporter status (say R4 type), then appreciating home currency will initially increase and then decrease the possibility of FDI while the status of suspending production become more likely. The simulation result in Panel (c) confirms the general pattern.

5.3. Exhibition 2: The Impact of Demand Volatility on International Investment

Our next focus is the impact of foreign demand uncertainty. Since as we have seen previously this result is robust for all types of scenarios, we just select the case in the Extension 1 as illustration. Our numerical solution illustrates that positive correlation between the demand and currency dynamics may produce positive impact of the demand uncertainty on foreign investment, the result is confirmed in Figure 13(left panel). Higher demand uncertainty causes accelerated FDI entry compared to the case with lower demand uncertainty, but this trend is reversed given sufficiently long period of time. This happens because of the relative dominance of "real option effect" and "revenue effect," as discussed before. Again, when the correlation is negative, high uncertainty always has negative impact on foreign investments because only "real option effect" remains.

5.4. Exhibition 3: The Impact of Exchange Rate Volatility on International Investment

Figure 14 plots the FDI decision under the impact of exchange rate variation. Similar to demand uncertainty we find that positive correlation leads to some positive or insignificant impact of high exchange rate uncertainty (left panel) while negative correlation usually entails negative impact of high currency volatility (right panel), particularly if we neglect the first 20 years. However, notice that the results are not so pronounced compared to the demand factor. This result is interestingly consistent with other empirical findings where the exchange rate volatility generally has insignificant impact on the FDI decision (see Choi and Jiang (2009) and Nguyen et al. (2018)). Such a coincidence can be attribute to the dominant impact of foreign demand relative to the that of exchange rate in our model framework.

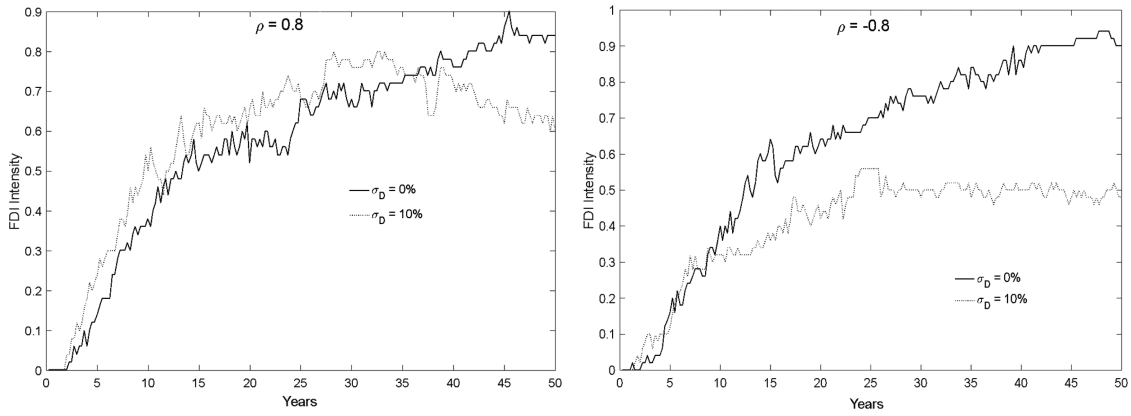


Figure 13: Depicts the impact of demand uncertainty on the FDI intensity for *positive correlation* (left) and *negative correlation* (right). The parameters are selected as $N = 100$, $\theta_0 = 1.5$, $S_0 = 3$, $\sigma = 0.2$. The solid line depicts the case of low demand volatility (0%) and the dotted line is for high demand volatility (10%).

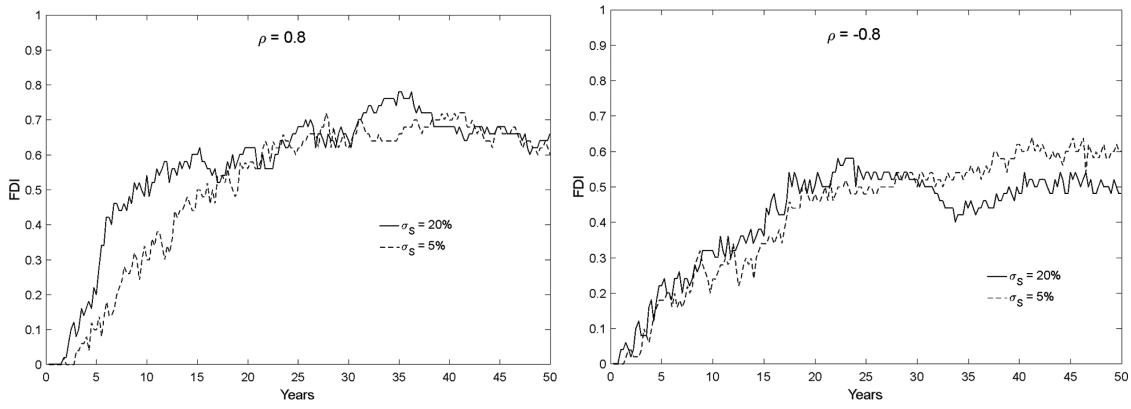


Figure 14: Depicts the impact of exchange rate uncertainty on the FDI intensity for positive correlation (left) and negative correlation (right). The parameters are selected as $N = 100$, $\theta_0 = 1.5$, $S_0 = 3$, $\sigma_D = 0.1$. The solid line depicts the case of high exchange rate volatility (20%) and the dotted line is for low exchange rate volatility (5%).

Additionally, Goldberg and Kolstad (1995) theoretically show that the positive correlation should increase the overseas production share. In our simulation exercise, we observe that the role of correlation has to be interpreted with the help of other factors such as demand or exchange rate volatilities. For instance, in the Figure 13, the positive correlation has nearly zero impact for the case of low demand volatility whereas it elevates FDI in the case of high demand volatility. In Figure 14, the positive correlation seemingly improves FDI for both high and low exchange rate volatilities. Therefore, the effect of correlation requires large resampling batches from simulations to construct statistical inference. However, this is beyond the scope of this paper, given our bidimensional setting.

6. Conclusion

In this paper we build a dynamic continuous-time model to characterize the (horizontal) foreign investment decision in the presence of both stochastic exchange rate and stochastic foreign demand. We show that exchange rate depreciation could have either positive, negative, or insignificant impact on FDI, depending on the source of the irreversible cost. If the investment expenditures are sourced in the destination country, then the appreciation of home currency always has positive impact on FDI; if the investment expenditures are sourced from home country, then there is a nonlinear relation between exchange rate return and foreign investment. Given in the real world, both cases exist, the aggregate results could vary.

We also show that both exchange rate and foreign demand volatility also could have mixed impact on FDI, there are two forces behind this phenomenon. First, the volatility itself can introduce real option effect that postpones foreign investments; Second, the revenue effect that makes the foreign affiliates more valuable, accelerates foreign investments. In empirical sense, the exact direction of uncertainty influence depends on the relative dominance between the two forces.

Our model contributes to the literature and encourages more insightful research in the future. In empirics, a larger dataset of firm-level foreign investments (for any trade partners) will contribute to our understanding on the ambiguous results so far. In theory, a dynamic model for bilateral economy with rich firm-level heterogeneity will lay more promising micro foundation. For instance, although exchange rate movement will be same for all firms the market demand will not necessarily be same because different industries have different specific uncertainties in their products' domain. Additionally, data from firm-level investments can provide more observations on lumpy investment spikes, which should be much closer to our real options setting instead of the data at country-level. Our paper also calls for the new direction in the application of Machine Learning to continuous-time models. Duarte et al. (2024) developed a deep policy learning algorithm for solving nonlinear high-dimensional continuous-time models in the fields of asset pricing, corporate finance and portfolio choice. In our classic finite difference method, the PDE is solved slowly due to iteration and approximation. The new deep learning algorithm could largely decrease computation and simulation time and incorporate more stochastic variables such as interest rate and inflation rate, both of which can also impact foreign exchange. Therefore, application of machine learning can embrace more economic factors and makes the results closer to the real economy.

Appendix A: Empirical Exercises for Section 3

The FDI data come from BEA and shows U.S. Direct Investment Abroad. BEA obtains these data from comprehensive mandatory surveys of U.S. multinational firms, and are available in considerable detail by country. Foreign direct investment abroad is defined by the BEA as ownership by a U.S. investor of at least 10% of a foreign business and measures the total outstanding level of U.S. direct investment abroad at yearend. Similarly, information on inward FDI into the U.S. is obtained from mandatory reporting by all U.S. business enterprises in which a foreign firm or individual owns, directly or indirectly, 10% or more of the voting securities of an incorporated U.S. business enterprise or an equivalent interest of an unincorporated U.S. business enterprise.

We use GDP growth volatility as a proxy for market demand. In previous literature, scholars have used GDP growth as demand proxy, for instance Goldberg and Kolstad (1995). However, the standard deviation of GDP growth for most countries is very trivial. We use GDP growth volatility instead. We obtain the GDP growth rate from the World Banks databank. We define the demand volatility as the five-year rolling standard deviation of GDP growth.

We include both the exchange rate return and the volatility in exchange rate returns as independent variables. Exchange rate is defined as the *number of foreign currencies per USD*. This definition remains the same whether U.S. is the home or host country. The volatility in exchange rate returns is defined as the standard deviation of exchange rate returns. The main independent and dependent variables are Winsorized for outliers at 99%. Other controls in the estimation include GDP growth rate, FDI as a percent of GDP and the level of inflation. We also include year and country controls. The data for all the controls are obtained from World Bank's World Development Indicators.

We include the correlation between the two volatilities—exchange rate returns and foreign demand growth. To obtain the correlation, we first estimate quarterly GDP growth volatilities and then obtain annual correlation between the volatility of exchange rate and the GDP growth volatility.

The table below provides the data summary of all the variables used in the estimation.

We test for stationarity in the main independent variables and the dependent variable using the Augmented Dickey Fuller test. We find the presence of unit root in the GDP growth series, and both the volatilities for some countries. To address the nonstationarity, we include first difference of the series where there is unit root.

Regression Analysis

Our estimation strategy is very similar to Goldberg and Kolstad (1995). However, we expand our study to seven countries instead of just two and study the data for a longer period of time using annual data instead of quarterly data. One caveat in our data is that we cannot identify the type of FDI—horizontal or vertical. This issue may be unavoidable since most MNCs engage in some combination of horizontal and vertical FDI. However, according to Fillat et al. (2015) most foreign sales are horizontal. We acknowledge that this caveat may cast some constraints on the explaining power of the regression.

Our reduced form equation that measures the relationship between FDI flow into the host country and the demand and exchange rate volatilities is given below.

$$\begin{aligned}\text{FDI}_{i,j,t} = {} & \beta_0 + \beta_1 \text{ER Returns}_{j,t} + \beta_2 \text{ER Volatility}_{j,t} + \beta_3 \text{GDP Growth Volatility}_{j,t} + \beta_4 \text{Correlation}_{j,t} \\ & + \gamma_j + \tau_i + \epsilon_{i,j,t}\end{aligned}$$

FDI flow is log linearized which leaves us with only the positive values of FDI flow.[7] $\text{FDI}_{i,j,t}$ shows the flow of FDI from country i to country j in time t. We include both home country controls and destination country controls. The estimation is an OLS regression run for each pair of countries. Since we do not have the FDI inflow information between the seven countries, we do

Table A.1: Summary of variables.

	Minimum	Maximum	Mean	25th Percentile	Median	75th Percentile
FDI Flow	−0.051	0.195	0.001	0.000	0.000	0.001
ER Returns	−0.294	0.223	−0.004	−0.061	−0.001	0.050
ER Volatility	0.009	0.180	0.082	0.059	0.079	0.102
GDP Growth Volatility	0.149	4.722	1.632	0.885	1.452	2.065
Correlation	−1.000	1.000	0.008	−0.850	−0.043	0.888
FDI as percent of GDP	−26.195	86.589	2.582	0.732	1.424	2.330
Inflation	−2.294	12.091	2.164	1.221	1.994	2.903
GDP growth rate	−5.619	7.259	2.366	1.567	2.518	3.678

[7]We also redo the regression with full FDI sample and the results are similar.

not use a panel estimation or a pooled regression. Again here we follow Goldberg and Kolstad (1995) and conduct pairwise estimation of the relationship. To facilitate our presentation, the summary of the estimation result with bilateral estimations is given below

Table A.2: OLS regression of log FDI Flow on exchange rate volatility and GDP growth volatility (with U.S. as the destination country).

Variables	(1) U.K.	(2) Japan	(3) Canada	(4) Germany	(5)† Netherlands	(6) Switzerland	(7) France
ER Returns	7.598*	5.353**	12.088	−4.968*	4.153	13.179**	7.689***
	(3.442)	(2.457)	(10.196)	(2.409)	(2.773)	(5.594)	(1.400)
ER Volatility	8.186	15.883	−9.642	−12.927	11.038	−16.743	−7.814
	(13.447)	(10.338)	(34.069)	(7.338)	(27.707)	(16.004)	(6.513)
GDP Growth Volatility	−0.155	−0.035	0.227	−0.769**	0.376	0.533	0.008
	(0.352)	(0.333)	(0.552)	(0.302)	(0.484)	(0.374)	(0.273)
Correlation	−0.477	−0.114	−0.435	0.560*	−0.235	0.294	−0.449*
	(0.373)	(0.397)	(0.473)	(0.288)	(0.576)	(0.503)	(0.247)
Constant	−6.175	−9.368***	−6.860**	−0.745	−7.614	−5.688*	−6.644***
	(3.509)	(2.206)	(2.548)	(3.055)	(4.615)	(2.563)	(1.183)
Observations	17	22	23	20	16	18	22
R^2	0.417	0.709	0.308	0.641	0.663	0.624	0.856

Dependent variable is logarithm of FDI Flow. The data for the estimation run from 1982 to 2018. ER Returns is the difference between the average annual exchange rate of one year from the previous year. ER Volatility is the volatility of the exchange rate estimated as the standard deviation of exchange rate returns. GDP Growth Volatility is the volatility of the GDP growth rate estimated as the standard deviation of GDP growth rate. Correlation is the correlation between GDP Growth Volatility and exchange rate returns volatility (the volatilities estimated from quarterly and monthly data, respectively). It is the correlation between country pairs—which means the correlation changes over years and over country pairs. The estimation includes FDI as a percent of GDP, one period lagged inflation, GDP, and GDP growth rate as controls. Lagged GDP volatility and exchange rate returns volatility are included where necessary to correct for unit root. † Includes lagged dependent variable to correct for unit root.
Robust standard errors in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.3: This table is similar to Table A.2 except here U.S. is the home country.

Variables	(1) U.K.	(2) Japan	(3) Canada	(4) Germany	(5) Netherlands	(6) Switzerland	(7) France
ER Returns	−2.009	3.637*	1.822	−10.437	−2.291	−8.928*	−1.143
	(7.351)	(1.645)	(21.699)	(5.860)	(3.180)	(3.804)	(2.223)
ER Volatility	8.478	14.244**	20.072	−20.011	8.697	−23.318**	2.496
	(28.699)	(4.501)	(79.174)	(14.511)	(7.787)	(8.677)	(7.450)
GDP Growth Volatility	0.067	−1.103***	0.432	0.250	−0.172	−1.812**	−0.387
	(1.222)	(0.234)	(0.830)	(0.398)	(0.536)	(0.614)	(0.599)
Correlation	−0.606	0.235	0.248	−0.431	−0.805	0.589	−0.289
	(0.936)	(0.258)	(0.508)	(0.434)	(0.732)	(0.358)	(0.367)
Constant	−10.881	−4.232*	−9.695*	−6.915**	−9.023***	−5.915*	−7.612***
	(10.576)	(1.975)	(3.904)	(2.338)	(2.431)	(2.610)	(1.294)
Observations	16	20	15	16	21	17	16
R^2	0.335	0.814	0.211	0.637	0.430	0.849	0.423

Robust standard errors in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Appendix B: Derivation of Equations (3) and (8)

Derivation of Equation (3)

Let us start with a general process

$$\frac{dS}{S} = \mu_S dt + \sigma_S dW^Q \tag{B.1}$$

We will prove $\mu_S = r_h - r_f$ under the risk-neutral measure, given S is denominated by the home currency value of foreign currency. Let H_t and F_t denote the share prices of the assets in the home money market and foreign money market, reported in units of home currency and foreign currency, respectively, and normalized so that the time-zero share prices are both 1. Then

$$H_t = exp(r_h t) \text{ and } F_t = exp(r_f t)$$

The share price of foreign money market at time t in home currency is $F_t S_t$. Solving the stochastic differential Equation (B.1) gives the explicit formula

$$F_t S_t = S_0 exp\left\{\left(r_f + \mu - \frac{1}{2}{\sigma_S}^2\right)t + \sigma_S W_t\right\} \tag{B.2}$$

Note the present value of share price at home currency is

$$exp(-r_h t)F_t S_t = S_0 exp\{(r_f - r_h + \mu)t\} exp\left\{-\frac{1}{2}{\sigma_S}^2 t + \sigma_S W_t\right\} \tag{B.3}$$

Since under $\mathbb{Q}$ measure, the discounted share price of the asset foreign money market in home currency must be a martingale, also note that the second exponential term is a martingale by itself, to see this, let's define function as $U(t) = exp\{-\frac{1}{2}\sigma^2 t + \sigma W_t\}$ with information filter $(F_t)_{t\geq 0}$:

$$\begin{aligned} E(U(t+s)|F_s) &= E\big(exp\{-\sigma^2(t+s)/2 + \sigma W_{t+s}\}|F_s\big) \\ &= exp\{-\sigma^2 s/2 + \sigma W_s\} E\big(exp\{-\sigma^2 t/2 + \sigma(W_{t+s} - W_s)\}|F_s\big) = U(s) \end{aligned}$$

Note that the above short proof requires condition $E(exp\{\sigma W\}) = exp\{\sigma^2/2\}$ which is established when W is a normal distribution with $N(0,\ \sigma^2)$. Thus, we obtain $\mu_S = r_h - r_f$

(Q.E.D.)

Derivation of Equation (8)

There are several approaches to derive this PDE. Here we follow a heuristic manner. Assume that the risks inherent in exchange rate S and foreign market demand θ are spanned by the market of existing securities. Let call these securities "currency" and "demand" for brevity. Recall that in a risk-neutral world, the instantaneous return of holding foreign assets V_F is equal to risk-free rate

$$E\left[\frac{dV}{V}\right] + \frac{\pi}{V} = r_f \tag{B.4}$$

The Itô's lemma says that $V = V(S, \theta)$ obeys

$$\begin{aligned} dV = &\left[\frac{1}{2}{\sigma_S}^2 S^2 \frac{\partial^2 V_F}{\partial S^2} + \frac{1}{2}{\sigma_\theta}^2 \theta^2 \frac{\partial^2 V_F}{\partial \theta^2} + \rho\sigma_S\sigma_\theta S\theta \frac{\partial^2 V_F}{\partial S \partial \theta} + (r_h - r_f)S\frac{\partial V_F}{\partial S} + \mu\theta\frac{\partial V_F}{\partial \theta}\right]dt \\ &+ {}_S S\frac{\partial V_F}{\partial S}dW + \sigma_\theta \theta \frac{\partial V_F}{\partial \theta}dZ \end{aligned}$$

Since the expectation of the last two items is zero, substitute it into Equation (B.4) we obtain Equation (8).

Appendix C: Numerical Methods

We use finite difference method to solve the equation. The mesh grid takes the form of centered discretization scheme and is constructed as (θ_i, S_j) with i = 1, 2….n and j = 1, 2….m, where $\theta_i = 0, \Delta\theta, 2\Delta\theta \ldots\ldots n\Delta\theta$ and $S_j = 0, \Delta S, 2\Delta S \ldots\ldots m\Delta S$, applying Tylor expansion in the error order of $O(h^2)$

$$\frac{\partial V}{\partial S} = \frac{V_{i,j+1} - V_{i,j-1}}{2\Delta S}$$

$$\frac{\partial V}{\partial \theta} = \frac{V_{i+1,j} - V_{i-1,j}}{2\Delta\theta}$$

$$\frac{\partial^2 V}{\partial \theta^2} = \frac{V_{i+1,j} - 2V_{i,j} + V_{i-1,j}}{(\Delta\theta)^2}$$

$$\frac{\partial^2 V}{\partial S^2} = \frac{V_{i,j+1} - 2V_{i,j} + V_{i,j-1}}{(\Delta S)^2}$$

$$\frac{\partial^2 V}{\partial S \partial \theta} = \frac{V_{i+1,j+1} - V_{i+1,j-1} - V_{i-1,j+1} + V_{i-1,j-1}}{4\Delta S \Delta\theta}$$

After applying the above equations we have

$$a_i V_{i-1,j} + b_{i,j} V_{i,j} + c_i V_{i+1,j} + d_j V_{i,j-1} + e_j V_{i,j+1} + f_{i,j}\left(V_{i+1,j+1} - V_{i+1,j-1} - V_{i-1,j+1} + V_{i-1,j-1}\right) = -\pi_X$$

with the following six definitions

$$a_i = \frac{{\sigma_\theta}^2\theta^2}{2(\Delta\theta)^2} - \frac{\mu\theta}{2\Delta\theta}; \quad b_{i,j} = -\frac{{\sigma_\theta}^2\theta^2}{(\Delta\theta)^2} - \frac{{\sigma_S}^2 S^2}{(\Delta S)^2} - r_h; \quad c_i = \frac{{\sigma_\theta}^2\theta^2}{2(\Delta\theta)^2} + \frac{\mu\theta}{2\Delta\theta}$$

$$d_j = \frac{{\sigma_S}^2 S^2}{2(\Delta S)^2} - \frac{(r_h - r_f)S}{2\Delta S}; \quad e_j = \frac{{\sigma_S}^2 S^2}{2(\Delta S)^2} + \frac{(r_h - r_f)S}{2\Delta S}; \quad f_{i,j} = \frac{\rho\sigma_S S\sigma_\theta\theta}{4\Delta S\Delta\theta}$$

The boundary conditions are set as $V_{1,j} = 0$ and $V_{i,1} = 0$

To start the iteration we first set an initial guess $V^0_{i,j} = max\left[V^0_X, V^0_F - I, 0\right]$, where V^0_X, V^0_F are the present value of all operating profit flows given the capacity is in operation

$$V^0_X(\theta, S) = E^Q \int_t^\infty e^{-r_h(s-t)}\pi_X ds = \frac{1}{4\gamma}\left(\frac{\tau S\theta^2}{r_f - 2\mu - {\sigma_\theta}^2 - 2\rho\sigma_S\sigma_\theta} - \frac{2\nu_h\theta}{r_h - \mu} + \frac{{\nu_h}^2}{\tau S(2r_h - r_f)}\right)$$

$$V^0_F = E^Q \int_t^\infty e^{-r_h(s-t)}\pi_F ds = \frac{S}{4\gamma}\left(\frac{\theta^2}{r_f - 2\mu - {\sigma_\theta}^2 - 2\rho\sigma_S\sigma_\theta} - \frac{2\nu_f\theta}{r_f - \mu - \rho\sigma_S\sigma_\theta} + \frac{{\nu_f}^2}{r_f}\right)$$

Note that there is a series of limitation that all denominators have to be positive so that there is no assets bubble, we concatenate all the limits to the following requirement

$$2r_h > r_f > 2\mu + {\sigma_\theta}^2 + 2\rho\sigma_S\sigma_\theta$$

Then the $V^{iter>0}_{i,j}$ is computed for each iteration and the computation will be ceased at the tolerance $V^{iter+1}_{i,j} - V^{iter}_{i,j} < \varepsilon$, where ε is the extremely small number. At the terminal knot since we assume $R(j)$ and $D(i)$ are sufficiently large we assume the 3-order derivative impact at $V_{n-1,m-1}$ is negligible

$$\frac{\partial^3 V_{n-1,\ 2:n}}{\partial {\theta_{n-1}}^3} = \frac{V_{i+2,j} - 2V_{i+1,j} + 2V_{i-1,j} - V_{i-2,j}}{2(\Delta\theta)^3} = 0$$

Thus, we have

$$V_{n+1,j} = 2V_{n,j} - 2V_{n-2,j} + V_{i-3,j}$$

$$\frac{\partial^3 V_{[2:m],\ m-1}}{\partial {S_{m-1}}^3} = \frac{V_{i,j+2} - 2V_{i,j+1} + 2V_{i,j-1} - V_{i,j-2}}{2(\Delta S)^3} = 0$$

Thus, we have

$$V_{i,m+1} = 2V_{i,m} - 2V_{i,m-2} + V_{i,m-3}$$

The system is then solved using the method of successive over-relaxation (SOR), a variant of the Gauss-Seidel method, which is a method for solving linear systems of equations. The SOR method is an iterative finite difference method that includes a relaxation factor $1 < \omega < 2$ with purpose being to accelerate convergence and we set it $\omega = 1.2$.

Acknowledgments: *Michi Nishihara was supported by the JSPS KAKENHI (Grant numbers JP20K01769, 23K20613, JP24K00272).*

References

Aabo, T., C. Pantzalis, and J. Park. 2016. "Multinationality as Real Option Facilitator—Illusion or Reality?" *Journal of Corporate Finance* **38**: 1–17.

Aray, H., and J. Gardeazabal. 2010. "Going Multinational under Exchange Rate Uncertainty." *Journal of International Money and Finance* **29**, no. 6: 1171–1191.

Belderbos, R., T. W. Tong, and S. Wu. 2019. "Multinational Investment and the Value of Growth Options: Alignment of Incremental Strategy to Environmental Uncertainty." *Strategic Management Journal* **40**, no. 1: 127–152.

Blonigen, 2005. "A Review of the Empirical Literature on FDI Determinants." *Atlantic Economic Journal* **33**, no. 4: 383–403.

Chi, T., J. Li, L. Trigeorgis, and A. Tsekrekos. 2019. "Real Options Theory in International Business." *Journal of International Business Studies* **50**, no. 4: 525–553.

Choi, J. J., and C. Jiang. 2009. "Does Multinationality Matter? Implications of Operational Hedging for the Exchange Risk Exposure?" *Journal of Banking & Finance* **33**, no. 11: 1973–1982.

Clark, P., N. Tamirisa, S. Wei, A. Sadikov, and L. Zeng. 2004. "A New Look at Exchange Rate Volatility and Trade Flows." IMF (Working paper 235). https://www.imf.org/external/pubs/nft/op/235/op235.pdf

Conconi, P., A. Sapir, and M. Zanardi. 2016. "The Internationalization Process of Firms: From Exports to FDI." *Journal of International Economics* **99**: 16–30.

Darby, J., A. Hallett, J. Irelan, and L. Piscitelli. 1999. "The Impact of Exchange Rate Uncertainty on the Level of Investment." *The Economic Journal* **109**, no. 454: 55–67.

Dixit, A., and R. S. Pindyck. 1994. *Investment Under Uncertainty*, Princeton, NJ: Princeton University Press.

Duarte, V., D. Duarte, and D. Silva. 2024. "Machine Learning for Continuous-Time Finance." *The Review of Financial Studies* **37**, no. 11: 3217–3271.

Fillat, J., and S. Garetto. 2015. "Risk, Returns and Multinational Production." *Quarterly Journal of Economics* **130**, no. 4: 2027–2073.

Fillat, J., S. Garetto, and L. Oldenski. 2015. "Diversification, Cost Structure, and the Risk Premium of Multinational Corporations." *Journal of International Economics* **96**, no. 1: 37–54.

Froot, K., and J. Stein. 1991. "Exchange Rates and Foreign Direct Investment: An Imperfect Capital Markets Approach." *The Quarterly Journal of Economics* **106**, no. 4: 1191–1217.

Helpman, E., M. Melitz, and S. Yeaple. 2004. "Export versus FDI with Heterogeneous Firms." *American Economic Review* **94**, no. 1: 300–316.

Garetto, S., L. Oldenski, and N. Ramondo. 2018. "Multinational Expansion in Time and Space." NBER (Working paper). https://www.nber.org/system/files/working_papers/w25804/revisions/w25804.rev0.pdf

Goldberg, L., and C. Kolstad. 1995. "Foreign Direct Investment, Exchange Rate Variability and Demand Uncertainty." *International Economic Review* **36**, no. 4: 855.

Jeanneret, A. 2016. "International Firm Investment under Exchange Rate Uncertainty." *Review of Finance* **20**, no. 5: 2015–2048.

Nguyen, Q., Kim, and T. Papanastassiou. 2018. "Policy Uncertainty, Derivatives Use, and Firm-Level FDI." *Journal of International Business Studies* **49**, no. 1: 96–126.

Rob, R., and N. Vettas. 2003. "Foreign Direct Investment and Exports with Growing Demand." *Review of Economic Studies* **70**, no. 3: 629–648.

Song, S., M. Makhija, and S. M. Kim. 2015. "International Investment Decisions under Uncertainty: Contributions of Real Options Theory and Future Directions." *Journal of Management & Organization* **21**, no. 6: 786–811.

Sung, H., and H. Lapan. 2000. "Strategic Foreign Direct Investment and Exchange-Rate Uncertainty." *International Economic Review* **41**, no. 2: 411–423.

Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 3, No. 1, 2025 DOI: 10.54116/jbdai.v3i1.49

Applications of Analytics in Disease Prediction Types

Cheng-Yi Tsai
Penn State University
czt5442@psu.edu

Satish Mahadevan Srinivasan
Penn State University
sus64@psu.edu

Abhishek Tripathi
The College of New Jersey
tripatha@tcnj.edu

Abstract

Predictive analytics has immense potential in disease-type classifications. The key is to identify the set of genetic and clinical variables that can serve as predictors for disease classification purposes. However, the predictive and the prescriptive models both suffer from high dimensionality of these predictors. Therefore, it becomes important to identify a subset of these genetic and clinical variables that can be used for disease-type predictions. Earlier studies identified a subset of 978 landmark genes that can infer the expression values of the remaining gene in the human genome with ∼81% accuracy. This study focused on understanding if there is any significant difference in the characteristics of the landmark and non-landmark genes. Several experiments were performed on diseased tissues that were classified across race, ethnicity, and disease types, with an objective to identify the number of differentially expressed genes within the landmark and non-landmark gene sets. Statistically, there was no conclusive evidence to support the hypothesis that there is a significant difference in the number of differentially expressed genes across the landmark and non-landmark gene sets.

Keywords *L1000 dataset analysis, landmark genes, non-landmark genes, differentially expressed genes, cancer tissues, RNA-Seq data.*

1. Introduction

Cancer is a disease that is characterized by uncontrolled cell growth. It is a heterogeneous disease that consists of many different subtypes. Early diagnosis of cancer type has become a priority for many cancer researchers because it can facilitate the subsequent genetic and clinical management of the patients. Cancer research is mainly focused on primarily identifying the cancer type and, secondarily, on classifying patients into high- or low-risk groups. These two tasks involve analyzing large datasets and building predictive and prescriptive models that can decode the interaction between both the clinical and genetic variables. Therefore, biomedical and bioinformatics research teams have started to rely heavily on machine learning (ML) and artificial intelligence (AI) techniques. These

techniques have been proven to model the progression and treatment of cancerous conditions. In addition, these techniques have the ability to detect key features from complex datasets.

Even though ML methods can help in detecting cancer types and help us understand the progression of the disease, an appropriate level of validation is still needed for these methods to be used in clinical practice. Studies in the past (Duncan et al. 2008; Liang et al. 2015; Danaee et al. 2017; Bailey et al. 2018; Huang et al. 2018; Saltz et al. 2018; Way et al. 2019) applied various ML techniques that focused on the impact of the genetic variables (genes) on the clinical responses. These techniques also have applications in cancer research. By using ML techniques, scientists can screen early stages of cancer progression by analyzing the genetic variables to find its nature before the symptoms show up. At the same time, with the advent of new technologies in the field of medicine, large amounts of cancer data have been collected and are available to the biomedical research community. Within the data lie complex patterns that can be mined efficiently by using the current state of the ML techniques. However, an accurate prediction of a disease outcome is one of the most interesting and challenging tasks for the biomedical research community. As a result, ML methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relationships between them from complex datasets, while they are able to effectively predict future outcomes of a cancer type (Kourou et al. 2015).

Advances in the area of personalized medicine are significantly fueled by advances in ML and AI techniques. Personalized medicine is important because it has increasingly been applied with success in clinical trials. Early detection of cancer increases the survival rate. Determining which genes contribute to decreased survival likelihood in cancer patients can provide clinically relevant biomarkers. This study was an effort in developing data analytics techniques to assess the RNA-sequencing (RNA-Seq) data from the National Cancer Institute (NCI) database (Tomczak et al. 2015). Early diagnosis of cancer, including cancer susceptibility, recurrence, and survival prediction, can be efficiently performed by using various ML and AI techniques. The availability of larger datasets that contain gene expression profiling captured over the period of time can significantly improve our ability for prognosis in cancer patients (Clayman et al. 2020a, 2020b).

The gene expression profiling measures which genes are expressing in a cell at any given moment. Gene expression profiling measures the messenger RNA (mRNA) levels, showing the pattern of genes expressed by a cell at the transcription level (Fielden and Zacharewski 2001). Gene expression profiling is used by a variety of researchers in the area of biomedical engineering, from molecular biologists to environmental toxicologists. This technology can provide accurate information on gene expression for the entire human genome. Different techniques are used to determine gene expressions, including DNA microarrays and sequencing technologies, for example, the RNA-Seq (Hurd and Nelson 2009).

The genome is a collection of biological information, but it is unable to disclose that information on its own. The initial product of the genome expression is the transcriptome. RNA-Seq is a state-of-the-art approach that can determine the quantity and sequences of RNA in a sample by using the next-generation sequencing. It analyzes the transcriptome, which indicates which of the genes in our DNA are turned on and off, and to what extent, and corresponding gene expression levels. RNA-Seq possesses the capability to measure the expression values of the genes across the transcriptome. RNA-Seq also promises to discover de novo transcriptome with high specificity in different species. It is a relatively new method and has already provided unprecedented insights into the transcriptional complexities of a variety of organisms.

RNA-Seq is a relatively modern approach used to generate read counts of complementary DNA in parallel to generate a comprehensive set of corresponding gene expression levels. Some ML models effectively generalize between microarray and RNA-Seq data. RNA-Seq and microarray-based predictive models can predict clinical outcomes with a similar performance. RNA-Seq better represents transcript expression patterns that map onto clinically and genetically generated cancer subgroups compared with microarray data (Zhang et al. 2015).

L1000 is a high-throughput gene expression assay that measures the mRNA transcript abundance of 978 "landmark" genes from human cells. Landmark gene expression levels measured with the L1000 microarray have been assessed in The Library of Integrated Cellular Signatures (LINCS), which uses expression of ∼1,000 landmark genes to infer ∼21,000 target genes with ∼81% accuracy. LINCS measured ∼1.4 million gene expression profiles of heterogeneous normal and diseased tissue. Computational analysis of large gene expression indicates that it would be feasible to derive sufficient information about the transcriptional state of a cell by measuring only a subset of expressed genes. In addition to that, the genome-wide expression analysis has shown that gene expression is highly correlated, with a small cluster of genes that exhibit similar expression patterns across cell states. The genes that are part of the landmark genes have an expression profile that has been determined as being informative to characterize the transcriptome and can be directly measured from the L1000 assay. These genes have a good predictive power for inferring the expression of other genes that are not directly measured in the assay (Chen et al. 2016; Clayman et al. 2020a, 2020b).

This study used the The Cancer Genome Atlas Program (TCGA)/NCI Genomic Data Commons (GDC) dataset to select RNA-Seq and clinical outcome data for the present analysis. The TCGA is a comprehensive set of studies compiled through the National Institutes of Health that includes genetic and clinical data within individual patient samples. TCGA data have been thoroughly assessed because TCGA data includes a large depth (large sample size) and breadth (heterogeneity of sample types and clinical data) for various applications, including predictive analytics and ML (Tomczak et al. 2015). The TCGA data, along with other cancer research data, are currently hosted through the GDC, a data repository initiated in June 2016 for its applications in precision medicine.

Previous studies (Clayman et al. 2020a; Liang et al. 2015; Quang et al. 2015; Duan et al. 2016; Chen et al. 2018; Tsagri et al. 2018; Duncan et al. 2008; Danaee et al. 2017; Kursa and Rudnicki 2010; Kogelman and Kadarmideen 2014; Petralia et al. 2016; Chen et al. 2016; Clayman et al. 2020b) used different predictive analytics methods, such as clustering methods, deep learning, and feature selection, to evaluate the impact of genes on clinical responses. Many of these studies (Clayman et al. 2020a; Duan et al. 2016; Chen et al. 2018; Tsagri et al. 2018; Duncan et al. 2008; Danaee et al. 2017; Kursa and Rudnicki 2010; Kogelman and Kadarmideen 2014; Petralia et al. 2016; Chen et al. 2016; Clayman et al. 2020b) have assessed the impact of clinical and genetic variables on clinical results such as metastases possibility and survival time. Some of the studies (Lin et al. 2018; Way et al. 2019; Bailey et al. 2018; Saltz et al. 2018; Malta et al. 2018) assessed heterogeneous datasets, including data from several cancer types. Others evaluated homogeneous datasets with data from a single cancer type.

However, studies in the past failed to understand the significance and role of landmark genes in disease-type predictions. The exact nature and characteristics of landmark genes are still unknown. It is unknown as to how different the landmark genes are when compared with non-landmark genes with respect to predicting disease types. By using statistical techniques, we explored if there is any significant difference in the characteristics of the landmark and non-landmark genes. The present study chose genes based on selection criteria, that is, landmark or non-landmark, and compared the ability of genes and/or gene sets to predict clinical outcomes.

This study sought to understand if there is any significant difference in the characteristics of the landmark and non-landmark genes. Earlier studies (Duncan et al. 2008; Chen et al. 2016; Danaee et al. 2017; Ramaker et al. 2017; Bailey et al. 2018; Chen et al. 2018; Huang et al. 2018; Way et al. 2018; Daoud and Mayo 2019; Clayman et al. 2020a) only focused on using the expression values of the landmark genes to determine the expression values of the non-landmark genes but did not discuss whether the landmark genes are any different from the non-landmark genes, that is, could we find a different set of non-landmark genes and say that they are similar in characteristics to the original set of identified landmark genes.

2. Literature Survey

Personalized medicine can be facilitated by analyzing both the genomic and clinical variables. Genes interact with one another and with the different clinical variables such as survival, cancer stage, gender, and age of diagnosis to determine the disease type. For example, let us consider cancer types, namely, prostate, breast, ovarian, and pancreas cancer. All these cancer types possess several genes in common, including the breast cancer 1, early onset (*BRCA1*) and *BRCA2*. The mutation in *BRCA1* and *BRCA2* is associated with a Gleason score $\geq$ 8, T3/T4 tumor stage, nodal involvement, and metastases at the time of diagnosis in prostate cancer patients (Castro et al. 2013). With another gene, *TP53,* the presence or absence of a *TP53* mutation has been identified as a predictor of survival in prostate cancer patients (Ecke et al. 2010). However, the clinical variables of prostate cancer patients such as the tumor state can be used to predict treatment resistance. Prostate cancer adenocarcinoma metastases possess greater treatment resistance as opposed to primary tumors and possess more de-differentiation of phenotypes. The prostate adenocarcinomas have a strong inverse relationship between stemness index and reduced leukocyte fractions, indicative of reduced immune response when tissue is more differentiated as indicated by mRNA expression-based stemness index response (Malta et al. 2018).

Studies in the past associated with the GDC/TCGA database compared distinct cancer subsets (Bailey et al. 2018) and specifically used deep learning to study immunohistochemical data (Saltz et al. 2018) and mRNA expression-based stemness index (Malta et al. 2018). One study assessed RNA-Seq data available on TCGA (Lim et al. 2020), and one studied a cancer pathway, Ras, across various cancer types by incorporating RNA-Seq data (Way et al. 2018). Also, RNA-RNA interactions have been explored for different cancer subtypes by using deep learning techniques (Dutil et al. 2018). However, not much has been reported with regard to the interactions between the RNA and clinical data for different cancer subtypes. A study on the GDC/TCGA database also compared distinct cancer subsets (Bailey et al. 2018). Twenty-six distinct computational tools and/or algorithms established driver genes that influence distinct cancer and/or cell types and anatomic sites within the TCGA dataset. These include algorithms such as a random forest algorithm used for predicting oncogenes and tumor suppressor genes from somatic

mutations. The consensus list/union of the gene sets generated through each of these 26 approaches were pooled for downstream analysis, which included methods that factored in weighting of genes based on performance for distinct cancer types (Bailey et al. 2018).

Genes play a very significant role in the progression of diseases. Some genes are predictive of cancer severity, whereas other genes, including *TP53,* are protective against the development of cancer. Mutations in *TP53* results in alteration in stress and cell-cycle transcriptional regulator genes in few cancer types, and the intensity of the alteration vary across other cancer types. Target genes that are either up- or downregulated in response to a *TP53* mutation involve functions such as cell-cycle inhibition, apoptosis, p53 regulation, and DNA damage response. Genes, for example, *TP53,* that influence pathways that regulate many other genes are especially important to consider when assessing clinical outcomes given that their expression can both directly and indirectly modulate cancer and/or tumor stage progression (Parikh et al. 2014).

The expression level of the genes obtained through RNA-Seq can be used to build predictive models that can predict the outcomes of diseases. A combination of genetic and clinical characteristics can increase the ability to predict overall survival of prostate cancer patients. Personalized medicine is important because it has increasingly been applied with success in clinical trials. In addition, early detection of cancer produces an increase in survival rate and consideration of clinical variables, along with RNA-Seq data, can be used to increase efforts at early detection of cancer (Clayman et al. 2020a). One of the studies focused on developing data analytics techniques to analyze the RNA-Seq and clinical data gathered from the NCI database. By using the data modalities on the genomic and clinical data obtained from the TCGA and by applying integrative clustering, Liang et al. (2015) reported effective differentiation of clinical subgroups for ovarian cancer. A study also investigated the relationship between genetic and clinical variables by accounting for both coding and non-coding genetic variants (Quang et al. 2015).

Computational costs of biological data analysis call for increasingly efficient methods of determining which genetic and clinical factors are most relevant for understanding the overall genetic and clinical profiles of human patients (Duan et al. 2016). This challenge is especially difficult given that distinct individuals can possess distinct profiles of genetic expression, and certain genetic conditions can be more readily captured than others when using a varying number of genetic features. Dimensionality reduction methods, such as random forest analysis, k-means clustering, and principal component analysis (PCA), are often used in tandem to capture essential elements of the data that explain larger datasets when using a subset of relevant features. Dimensionality reduction methods such as PCA are effective methods of data representation when linear relationships are present. PCA can detect multiple types of cancer while also selecting relevant features (Chen et al. 2018). A study to predict cancer outcomes (Tsagri et al. 2018) applied feature selection methods, including PCA and Boruta random forest, for dimensionality reduction. The utility of the *k*-means clustering for protein expression and cancer outcomes is demonstrated in the study by Duncan et al. 2008. Selection methods were further refined by applying the random forest decision tree classifier to determine a smaller subset of important genes to use for downstream analysis in several clustering methods, including *k*-means, partition around medoids, and res-hierarchical clustering. These clustering methods were used to generate subsets within the dataset in an unsupervised manner.

One study used deep learning to assess the entire search space of gene expression levels from RNA-Seq data (Danaee et al. 2017). Another study implemented dimensionality reduction and feature selection methods to reduce computation and model complexity. When dealing with results of gene expression measurements in the context of cancer, identification of a minimal-optimal set of genes related to cancer is often useful for establishing genetic markers (Kursa and Rudnicki 2010). In this way, Boruta analysis can be applied specifically to the approach of identifying the minimal-optimal set of genes as a selection method to restrict analysis to relevant genes. Previous studies implemented thresholding, weighting (Kogelman and Kadarmideen 2014), and networks analysis (Petralia et al. 2016) to assess whether biologically relevant interactions can improve model performance.

A set of 978 landmark genes has been established as predictors of the remaining genes in a microarray dataset analyzed by Chen et al. (2016). When applying 978 landmark genes as inputs, a deep learning method (D-GEX) results in lower error compared with linear regression in predicting expression of 81.31% of target genes in an independent RNA-Seq–based GTEx dataset (Chen et al. 2016). As an extension of the analysis performed by Chen et al. (2016), these 978 landmark genes from the L1000 dataset were selected from the GDC's RNA-Seq dataset to assess whether 978 landmark genes improve clustering (Clayman et al. 2020a).

Dimensionality reduction methods such as PCA are used to select relevant features. In addition to that, the unsupervised learning technique, *k*-means clustering performs well when applied to data with low effective dimensionality. Our previous study (Clayman et al. 2020b) showed that 978 landmark genes better differentiated *k*-means clusters compared with 978 randomly selected non-landmark genes. *K*-means clusters generated from the landmark genes show more separation of cluster groups when plotted against the first two principal components, which capture a greater proportion of variation for the 978 landmark genes (Clayman et al. 2020b). Analysis of these results

suggests that the 978 landmark genes better represent the overall genetic profile of these heterogeneous samples. However, clustering results varied when using the 978 landmark genes versus the 978 non-landmark genes as features, depending on whether clustering was performed on the heterogeneous versus the homogeneous datasets. For the heterogeneous dataset, the percentage of variation captured by each of the first two principal components was greater for the 978 landmark genes (PCA1, 13.1%; PCA2, 9.2%) versus the 978 non-landmark genes (PCA1, 9.4%; PCA2, 6.2%), with similar results for the homogeneous dataset. Variability, depending on the set of genes selected, is also depicted based on the distinct appearance of cluster plots, which possess more visual overlap and greater between-cluster sum of squares for the non-landmark genes compared with the landmark genes for both the homogeneous and heterogeneous datasets. *K*-means clustering results coincide with the clinical variable of the Ann Arbor cancer stage to a greater extent when using non-landmark genes as features compared with landmark genes (Clayman et al. 2020b).

The study by Clayman et al. (2020a) depicted the use of 978 landmark genes as a more effective method of identifying distinct clusters of individuals according to visualization of data clusters against the first two principal components of the data when assessing large heterogeneous datasets. Clusters in these plots are more distinct compared with cluster plots generated by using 978 randomly selected non-landmark genes in the dataset, which supports the use of these landmark genes as a representation of the genetic profile of these samples when assessing heterogeneous datasets (Clayman et al. 2020a; Chen et al. 2016). In contrast, non-landmark genes capture more of the variation in the data for homogeneous and heterogeneous datasets studied here. Despite this, the non-landmark genes allow for clustering into groups more consistent with clinical variables for the homogeneous dataset compared with the 978 landmark genes. Certain genes or clinical variables can be more predictive of clustering results than others. When assessing the separation of groups, the role of sets of individual genes and clinical variables can be examined further. Cluster analysis can be used to inform future studies on the ability of genes to predict clinical variables as well as the ability of clinical variables to characterize clusters derived from gene expression results, as examined in this study. This can be especially relevant toward applications for personalized medicine such as treatment responsiveness, depending on the combination of genetic and clinical variables (Clayman et al. 2020a). Predictive models of cancer outcomes can be built by specific protein expression levels with RNA-Seq. The overall survival of prostate cancer patients can be precisely predicted by genetic and clinical characteristics (Clayman et al. 2020a, 2020b). Personalized medicine has become a new trend because there are many successful cases in clinical trials with personalized medicine. Moreover, the survival rate can be increased by early detection of cancer with clinical variables and RNA-Seq data (Clayman et al. 2020a, 2020b).

Other studies evaluated histopathologic imaging data (Ash et al. 2018; Saltz et al. 2018), multi-omics data (Liang et al. 2015; Chaudhary et al. 2018; Lin et al. 2018; Way et al. 2019), mRNA data (Azarkhalili et al. 2018), microarray data (Daoud and Mayo 2019), or RNA-Seq data (Danaee et al. 2017) from the TCGA commons to predict clinical outcomes by using deep learning methods, including convolutional and variational autoencoders. Studies implemented other ML techniques (Huang et al. 2018), support vector machines (Bailey et al. 2018), ensemble methods (Way et al. 2019; Bailey et al. 2018), construction of latent dimensionalities and PCA (Way et al. 2019), feature selection, and clustering (Liang et al. 2015) to assess the impact of genes on clinical responses. A study by Duncan et al. (2008) applied *k*-means clustering to assess protein expression and cancer outcomes. PCA can be applied in predictive analysis of multiple types of cancer by selecting relevant features and capturing linear relationships in the data to reduce the dimensionality of data (Chen et al. 2018). Some of these studies evaluated homogeneous datasets that contain data from a single cancer type, such as prostate cancer (Saltz et al. 2018), breast cancer (Danaee et al. 2017), liver cancer (Chaudhary et al. 2018), lung adenocarcinoma (Chaudhary et al. 2018), and acute myeloid leukemia (Lin et al. 2018). Other studies assessed heterogeneous datasets with data from multiple tumor types (Lin et al. 2018) or multiple cancer types (Ash et al. 2018; Azarkhalili et al. 2018; Bailey et al. 2018; Huang et al. 2018; Way et al. 2018). A study by Petralia et al. (2016) evaluated gene and protein networks within TCGA breast cancer data using the random forest classifier. Previous studies of the GDC used feature selection to reduce the set of genes used for predicting clinical outcomes. Landmark genes have not been extensively used for assessing the GDC dataset and have not been assessed for further feature selection approaches to further reduce this set of genes for predictive analysis (Clayman et al. 2020a, 2020b).

3. Methods and Materials

A total of three datasets were used in this study. The clinical and RNA-Seq dataset (dataset 1) was obtained from the NCI's GDC repository. This dataset consists of clinical and genetic information for tissues of 55 cancer types. In total, there are 13,122 observations (instances) of 83 clinical variables and >20,000 genetic variables. Two L1000 datasets, namely, the microarray version of the L1000 dataset (dataset 2) and the RNA-Seq version of the L1000 dataset (dataset 3) were also analyzed in this study. A brief introduction to all three datasets is provided here.

Table 1: The number of tissue samples per disease types in dataset 1.

Disease (Cancer) Type	No. Samples
Breast	1,485
Kidney	1,448
Brain	759
Colon	675
Prostate gland	660
Bladder	488
Skin	474
Stomach	460
Pancreas	188
Testis	165

3.1. Datasets

3.1.1. Dataset 1

Of 13,122 diseased tissues of 55 disease types, a random subset of 6,802 diseased tissues across 10 different disease types were analyzed in this study. A total of ∼22k genetic variables were considered as predictors for each diseased tissue. The total numbers of tissue samples across each disease type considered in this study are listed in Table 1.

Due to the high dimensionality of the datasets, the descriptions of the individual predictors are not provided here.

3.1.2. Dataset 2

The L1000 microarray-based dataset in the GCTx format contained expression data of 22,268 (∼22k) genes (rows) across 129,158 tissues (columns). Of the ∼22k rows, the first 978 rows were the landmark genes and the remaining 21,290 rows were the non-landmark genes whose expression values were predicted by using the landmark genes. A sample dataset that consisted of ∼22k genes across 6,802 diseased tissues was obtained by matching the tissue identifier in both dataset 1 and dataset 2. To eliminate the variability, the sample dataset was quantile normalized into the numerical range between 4 and 15, that is, the expression values of the genes were in the range between 4 and 15.

3.1.3. Dataset 3

The L1000 RNA-Seq-based dataset contained expression data of 22,268 (∼22k) genes (rows) across 129,158 tissues (columns). Of the ∼22k rows, the first 978 rows were the landmark genes and the remaining 21,290 rows were the non-landmark genes. A sample dataset that consisted of ∼22k genes across 6,802 diseased tissues was obtained by matching the tissue ids in both dataset 1 and dataset 3. To eliminate the variability, the sample dataset was quantile normalized across all samples such that they know all have the same distribution (e.g., same mean ± standard deviation [SD]).

3.2. Tools and Techniques

3.2.1. Analysis of variance

It is a statistical tool used to detect differences between experimental group means. Analysis of variance (ANOVA) is performed in experimental designs with one dependent variable that is a continuous parametric numerical outcome measure, and multiple experimental groups within one or more independent (categorical) variables. The independent variables are called factors, and groups within each factor are referred to as levels. ANOVA, similar to linear regression and general linear models, quantifies the relationship between the dependent variable and the independent variable(s). There are three different general linear models for ANOVA: the fixed effects model, which makes inferences that are specific and valid only to the populations and treatments of the study; the random effects model, which makes inferences about levels of the factor that are not used in the study, that is, this model pertains to random effects within levels, and makes inferences about a population's random variation; and the mixed effects model, which contains both the fixed and the random effects (Sawyer 2009).

Assumptions for ANOVA: a data set should meet the following criteria before performing ANOVA (Sawyer 2009):

Parametric data: A parametric ANOVA requires parametric data (ratio or interval measures). There are nonparametric, one-factor versions of ANOVA for nonparametric ordinal (ranked) data, specifically the Kruskal-Wallis test for independent groups and the Friedman test for repeated measures analysis.

Normally distributed data within each group: The fundamental assumption of parametric ANOVA is that each group of data (each level) be normally distributed. The Shapiro-Wilk test is commonly used to test for normality for group sample sizes (N) < 50 and the D'Agnostino modification is useful for larger samplings (N > 50).

Homogeneity of variance within each group: Because ANOVA compares normal distribution curves of datasets, these curves need to be similar to each other in shape and width for the comparison to be valid. In other words, the amount of data dispersion (variance) needs to be similar between groups. Two commonly invoked tests of homogeneity of variance are by Levene and by Brown and Forsthye.

Independent observations: A general assumption of parametric analysis is that the value of each observation for each subject is independent of the value of any other observation. For independent groups designs, this issue is addressed with random sampling, random assignment to groups, and experimental control of extraneous variables.

Most commercially available statistics programs perform normality and homogeneity of variance tests. Determination of the parametric nature of the data and soundness of the experimental design is the responsibility of the investigator, reviewers, and critical readers of the literature (Sawyer 2009).

Robustness of ANOVA to violations of normality and variance assumptions: ANOVA tests can handle moderate violations of normality and equal variance if there is a large enough sample size and a balanced design. The validity of ANOVA is said to be "robust" in the face of violations of normality assumptions if there is an adequate sample size. ANOVA is more sensitive to violations of the homogeneity of variance assumption, but this is mitigated if sample sizes of factors and levels are equal or nearly so. If normality and homogeneity of variance violations are problematic, then there are three options: transform the data to best mitigate the violation; use one of the nonparametric ANOVAs, but at the cost of reduced power and being limited to one-factor analysis; or identify outliers in the dataset by using formal statistical criteria. In that case, use caution in deleting outliers from the dataset; such decisions need to be justified and explained. Removal of outliers will reduce deviations from normality and homogeneity of variance (Sawyer 2009).

3.2.2. Kruskal Wallis H test

This test is a nonparametric alternative to the one-way ANOVA. The Kruskal Wallis H test is used when the assumptions for ANOVA are not met. This test is also referred to as one-way ANOVA on ranks because the ranks of the data values are used in the test rather than the actual data points. This test determines whether the medians of two or more groups are different. The hypotheses for the test are the following:

H0: Population medians are equal.

H1: Population medians are not equal.

The Kruskal-Wallis H test is more suitable for analysis of the dataset in which the sample size is small (<30). For the dataset that is not normally distributed and contains some strong outliers, it is more appropriate to use ranks rather than actual values to avoid the testing being affected by the presence of outliers or by the non-normal distribution of data. This test also assumes that the observations are independent of each other. The Kruskal Wallis H test will determine if there is a significant difference between groups. However, this test cannot determine which groups are different. To determine which groups are significantly different, a post hoc test needs to be performed.

The assumptions for the Kruskal Wallis H test are the following:

- The test is more commonly used when an independent variable has three or more levels.
- The scales for the dependent variable are either ordinal, ratio, or interval.
- All observations should be independent; in other words, there should be no relationship between the members in each group or between groups.
- All groups should have the same shape distributions.

3.2.3. Paired Wilcoxon signed-rank test

The nonparametric analog of the *t*-test is the Wilcoxon signed-rank test and is used when the one-sample *t*-test assumptions are violated. The pairwise Wilcoxon signed-rank test is performed as a post hoc test to determine which groups are significantly different from other groups. The assumptions of the Wilcoxon signed-rank test are as follows:

- The differences between the data values are continuous (not discrete).
- The distribution of each difference (of the data values) is symmetric.
- The differences of the data values are mutually independent.
- The differences of the data values all have the same median.
- The measurement scale of the data value is interval.

In summary, parametric tests are more commonly used than are nonparametric tests. However, parametric tests require an important assumption, which is the assumption of normality. This means that the distribution of sample means is normally distributed. But, when this assumption is not satisfied, the parametric tests can be misleading. In such situations, nonparametric tests are the available alternative. The nonparametric tests are statistical methods based on signs and ranks. When used, nonparametric tests convert the original data into the order size instead of using the original data value and only uses the rank or signs. Although this can result in the loss of information, but, when the data are not normal, the nonparametric analysis has more statistical power than the parametric analysis. In particular, when the means of the sample group are not normally distributed and when the variances are equal across groups, then nonparametric statistical techniques are excellent alternatives. Another advantage of using the nonparametric test is that it is not sensitive to outliers (Nahm 2016).

3.2.4. Experimental design strategies

This study focused on developing data analytics techniques to assess the genetic and clinical data gathered from the GDC. This is a relevant area of research given that these research techniques have applications in analysis of bioinformatics datasets in general.

3.2.5. Data collection and integration of genetic and clinical GDC data

Data were downloaded from the GDC by using the GDC Data Portal. RNA-Seq data for each cancer type were appended into a dataframe, which included a corresponding sample id for later integration of RNA-Seq and clinical data, which allowed for individual subject-level data analysis. The microarray (LINCS L1000) dataset, which consisted of the expression values of >20,000 genes across 129k samples was collected from the Gene Expression Omnibus (GEO) repository. This dataset is curated by the Broad Institute, which is publicly available in the GEO repository.

3.2.6. Data normalization

The DeSeq Bioconductor package in R (Love et al. 2014) was used to normalize the RNA-Seq dataset for all the cancer types. This normalization method accounts for each gene length as well as the number of observations in the dataset.

3.3. Experiments

3.3.1. Experiment A: Perform ANOVA and the Kruskal Wallis H test to compare the number of differentially expressed genes within the landmark and non-landmark gene sets across different tissue samples (6,802 tissue samples)

Initially, ANOVA is performed on all the genes within the landmark gene set and the non-landmark gene sets to determine if a gene within the set is or is not differentially expressed (Koch et al. 2018). The null and the alternate hypothesis of the ANOVA is given as follows:

H_0 = the gene is not differentially expressed

H_a = the gene is differentially expressed

A significant p-value ($p < 0.05$, given $\alpha = 0.05$) that results on an ANOVA test would indicate that the gene is differentially expressed (Koch et al. 2018). For each of the gene sets (landmark and non-landmark), the total numbers of genes that are differentially expressed is identified by performing ANOVA. Then, when using the Kruskal Wallis H test, it is determined if the number of genes that is differentially expressed within the different sets are or are not similar. If the null hypothesis is rejected, then it can be concluded that at least one of the sets has a different number of differentially expressed genes compared with the others. The design strategy for this experiment is summarized in Figure 1.

Note here that, due to the small sample size (n = 15), the nonparametric test was performed because we could not assume that the distribution of sample means is normal. The nonparametric Kruskal Wallis H test was performed here because the variables were measured on a continuous scale, the independent variable consists of two or more categorical, independent, or unrelated group, and there is no relationship between the observations in each group (Nahm 2016).

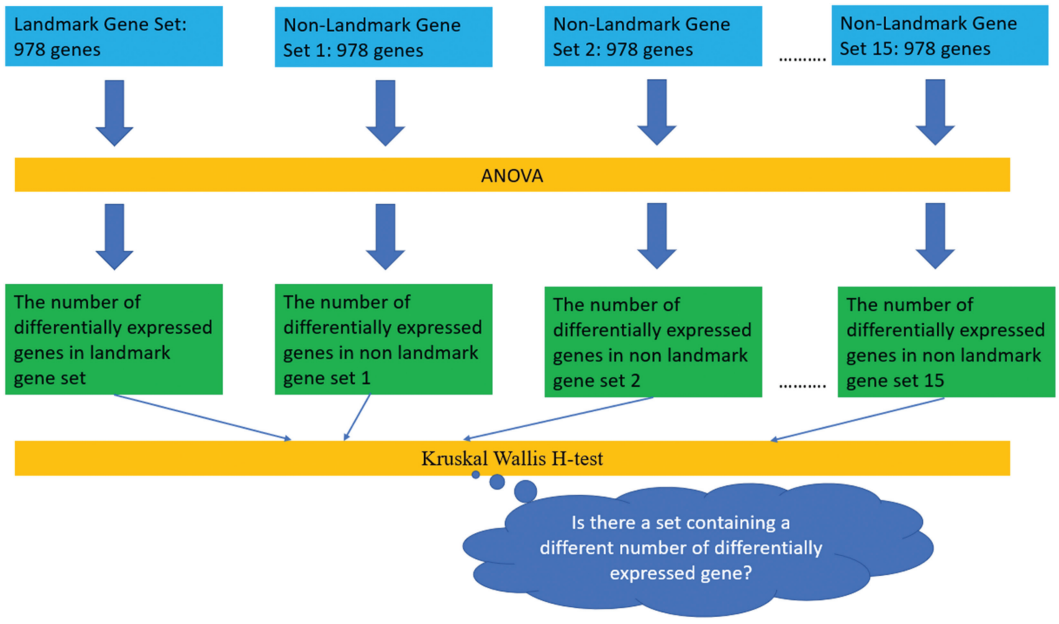


Figure 1: Design strategy for experiment A.

3.3.2. Experiment B: Perform an ANOVA and a Kruskal Wallis H test to compare the number of differentially expressed genes within the landmark and non-landmark gene sets across different tissue samples classified by race, ethnicity, and disease types

To begin with, the diseased tissues were classified (split) by categories either by race or ethnicity, or by disease types. Within subcategories (e.g., Asian, White, Black) of each category (e.g., race), ANOVA was performed on all the genes within the landmark gene set and the non-landmark gene sets to determine if a gene within the set was or was not differentially expressed. For each subcategory within a category, ANOVA was performed to identify the number of differentially expressed genes within each of the gene sets (landmark and non-landmark) (Koch et al. 2018). For each subcategory, within each category, the Kruskal Wallis H test was used to determine if there was a significant difference in the number of differentially expressed genes. If the null hypothesis is rejected, then it can be concluded that at least one or more subcategories have a different number of differentially expressed genes. Also, the Kruskal Wallis H test was performed within each category to determine if the number of genes that were differentially expressed within the different gene sets across different subcategory were or were not similar. If the null hypothesis is rejected, then it can be concluded that at least one of the sets within the gene sets (landmark or non-landmark gene sets) has a different number of differentially expressed genes across different subcategories. Finally, the pairwise Wilcox signed-rank test was used to identify the subcategory or the gene set that was different from the others in terms of the number of differentially expressed genes. The pairwise Wilcox signed-rank test is a post hoc test because the Kruskal Wallis H test is an omnibus test statistic. It cannot tell which specific groups of the independent variable are statistically significantly from each other. The design strategy for the experiments conducted in this section is summarized in Figure 2.

3.3.3. Experiment C: Perform correlation studies to determine the pairwise correlation range of the genes in the landmark and non-landmark gene sets

In this experiment, the expression values of the 978 landmark genes are compared with the expression values of the 15 different randomly selected 978 non-landmark genes set across ~129k tissue samples. An ANOVA was performed on a dataset that consisted of randomly selected 100 correlation values between gene pairs from both the landmark gene set and the 15 different non-landmark gene set. Here, rejecting the null hypothesis would indicate that there is no significant difference in the correlation values of the gene pairs across the landmark and non-landmark gene sets. Here, the parametric test ANOVA is performed on the dataset (16 gene sets), which consisted of the randomly selected 100 correlation values between gene pairs because the sample size (n = 100) was good enough to assume that the sample was taken from the normally distributed population, that is, each sample was drawn independently of the other samples, the variance in different groups (gene sets) was the same, and the correlation values in each group was continuous. Visualizations, such as boxplots, would highlight the variations in the correlations of the gene pairs across the different set of genes. The design strategy for the experiment conducted in this section is summarized in Figure 3.

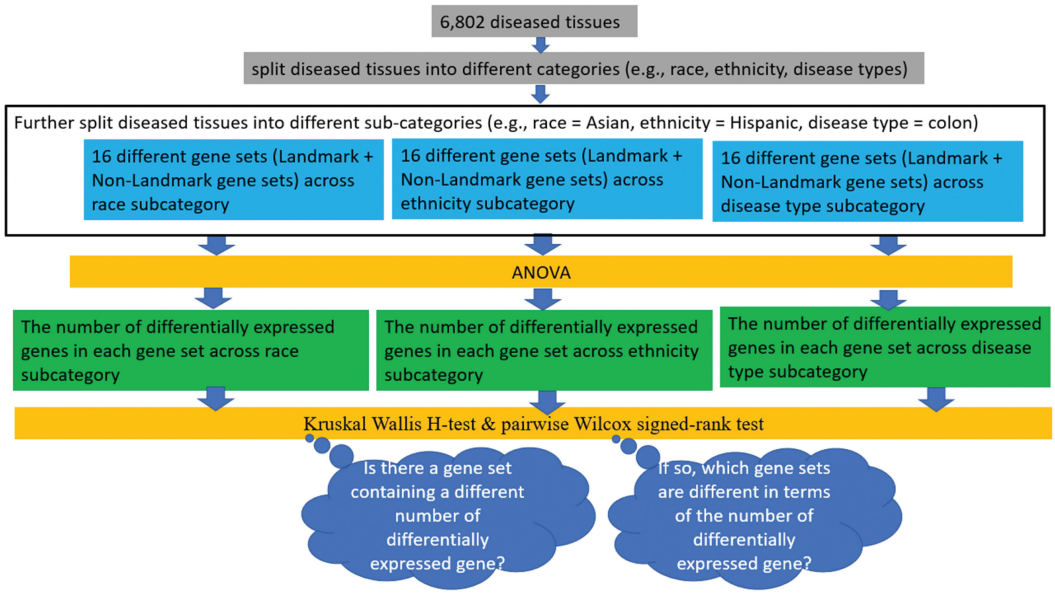


Figure 2: Design strategy for experiment B.

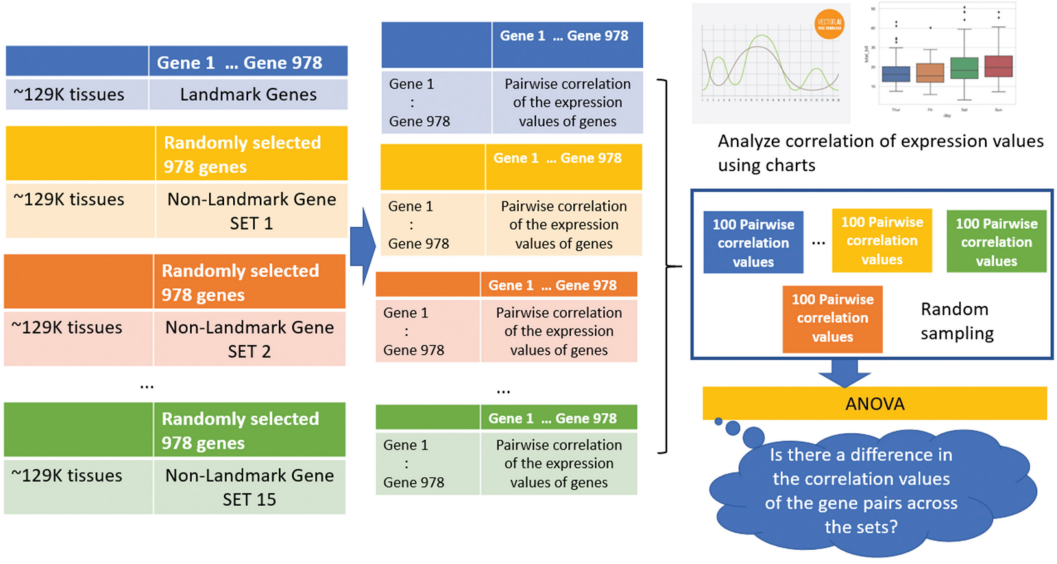


Figure 3: Design strategy for experiment C.

4. Results and Discussions

This study sought to address if there is any significant difference in the characteristics of the landmark and non-landmark genes. In an attempt to address the above-mentioned objectives, a total of three experiments were designed. Here, we present the results obtained from all three experiments and also discuss the inferences gathered from the results obtained.

To begin with, 16 different sets of 978 genes were obtained from dataset 1. One of the sets included the 978 landmark genes and the remaining 15 sets included the randomly chosen 978 genes out of the pool of non-landmark genes. When randomly choosing the genes for each set, it was ensured that none of the genes were duplicated within and across the sets (Figure 4). The tissue samples within dataset 1 were further portioned across race, ethnicity, and disease (cancer) types.

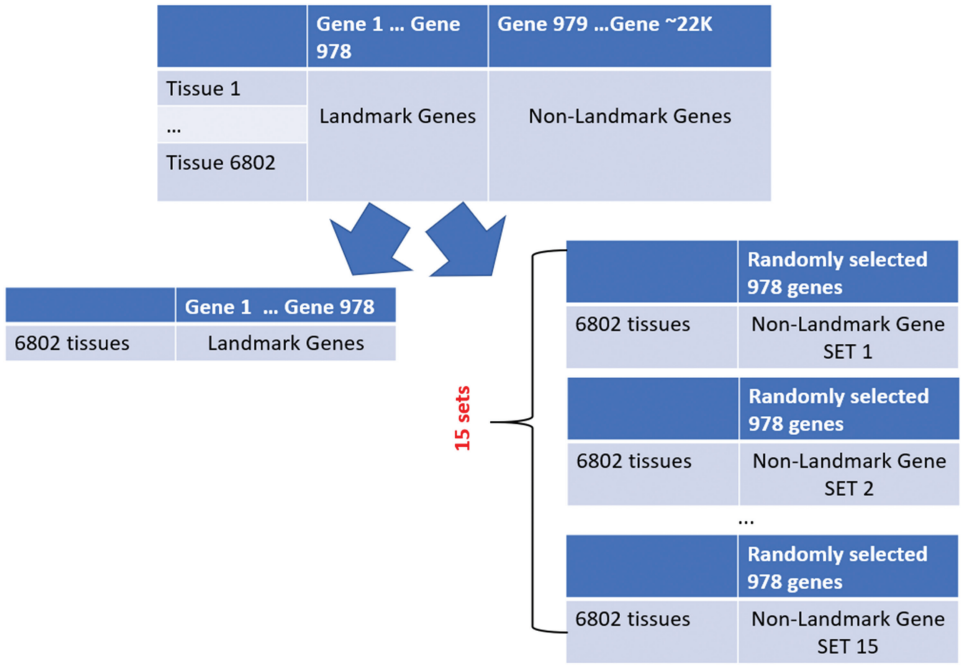


Figure 4: Sixteen sets of 978 genes used in experiments A–C.

Across the 16 sets of genes, ANOVA was performed to analyze the expression values of the genes across 6,802 diseased tissue samples of 10 cancer types. The objective was to determine how many genes in each of the sets were differentially expressed (refer to experiment A). Identifying the differentially expressed genes is critical because they are assumed to be the driving force and/or the molecular biomarkers of different phenotypes (Zhao et al. 2018). Within the landmark and the non-landmark sets of 978 genes, it is important to determine if there is a statistical difference in the number of differentially expressed genes. If there is a statistical difference in the number of differentially expressed genes between one or more sets, then it would indicate that the characteristics of one set is different from the other. The total number of genes, of the 978 genes, that were differentially expressed across the landmark set and the 15 different non-landmark sets in three different samples, namely, sample 1, sample 2, and sample 3, are highlighted in Table 2. The number of differentially expressed genes across each sample for both the landmark gene set and the non-landmark gene sets were obtained by considering different p-values for the ANOVA, that is, for sample 1, sample 2, and sample 3, the p-values were <0.05, <0.01, and <0.1, respectively.

In the landmark gene set, 99.9% of the genes were differentially expressed across the 6,802 diseased tissue samples. ANOVA resulted in a p-value of all the 977 genes to be < 0.1. However, the number of genes that were differentially expressed across the 15 sets of non-landmark genes ranged between 973 and 978 (Table 2). The Kruskal Wallis H test was performed to determine if there is a significant difference in the number of differentially expressed genes across the 16 gene sets. Here, three different samples, namely, sample 1 ($p < 0.05$), sample 2 ($p < 0.01$), and sample 3 ($p < 0.1$), were considered for the analysis (Table 2). Across the 16 different gene sets, no significant difference in the number of differentially expressed genes was observed at $\alpha = 0.05$. The Kruskal-Wallis H test resulted in a p-value of 0.1759. This implies that both the landmark gene set and the 15 different set of randomly chosen non-landmark gene sets, both have similar numbers of differentially expressed genes.

The total number of differentially expressed genes across the 16 gene sets (the landmark set and 15 non-landmark sets) at $p < 0.05$ for the different races, namely, Asian, White, and Black, are recorded in Table 3. Descriptive statistics across the races indicated that relatively more genes were differentially expressed in the White race (mean $\pm$ SD, 977 $\pm$ 1.61) than in the Black (mean $\pm$ SD, 715 $\pm$ 117) and Asian (mean $\pm$ SD, 656 $\pm$ 71.9) races.

The Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the three races. A significant difference was observed in the number of differentially expressed genes across the three races at $\alpha = 0.05$. The Kruskal-Wallis H test resulted in a p-value of 7.767e-08. In addition to that, a pairwise Wilcox signed-rank test was performed to determine which group of races differed from each other in terms of the number of differentially expressed genes. At $\alpha = 0.05$, a significant difference in the number of differentially expressed genes was observed for the White race, with the p-value of 1.9e-06 against the Asian and Black race (refer to experiment B).

Table 2: The number of genes that were differentially expressed across the landmark and non-landmark gene sets.

Gene Sets	Sample 1: $p < 0.05$	Sample 2: $p < 0.01$	Sample 3: $p < 0.1$
L	977	977	977
NL set 1	976	976	976
NL set 2	973	973	973
NL set 3	977	977	977
NL set 4	978	978	978
NL set 5	978	978	978
NL set 6	978	978	978
NL set 7	977	976	978
NL set 8	978	976	978
NL set 9	978	978	978
NL set 10	977	975	977
NL set 11	977	975	977
NL set 12	976	976	976
NL set 13	976	974	976
NL set 14	976	975	976
NL set 15	976	974	977

L, landmark genes; NL, non-landmark genes.

Table 3: The number of genes that were differentially expressed across the landmark and non-landmark gene sets for different races at $p < 0.05$.

	$p < 0.05$		
Gene Sets	Asian	White	Black
L	738	978	870
NL set 1	778	978	811
NL set 2	784	977	848
NL set 3	711	978	824
NL set 4	670	978	799
NL set 5	643	978	754
NL set 6	592	977	609
NL set 7	553	974	522
NL set 8	590	978	570
NL set 9	702	978	799
NL set 10	591	973	636
NL set 11	630	976	644
NL set 12	699	978	823
NL set 13	624	976	746
NL set 14	606	976	631
NL set 15	588	975	550

The total number of differentially expressed genes across the 16 gene sets (the landmark set and 15 non-landmark sets) at $p < 0.01$ for the different races, namely, Asian, White, and Black, is recorded in Table 4. Descriptive statistics across the races indicates that relatively more genes were differentially expressed in the White race (mean $\pm$ SD, 975 $\pm$ 2.73) than in the Black (mean $\pm$ SD, 603 $\pm$ 132) and Asian (mean $\pm$ SD, 515 $\pm$ 77.1) races.

The Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the three races. A significant difference was observed in the number of differentially expressed genes across the three races at $\alpha = 0.05$. The Kruskal-Wallis H test resulted in a p-value of 6.584e-08. In addition to that, a pairwise Wilcox signed-rank test was performed to determine which group of races differed from

Table 4: The number of genes that were differentially expressed across the landmark and non-landmark gene sets for different races at $p < 0.01$.

$p < 0.01$			
Gene Sets	Asian	White	Black
L	589	978	785
NL set 1	648	978	717
NL set 2	665	976	741
NL set 3	574	978	746
NL set 4	510	978	696
NL set 5	472	978	641
NL set 6	450	976	479
NL set 7	413	970	385
NL set 8	456	974	444
NL set 9	575	977	694
NL set 10	460	971	523
NL set 11	502	972	530
NL set 12	556	978	718
NL set 13	487	975	629
NL set 14	447	973	505
NL set 15	436	975	422

each other in terms of the number of differentially expressed genes. At $\alpha = 0.05$, a significant difference in the number of differentially expressed genes was observed for the White race, with the p-value of 2.1e-06 against the Asian and Black races (refer to experiment B).

The total number of differentially expressed genes across the 16 gene sets (the landmark set and 15 non-landmark sets) at $p < 0.1$ for the different races, namely, Asian, White, and Black, are recorded in Table 5. Descriptive

Table 5: The number of genes that were differentially expressed across the landmark and non-landmark gene sets for different races at $p < 0.1$.

$p < 0.1$			
Gene Sets	Asian	White	Black
L	799	978	912
NL set 1	835	978	850
NL set 2	840	978	881
NL set 3	779	978	863
NL set 4	743	978	838
NL set 5	713	978	801
NL set 6	659	978	663
NL set 7	642	975	581
NL set 8	665	978	631
NL set 9	761	978	840
NL set 10	674	976	693
NL set 11	700	976	709
NL set 12	764	978	863
NL set 13	712	977	797
NL set 14	676	976	697
NL set 15	670	977	620

statistics across the races indicated that relatively more genes were differentially expressed in the White race (mean ± SD, 977 ± 1.01) than in the Black (mean ± SD, 765 ± 107) and Asian (mean ± SD, 727 ± 63.6) races.

The Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the three races. A significant difference was observed in the number of differentially expressed genes across the three races at $\alpha = 0.05$. The Kruskal-Wallis H test resulted in a p-value of 9.975e-08. In addition to that, a pairwise Wilcox signed-rank test was performed to determine which group of races differed from each other in terms of the number of differentially expressed genes. At $\alpha = 0.05$, a significant difference in the number of differentially expressed genes was observed for the White race, with the p-value of 1.6e-06, against the Asian and Black races (refer to experiment B).

For the above observations, it is conclusive that the number of differentially expressed genes across the races was significantly different, at $p < 0.01$ (see Table 4), $p < 0.05$ (see Table 3), and at $p < 0.1$ (see Table 5). Within the three races, the number of differentially expressed genes was significantly different for White race when compared with the Asian and Black races, at $p < 0.01$ (see Table 4), $p < 0.05$ (see Table 3), and at $p < 0.1$ (see Table 5).

Finally, the Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the 16 gene sets across the three different races. The Kruskal-Wallis H test resulted in a p-value of 0.7449 for $p < 0.01$, which indicated that there was no significant difference in the number of differentially expressed genes across the 16 different gene sets, that is, irrespective of landmark or non-landmark genes, both types of the genes are similar across the races in terms of the constitution of the number of differentially expressed genes. Similarly, for the $p < 0.05$ and $p < 0.1$, the Kruskal-Wallis H test resulted in a p-value of >0.749, which indicated that, across the races, there was no significant difference in the number of differentially expressed genes across the 16 different gene sets (refer to experiment B).

The total number of differentially expressed genes across the 16 gene sets (the landmark set and 15 non-landmark sets) at $p < 0.05$ for the two ethnic groups, namely, Hispanic and non-Hispanic, is recorded in Table 6. Descriptive statistics across the ethnic groups indicate that relatively more genes are differentially expressed in the non-Hispanic group (mean ± SD, 977 ± 1.31) than in the Hispanic group (mean ± SD, 506 ± 106).

The Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the two ethnic groups. A significant difference was observed in the number of differentially expressed genes across the two ethnic groups at $\alpha = 0.05$. The Kruskal-Wallis H test resulted in a p-value of 1.207e-06. Thus, at $\alpha = 0.05$, a significant difference in the number of differentially expressed genes was observed between the Hispanic and non-Hispanic ethnic groups (refer to experiment B).

Table 6: The number of genes that were differentially expressed across the landmark and non-landmark gene sets for different ethnic groups at $p < 0.05$.

$p < 0.05$		
Gene Sets	Hispanic	Non-Hispanic
L	658	978
NL set 1	620	978
NL set 2	650	976
NL set 3	601	977
NL set 4	555	978
NL set 5	484	978
NL set 6	397	977
NL set 7	331	976
NL set 8	402	977
NL set 9	581	978
NL set 10	414	974
NL set 11	441	975
NL set 12	612	978
NL set 13	516	975
NL set 14	447	978
NL set 15	391	977

Finally, the Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the 16 gene sets across the two ethnic groups. The Kruskal-Wallis H test resulted in a p-value of 0.989 for $p < 0.05$, which indicated that there was no significant difference in the number of differentially expressed genes across the 16 different gene sets, that is, irrespective of landmark or non-landmark genes, both types of the genes are similar across ethnic groups in terms of the constitution of the number of differentially expressed genes (refer to experiment B).

The total numbers of differentially expressed genes across the 16 gene sets (the landmark set and 15 non-landmark sets) at $p < 0.01$ for the two ethnic groups, namely, Hispanic and non-Hispanic, are recorded in Table 7. Descriptive statistics across the ethnic groups indicated that relatively more genes were differentially expressed in the non-Hispanic group (mean ± SD, 976 ± 2.28) than in the Hispanic group (mean ± SD, 361 ± 97.8).

The Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the two ethnic groups. A significant difference was observed in the number of differentially expressed genes across the two ethnic groups at $\alpha = 0.05$. The Kruskal-Wallis H test resulted in a p-value of 1.289e-06. Thus, at $\alpha = 0.05$, a significant difference in the number of differentially expressed genes was observed between the Hispanic and non-Hispanic ethnic groups (refer to experiment B).

Finally, the Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the 16 gene sets across the two ethnic groups. The Kruskal-Wallis H test resulted in a p-value of 0.979 for $p < 0.01$, which indicated that there was no significant difference in the number of differentially expressed genes across the 16 different gene sets, that is, irrespective of landmark or non-landmark genes, both types of the genes are similar across ethnic groups in terms of the constitution of the number of differentially expressed genes (refer to experiment B).

The total number of differentially expressed genes across the 16 gene sets (the landmark set and 15 non-landmark sets) at $p < 0.1$ for the two ethnic groups, namely, Hispanic and non-Hispanic, are recorded in Table 8. Descriptive statistics across the ethnic groups indicated that relatively more genes were differentially expressed in the non-Hispanic group (mean ± SD, 977 ± 0.931) than in the Hispanic group (mean ± SD, 590 ± 107).

The Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the two ethnic groups. A significant difference was observed in the number of differentially expressed genes across the two ethnic groups at $\alpha = 0.05$. The Kruskal-Wallis nonparametric test resulted in a p-value of 1.109e-06. Thus, at $\alpha = 0.05$, a significant difference in the number of differentially expressed genes was observed between the Hispanic and non-Hispanic ethnic groups (refer to experiment B).

Table 7: The number of genes that were differentially expressed across the landmark and non-landmark gene sets for different ethnic groups at $p < 0.01$.

$p < 0.01$		
Gene Sets	Hispanic	Non-Hispanic
L	503	978
NL set 1	479	978
NL set 2	501	976
NL set 3	434	977
NL set 4	419	978
NL set 5	297	978
NL set 6	268	976
NL set 7	214	974
NL set 8	273	977
NL set 9	432	978
NL set 10	269	971
NL set 11	278	974
NL set 12	454	978
NL set 13	368	973
NL set 14	310	975
NL set 15	275	973

Table 8: The number of genes that were differentially expressed across the landmark and non-landmark gene sets for different ethnic groups at $p < 0.1$

	$p < 0.1$	
Gene Sets	Hispanic	Non-Hispanic
L	752	978
NL set 1	703	978
NL set 2	720	977
NL set 3	682	977
NL set 4	638	978
NL set 5	582	978
NL set 6	471	978
NL set 7	410	976
NL set 8	477	977
NL set 9	661	978
NL set 10	503	976
NL set 11	530	977
NL set 12	700	978
NL set 13	605	975
NL set 14	533	978
NL set 15	476	977

Finally, the Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the 16 gene sets across the two ethnic groups. The Kruskal-Wallis H test resulted in a p-value of 0.992 for $p < 0.1$, which indicated that there was no significant difference in the number of differentially expressed genes across the 16 different gene sets, that is, irrespective of landmark or non-landmark genes, both types of the genes are similar across ethnic groups in terms of the constitution of the number of differentially expressed genes (refer to experiment B).

The total number of differentially expressed genes across the 16 gene sets (the landmark set and 15 non-landmark sets) at $p < 0.05$ for the 10 disease (cancer) types, namely, colon, brain, bladder, skin, breast, kidney, prostate, stomach, testis, and pancreas, are recorded in Table 9. Descriptive statistics across the disease type are recorded in Table 10.

The Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the 10 disease types. A significant difference was observed in the number of differentially expressed genes across the 10 disease types at $\alpha = 0.05$. The Kruskal-Wallis H test resulted in a p-value of <2.2e-16. Thus, at $\alpha = 0.05$, a significant difference in the number of differentially expressed genes was observed between the 10 different disease types (refer to experiment B).

Finally, the Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the 16 gene sets across the 10 disease types. The Kruskal-Wallis H test resulted in a p-value of 0.999 for $p < 0.1$, which indicated that there was no significant difference in the number of differentially expressed genes across the 16 different gene sets, that is, irrespective of landmark or non-landmark genes, both types of the genes were similar across the different disease types in terms of the constitution of the number of differentially expressed genes (refer to experiment B).

A correlation study was performed to differentiate the characteristics of the landmark and the non-landmark genes in the L1000 dataset (dataset 3). The pairwise correlation of the expression values of the 978 landmark genes across ~129,000 tissue samples were compared against the expression values of the 978 non-landmark genes across 15 different randomly selected set of non-landmark genes across the ~129,000 tissue samples. The results of the correlation study are demonstrated in Figure 5. The blue-colored line represents the range of the correlation values between a pair of genes in the landmark set. The remaining colored lines represent the range of the correlation values between the pair of genes in the different non-landmark gene sets. The range of the correlation values of the gene pairs within the landmark set were between [–0.8, –0.4] and [0.4, 0.8]. However, for the other sets of non-landmark genes, the range of correlation values between the pair of genes were almost similar, without any distinctive patterns (Figure 5).

Table 9: The number of genes that were differentially expressed across the landmark and non-landmark gene sets for different disease types at $p < 0.05$.

Gene Set	Colon	Brain	Bladder	Skin	Breast	Kidney	Prostate	Stomach	Testis	Pancreas
L	79	8	130	6	4	53	13	86	1	1
NL set 1	99	12	102	4	3	52	10	88	3	2
NL set 2	73	7	112	7	7	72	13	110	3	1
NL set 3	86	17	109	5	7	66	14	106	10	4
NL set 4	92	19	122	8	7	41	40	95	9	3
NL set 5	80	23	109	16	17	63	27	95	15	3
NL set 6	91	14	85	16	11	48	20	67	29	5
NL set 7	96	29	73	24	18	37	18	48	24	4
NL set 8	79	19	70	22	10	40	7	63	28	1
NL set 9	103	7	100	10	8	58	25	91	15	0
NL set 10	94	16	60	17	6	54	18	69	25	3
NL set 11	110	28	77	18	9	54	16	78	18	3
NL set 12	102	11	106	9	1	70	15	99	10	2
NL set 13	102	16	78	11	8	42	22	62	7	4
NL set 14	91	20	71	16	9	45	21	74	24	4
NL set 15	96	19	73	16	6	54	16	47	19	4

Table 10: Descriptive statistics for different disease types.

Disease Type	Mean ± Standard Deviation
Colon	92.1 ± 10.4
Brain	16.6 ± 6.69
Bladder	92.3 ± 21.3
Skin	12.8 ± 6.12
Breast	8.19 ± 8.19
Kidney	53.1 ± 10.7
Prostate gland	18.4 ± 7.78
Stomach	79.9 ± 19.4
Testis	15 ± 9.26
Pancreas	2.75 ± 1.44

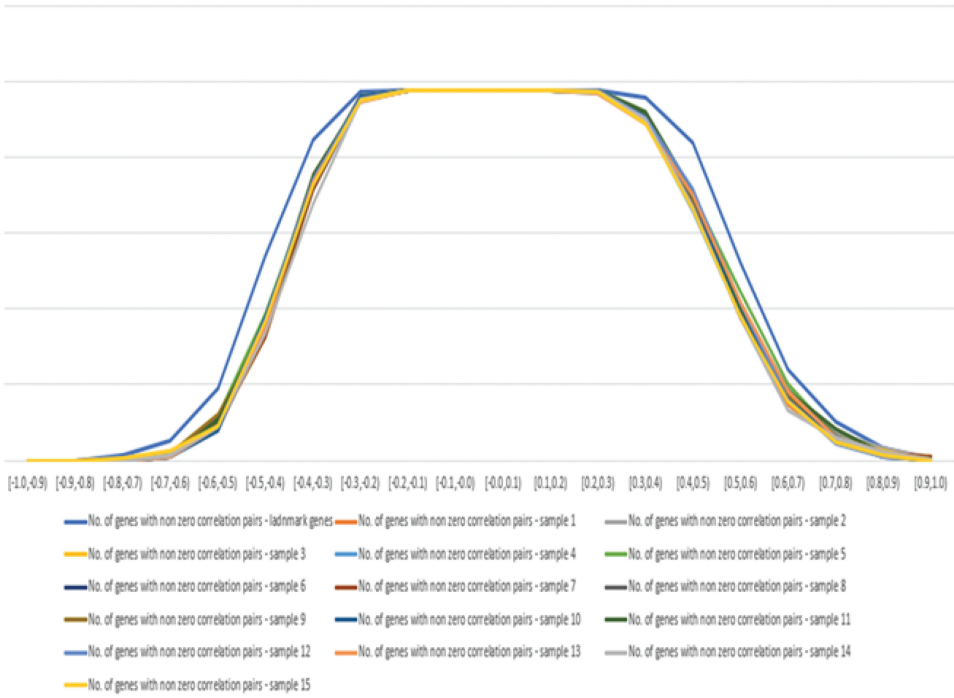


Figure 5: Pairwise correlation of genes in the landmark and non-landmark gene sets.

One-way ANOVA was performed on a dataset that contained randomly selected 100 correlation values between the pair of genes from both the landmark gene set and the 15 non-landmark gene sets. At $\alpha = 0.05$, one-way ANOVA resulted in a p-value of 0.999, which suggests not to reject the null hypothesis and conclude that there was no significant difference in the correlation values of the gene pairs across the 16 gene sets, that is, there was no evidence that the correlation values of the gene pairs in both the landmark set and the non-landmark sets are any different (refer to experiment C).

The boxplot of the correlation values of the gene pairs across the 16 gene sets are shown in Figure 6. The red-colored boxplot represents the landmark gene set, and the remaining colored boxplots represent the different non-landmark gene sets. The boxplots of the different gene set clearly highlights a slight variation in the correlation values of the gene pairs. However, there are no definitive patterns to clearly differentiate the correlation values of the gene pairs in both the landmark and non-landmark gene sets.

Based on all the experiments conducted so far, there were no observations that support the fact that landmark and non-landmark genes are different from each other.

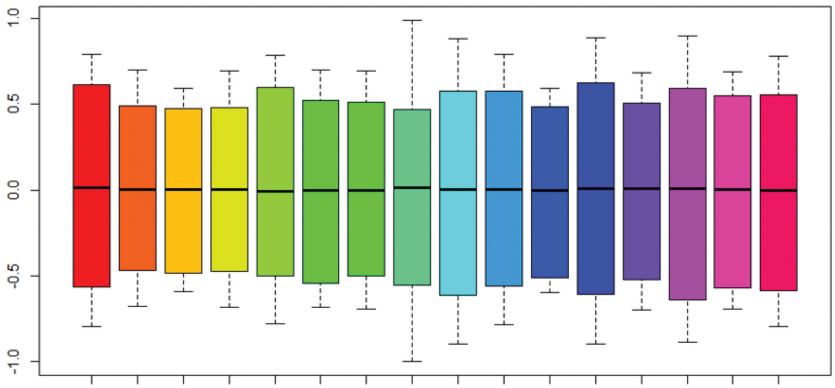


Figure 6: A boxplot of the correlation values of the gene pairs across the 16 gene sets.

5. Conclusion and Future Direction

This study aimed at understanding if there is any significant difference in the characteristics of the landmark and the non-landmark genes. Studies in the past (Duncan et al. 2008; Chen et al. 2016; Danaee et al. 2017; Ramaker et al. 2017; Bailey et al. 2018; Chen et al. 2018; Huang et al. 2018; Way et al. 2018; Daoud and Mayo, 2019; Clayman et al. 2020a) only focused on using the expression values of the landmark genes to determine the expression values of the non-landmark genes but did not discuss whether the landmark genes are any different from the non-landmark genes, that is, could we find a different set of non-landmark genes and say that they are similar in characteristics to the original set of identified landmark genes. The two experiments, namely the experiment A (see Figure 1) and experiment B (see Figure 2), indicated that there is no significant difference in the characteristics of the landmark and non-landmark genes. Across the landmark gene set and the 15 different randomly chosen non-landmark gene sets of similar size, no significant difference was observed in the number of differentially expressed genes across race, ethnicity, and disease types. On analyzing the correlation of the gene pairs within the landmark gene set and the 15 different randomly chosen non-landmark gene sets of similar size, it was observed that landmark gene pairs had slightly more range of correlation values compared with the other 15 sets of non-landmark gene pairs. However, the statistical test concluded that there was no evidence that the correlation values of the gene pairs in both the landmark gene set, and the non-landmark gene sets were any different (refer to experimental design C).

In this study, we only considered 16 sets of 978 genes, that is, one set of landmark genes identified in the work by Chen et al. (2016) and Clayman et al. (2020a, 2020b), and the 15 sets of randomly chosen genes labeled as non-landmark genes. The 15 sets, each contained 978 genes, were randomly chosen of the remaining $\sim$21,000 genes. Choosing a set of random 978 genes without repetition (non-landmark) of the remaining $\sim$21,000 genes is a complex combinatorial problem, and comparing all of those combinations against the established set of landmark genes is computationally intensive. Therefore, we chose a sample set of 15 equally sized non-landmark gene sets and used the non-parametric statistical tests to determine if there was any significant difference in the characteristics of the landmark and the non-landmark genes.

This study made a significant contribution to the field of personalized medicine. This field strives on the objective of determining the contributing genetic variables or the genes that are clinically relevant biomarkers for diseases. Large-scale availability of the gene expression profiling and clinical data related to carcinomas diseases provided us with the opportunity to explore and identify the significant variables (clinical and genetic) that could clearly characterize one or more diseases. At the same time, the advances in ML and AI techniques have made it possible to model the clinical and genetic variables to understand the relationships between these variables and the clinical outcomes. Even though both clinical and genetic variables are important to understand the clinical outcomes, the goal was to focus on identifying the genetic variables that can serve as clinically relevant biomarkers. This is because the clinical variables are mainly associated with the manifestations of the disease; that is, they are highly corelated with the disease types, and, also, it is a measurement that is captured after the fact. However, the genetic characteristics of an individual organism in a species or population, that is, genetic predisposition has a direct influence on disease development under the influence of environmental conditions.

Future studies can assess the potentiality of the linear combination or the principal components of the landmark and non-landmark gene clusters for disease-type predictions. This will be performed by implementing statistical, data mining, and ML techniques to extract patterns from the data as well as building predictive models. Effects of gene-

gene interactions for various types of cancer diseases should also be further assessed by using survival analysis, given that gene interactions are predictive of clinical outcomes. Various cancer types should be assessed to determine genes relevant to specific disease or cancer types.

Future studies may also build on this analysis by using predictive analytics techniques to further develop the understanding of how to investigate the relationship between genetic and clinical variables by accounting for both coding and noncoding genetic variants (Quang et al. 2015). This may be especially relevant toward applications for personalized medicine such as treatment responsiveness, depending on the combination of genetic and clinical variables. Future studies can assess whether clustering results based on gene expression levels can predict various disease types.

References

Ash, J. T., G. Darnell, D. Munro, and B. E. Engelhardt. 2018. "Joint Analysis of Gene Expression Levels and Histological Images Identifies Genes Associated with Tissue Morphology." Accessed March 12, 2025. https://doi.org/10.1101/458711

Azarkhalili, B., A. Saberi, H. Chitsaz, and A. Sharifi-Zarchi. 2018. "DeePathology: Deep Multi-Task Learning for Inferring Molecular Pathology from Cancer Transcriptome." Preprint, submitted August 2019. http://arxiv.org/abs/1808.02237

Bailey, M. H., C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, et al. 2018. "Comprehensive Characterization of Cancer Driver Genes and Mutations." *Cell* **173**, no. 2: 371–385. doi: 10.1016/j.cell.2018.02.060

Castro, E., C. Goh, D. Olmos, E. Saunders, D. Leongamornlert, M. Tymrakiewicz, et al. 2013. "Germline *BRCA* Mutations Are Associated with Higher Risk of Nodal Involvement, Distant Metastasis, and Poor Survival Outcomes in Prostate Cancer." *J Clin Oncol* **31**, no. 14: 1748–1757. doi: 10.1200/JCO.2012.43.188

Chaudhary, K., O. B. Poirion, L. Lu, and L. X. Garmire. 2018. "Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer." *Clinical Cancer Research* **24**, no. 6: 1248–1259. doi: 10.1158/1078-0432.CCR-17-0853

Chen, X., J. Xie, and Q. Yuan. 2018. "A Method to Facilitate Cancer Detection and Type Classification from Gene Expression Data using a Deep Autoencoder and Neural Network." Accessed March 12, 2025. *ArXiv* 1812.08674 [Cs, Stat]. https://api.semanticscholar.org/CorpusID:56517070

Chen, Y., Y. Li, R. Narayan, A. Subramanian, and X. Xie. 2016. "Gene Expression Inference with Deep Learning." *Bioinformatics* **32**, no. 12: 1832–1839. doi: 10.1093/bioinformatics/btw074

Clayman, C. L., S. M. Srinivasan, R. S. Sangwan. 2020a. "Cancer Survival Analysis Using RNA Sequencing and Clinical Data." *Procedia Computer Science* **168**: 80–87. doi: 10.1016/j.procs.2020.02.261

Clayman, C. L., S. M. Srinivasan, R. S. Sangwan. 2020b. "K-Means Clustering and Principal Components Analysis of Microarray Data of L1000 Landmark Genes." *Procedia Computer Science* **168**: 97–104. doi: 10.1016/j.procs.2020.02.265

Danaee, P., R. Ghaeini, and D. A. Hendrix. 2017. "A Deep Learning Approach for Cancer Detection and Relevant Gene Identification." *Pacific Symposium on Biocomputing 2017*. Accessed March 12, 2025. https://doi.org/10.1142/9789813207813_0022

Daoud, M., and M. Mayo. 2019. "A Survey of Neural Network-based Cancer Prediction Models from Microarray Data." *Artificial Intelligence in Medicine* **92**: 204–214. doi: 10.1016/j.artmed.2019.01.006

Duan, Q., S. P. Reid, N. R. Clark, Z. Wang, N. F. Fernandez, A. D. Rouillard, et al. 2016. "L1000CDS^2: LINCS L1000 Characteristic Direction Signatures Search Engine." *NPJ Systems Biology and Applications* **2**: 16015. doi: 10.1038/npjsba.2016.15

Duncan, R., B. Carpenter, L. C. Main, C. Telfer, and G. I. Murray. 2008. "Characterisation and Protein Expression Profiling of Annexins in Colorectal Cancer." *British Iournal of Cancer* **98**, no. 2: 426–433. doi: 10.1038/sj.bjc.6604128

Dutil, F., J. P. Cohen, M. Weiss, G. Derevyanko, and Y. Bengio. 2018. "Towards Gene Expression Convolutions using Gene Interaction Graphs." International Conference on Machine Learning Workshop on Computational Biology, 2018. Preprint, submitted Jun 18. Accessed March 12, 2025. http://arxiv.org/abs/1806.06975

Ecke T. H., H. H. Schlechte, K. Schiemenz, M. D. Sachs, S. V. Lenk, B. D. Rudolph, et al. 2010. "TP53 Gene Mutations in Prostate Cancer Progression." *Anticancer Research* **30**, no. 5: 1579–1586.

Fielden, M. R., and T. R. Zacharewski. 2001. "Challenges and Limitations of Gene Expression Profiling in Mechanistic and Predictive Toxicology." *Toxicological Sciences* **60**, no. 1: 6–10. doi: 10.1093/toxsci/60.1.6

Huang, S., N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu. 2018. "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics." *Cancer Genomics & Proteomics* **15**, no. 1: 41–51. doi: 10.21873/cgp.20063

Hurd, P. J., and C. J. Nelson. 2009. "Advantages of Next-Generation Sequencing Versus the Microarray in Epigenetic Research." *Briefings in Functional Genomic & Proteomics* **8**, no. 3: 174–183. doi: 10.1093/bfgp/elp013

Koch, C. M., S. F. Chiu, M. Akbarpour, A. Bharat, K. M. Ridge, E. T. Bartom, et al. 2018. "A Beginner's Guide to Analysis of RNA Sequencing Data." *American Journal of Respiratory Cell and Molecular Biology* **59**, no. 2: 145–157. doi: 10.1165/rcmb.2017-0430TR

Kogelman, L. J. A., and H. N. Kadarmideen. 2014. "Weighted Interaction SNP Hub (WISH) Network Method for Building Genetic Networks for Complex Diseases and Traits Using Whole Genome Genotype Data." *BMC Systems Biology* **8**, no. Suppl 2: S5. doi: 10.1186/1752-0509-8-S2-S5

Kourou, K., T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. 2015. "Machine Learning Applications in Cancer Prognosis and Prediction." *Computational and Structural Biotechnology Journal* **13**: 8–17. doi: 10.1016/j.csbj.2014.11.005

Kursa, M. B., and W. R. Rudnicki. 2010. "Feature Selection with the Boruta Package." *Journal of Statistical Software* **36**, no. 11: 1–13. doi: 10.18637/jss.v036.i11

Liang, M., Z. Li, T. Chen, and J. Zeng. 2015. "Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **12**, no. 4: 928–937. doi: 10.1109/TCBB.2014.2377729

Lim, S., S. Lee, I. Jung, S. Rhee, and S. Kim. 2020. "Comprehensive and Critical Evaluation of Individualized Pathway Activity Measurement Tools on Pan-Cancer Data." *Briefings in Bioinformatics* **21**, no. 1, 36–46. doi: 10.1093/bib/bby097

Lin, M., V. Jaitly, I. Wang, Z. Hu, L. Chen, M. A. Wahed, et al. 2018. "Application of Deep Learning on Predicting Prognosis of Acute Myeloid Leukemia with Cytogenetics, Age, and Mutations." Preprint, submitted Oct 30. Accessed March 12, 2025. https://arxiv.org/abs/1810.13247

Love, M. I., W. Huber, and S. Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2." *Genome Biology* **15**, no. 550: 1–21. doi: 10.1186/s13059-014-0550-8

Malta, T. M., A. Sokolov, A. J. Gentles, T. Burzykowski, L. Poisson, J. N. Weinstein, et al. 2018. "Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation." *Cell* **173**, no. 2: 338–354. doi: 10.1016/j.cell.2018.03.034

Nahm, F. S. 2016. "Nonparametric Statistical Tests for the Continuous Data: The Basic Concept and the Practical Use." *Korean Journal of Anesthesiology* **69**, no. 1: 8–14. doi: 10.4097/kjae.2016.69.1.8

Parikh, N., S. Hilsenbeck, C. J. Creighton, T. Dayaram, R. Shuck, E. Shinbrot, et al. 2014. "Effects of *TP53* Mutational Status on Gene Expression Patterns Across 10 Human Cancer Types." *The Journal of Pathology* **232**, no. 5: 522–533. doi: 10.1002/path.4321

Petralia, F., W.-M. Song, Z. Tu, and P. Wang. 2016. "New Method for Joint Network Analysis Reveals Common and Different Coexpression Patterns among Genes and Proteins in Breast Cancer." *Journal of Proteome Research* **15**, no. 3: 743–754. doi: 10.1021/acs.jproteome.5b00925

Quang, D., Y. Chen, and X. Xie. 2015. "DANN: A Deep Learning Approach for Annotating the Pathogenicity of Genetic Variants." *Bioinformatics* **31**, no. 5: 761–763. doi: 10.1093/bioinformatics/btu703

Ramaker, R. C., B. N. Lasseigne, A. A. Hardigan, L. Palacio, D. S. Gunther, R. M. Myers, et al. 2017. "RNA Sequencing-Based Cell Proliferation Analysis Across 19 Cancers Identifies a Subset of Proliferation-Informative Cancers with a Common Survival Signature." *Oncotarget* **8**, no. 24: 38668–38681. doi: 10.18632/oncotarget.16961

Saltz, J., R. Gupta, L. Hou, T. Kurc, P. Singh, V. Nguyen, et al. 2018. "Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images." *Cell Reports*, **23**, no. 1: 181–193. doi: 10.1016/j.celrep.2018.03.086

Sawyer, S. F. 2009. "Analysis of Variance: The Fundamental Concepts." *Journal of Manual & Manpulative Therapy* **17**, no. 2: 27–38. doi: 10.1179/jmt.2009.17.2.27E

Tomczak, K., P. Czerwińska, and M. Wiznerowicz. 2015. "The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge." *Contemporary Oncology (Pozn)* **19**, no. 1A: A68–A77. doi: 10.5114/wo.2014.47136

Tsagri, M., Z. Papadovasilakis, K. Lakiotaki, and I. Tsamardinos. 2018. "Efficient Feature Selection on Gene Expression Data: Which Algorithm To Use? Accessed March 12, 2025. https://www.biorxiv.org/content/10.1101/431734v1

Way, G. P., F. Sanchez-Vega, K. La, J. Armenia, W. K. Chatila, A. Luna, et al. 2018. "Machine Learning Detects Pan-Cancer Ras Pathway Activation in the Cancer Genome Atlas." *Cell Reports* **23**, no. 1: 172–180. doi: 10.1016/j.celrep.2018.03.046

Way, G. P., M. Zietz, V. Rubinetti, D. S. Himmelstein, and C. S. Greene. 2019. "Sequential Compression of Gene Expression Across Dimensionalities and Methods Reveals no Single Best Method or Dimensionality." Accessed March 12, 2025. https://www.biorxiv.org/content/10.1101/573782v2.full.pdf+html

Zhang, W., Y. Yu, F. Hertwig, J. Thierry-Mieg, W. Zhang, D. Thierry-Mieg, et al. 2015. "Comparison of RNA-Seq and Microarray-Based Models for Clinical Endpoint Prediction." *Genome Biology* **16**: 133. doi: 10.1186/s13059-015-0694-1

Zhao, B., A. Erwin, and B. Xue. 2018. "How Many Differentially Expressed Genes: A Perspective from the Comparison of Genotypic and Phenotypic Distances." *Genomics* **110**, no. 1: 67–73. doi: 10.1016/j.ygeno.2017.08.007

Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 3, No. 1, 2025 DOI: 10.54116/jbdai.v3i1.45

A Holistic Approach to Subject Correlation Analysis in Secondary Education

Buddhi Ayesha
Rowan University
rathna55@rowan.edu

Adessha Jayasooriya
University of Moratuwa
adeesha.14@cse.mrt.ac.lk

Wishmitha Mendis
University of Moratuwa
wishmitha.14@cse.mrt.ac.lk

Bhanuka Mahanama
University of Moratuwa
bhanuka.14@cse.mrt.ac.lk

Malaka Dayasiri
University of Moratuwa
malaka.14@cse.mrt.ac.lk

Umashanger Thayasivam
Rowan University
thayasivam@rowan.edu

Uthayasanker Thayasivam
University of Moratuwa
rtuthaya@cse.mrt.ac.lk

ABSTRACT

This study presents a holistic investigation of subject correlations in secondary education by drawing on performance data from more than 600 students across grades 6 to 8 in Sri Lanka. By using correlation analysis, regression models, factor analysis, and hierarchical clustering, we reveal key interrelationships among core subjects, such as mathematics, science, and language studies, alongside broader disciplines, such as citizenship education and art. Our results confirm the robust influence of reading proficiency on science achievement, outpacing the traditionally studied mathematics-science link, and underscores the value of language skills in mastering diverse subjects. Factor analysis identifies a dominant general academic ability that spans multiple areas, particularly language and humanities, whereas clustering underscores that some subjects, such as art and practical and technical skills, cluster distinctly. These findings advocate for interdisciplinary teaching methods and targeted interventions, shedding light on students' varied learning trajectories and informing policy to enhance overall educational outcomes.

Keywords: *Educational data mining, subject correlation, holistic analysis, factor analysis, hierarchical clustering.*

1. Introduction

Exploring associations among academic subjects is a significant area of research in educational data mining, with correlation analysis widely used to investigate these relationships. This study presents a holistic approach to identifying subject correlations at the secondary school level, which encompasses all significant aspects of the educational experience (Mahmoudi et al. 2012). By analyzing correlations across all subjects, we aim to improve the learning experience for middle school students and inform educational policy changes.

One of the main challenges of a holistic approach is the high dimensionality due to the increased number of subjects, which requires a large sample size for accurate clustering and analysis. Although significant work has been done on subject-level correlations (Wang 2005), most studies focused on specifically selected subjects, potentially overlooking influences from other areas. Because an academic term can include many subjects, disregarding some subjects may lead to incomplete or biased results. Therefore, a holistic analysis is necessary to accurately reflect the true nature of subject interrelationships.

This paper presents a framework and methodology for analyzing correlations among different subjects by using a holistic approach. By using advanced data mining techniques, we identify patterns and correlations that may be missed in studies confined to part of the syllabus (Wang 2005; O'Reilly and McNamara 2007; Maerten-Rivera et al. 2010). Recent studies emphasize the effectiveness of holistic and interdisciplinary methods in educational settings, such as advanced machine learning algorithms for predictive analysis (Yağcı 2022), holistic educational frameworks (Miseliunaite et al. 2022), and the integration of science, technology, engineering, arts, and mathematics (STEAM) into education (Marín-Marín et al. 2021). These findings affirm the need for a holistic perspective in educational research to enhance learning outcomes across various environments.

This paper is organized as follows: Section 2 provides a comprehensive review of related studies on correlation analysis in education. Section 3 details the methodology adopted for this research, including data collection, preprocessing, and analytical techniques. Section 4 presents the empirical findings and offers a critical discussion of the results. Section 5 draws the conclusions of the study and highlights potential directions for future research.

2. Literature Review

The exploration of subject correlations in education has been a focal point for researchers aiming to enhance student performance and inform educational policies. However, traditional studies often focus on specific subject pairs or small sample sizes, potentially overlooking the complex interrelations within a broader curriculum. This literature review critically examines existing research on subject correlation analysis, highlights the methodologies used, and identifies gaps that the current study aims to address.

2.1. Subject Correlations and Academic Performance

Understanding how proficiency in one subject area may influence or predict performance in another is a key concern in educational research. For example, Ünal et al. (2023) conducted a meta-analysis that explored the sources of correlation between reading and mathematics achievement. They concluded that both domain-general cognitive abilities and domain-specific skills contribute to the observed correlations, highlighting the multifaceted nature of academic performance. Similarly, Jindra et al. (2022) provided observational evidence of the reciprocal relationship between reading and mathematics proficiency over time.

O'Reilly and McNamara (2007) conducted a longitudinal study that investigated how reading comprehension skills impact science achievement. By using hierarchical linear modeling, they found that reading skills significantly predict science performance, especially in understanding complex scientific texts. Cruz Neri et al. (2021) also emphasized that language proficiency is crucial for students to articulate scientific concepts effectively, which suggests that interventions to improve language skills could positively affect science achievement.

The relationship between mathematics and science has also been extensively studied. Wang (2005) analyzed data from eighth-grade students by using structural equation modeling and found a bi-directional relationship between mathematics and science achievement. Strong mathematical skills provide a foundation for understanding scientific concepts, whereas engagement in science reinforces mathematical understanding. However, these studies often focus on specific grades or education levels, which may limit the generalizability of their findings.

2.2. Methodological Approaches in Subject Correlation Studies

Methodologies in subject correlation research have evolved to incorporate more sophisticated statistical and data mining techniques. Traditional methods such as Pearson and Spearman correlation coefficients are commonly used

for initial exploratory analyses (Barnard-Brak et al. 2017). Although effective for measuring linear and monotonic relationships, these methods may not capture the complexity inherent in educational data, which often involve high-dimensional and non-linear interactions.

Advanced statistical methods, such as factor analysis and structural equation modeling, have been used to uncover latent variables that influence multiple subjects simultaneously. For instance, Barnard-Brak et al. (2017) used confirmatory factor analysis to examine underlying constructs that affect reading and mathematics proficiency, revealing significant roles of cognitive and non-cognitive factors.

Machine learning techniques are increasingly adopted in educational data mining to predict student performance and identify patterns in large datasets. Yağcı (2022) applied ensemble learning methods to predict academic success, demonstrating that algorithms such as random forests and gradient boosting outperform traditional regression models in handling complex, non-linear relationships.

Clustering algorithms, such as hierarchical clustering and k-means, have been used to group students or subjects based on performance metrics. Mahanama et al. (2018) used hierarchical clustering to identify subject groupings, which provides insights into curriculum development and personalized learning strategies.

Despite these advancements, many studies remain limited in scope, often focusing on specific subjects or small, homogeneous samples. External factors such as socioeconomic status, parental education, and learning styles are frequently overlooked, potentially confounding the relationships among subjects (Beylik and Genç Kumtepe 2021).

2.3. Holistic Educational Frameworks and Interdisciplinary Approaches

The shift toward holistic education emphasizes integrating personal, social, and academic development. Mahmoudi et al. (2012) argue that holistic approaches consider the whole learner by promoting interconnected learning experiences across disciplines. This perspective aligns with the integration of STEAM education, which advocates for interdisciplinary learning to foster creativity and critical thinking (Marín-Marín et al. 2021).

Recent studies support the effectiveness of holistic and interdisciplinary methods. Miseliunaite et al. (2022a) conducted a systematic literature review and found that holistic education frameworks contribute to improved student engagement and motivation. The inclusion of arts in science, technology, engineering, and mathematics (STEM) education (forming STEAM) has been shown to enhance problem-solving skills and innovation (Yakman and Lee 2012).

However, implementing holistic education faces challenges, such as curriculum rigidity and assessment practices that favor subject-specific achievements. Miseliunaite et al. (2022) highlight the need for systemic changes to fully realize the benefits of holistic educational approaches.

2.4. Critical Analysis of Existing Research

Although significant progress has been made, several gaps persist in the literature:

Limited Scope of Studies: Many studies focus narrowly on specific subject pairs or education levels, which limit the applicability of findings across different contexts. This narrow focus may lead to incomplete understandings of how various subjects interrelate within a comprehensive curriculum.

Methodological Constraints: Traditional statistical methods may not adequately capture the complex, non-linear relationships in educational data. There is a need for more sophisticated analytical techniques that can handle high-dimensional data and uncover deeper insights.

Underrepresentation of External Factors: Socioeconomic status, parental education, and learning styles are often underrepresented in analyses, despite their significant impact on student performance. Ignoring these factors can result in biased findings and limit the effectiveness of proposed educational interventions.

Geographic and Cultural Limitations: The majority of research is concentrated in Western countries, which may not account for cultural and educational differences in other regions. This limits the generalizability of findings and the development of globally applicable educational strategies.

2.5. Contribution of the Current Study

The present study addresses the aforementioned gaps by adopting a holistic approach to analyzing subject correlations in secondary education. Key contributions include the following:

Comprehensive Subject Analysis: By analyzing correlations across all subjects within the curriculum, the study provides a more complete picture of subject interrelationships, identifying overarching patterns and highly correlated subject categories.

Advanced Methodological Framework: Using a combination of correlation analysis, regression, factor analysis, and hierarchical clustering allows for a nuanced examination of complex relationships in the data. This methodological rigor enhances the reliability of the findings.

Inclusion of External Factors: The study incorporates variables such as socioeconomic status, parental education, and learning styles, providing a more comprehensive understanding of factors that influence student performance.

Diverse Sample and Context: By collecting data from more than 600 students across urban, suburban, and rural regions in Sri Lanka, the study adds valuable insights from a non-Western context, which contributes to the global discourse on educational data mining.

2.6. Relevance to Educational Policy and Practice

The findings of this study have practical implications for educators and policymakers. By identifying significant patterns of subject correlations, the research can inform curriculum development, teaching strategies, and resource allocation. Emphasizing a holistic approach aligns with contemporary educational goals of fostering well-rounded learners equipped with interdisciplinary skills.

In summary, although existing research has laid the groundwork for understanding subject correlations in education, there is a clear need for more holistic, methodologically robust studies that consider a wider range of subjects and external factors. The current study sought to fill this gap by offering a comprehensive analysis that can contribute to improved educational strategies and student outcomes.

3. Methodology

This study used a comprehensive methodological framework designed to holistically analyze subject correlations in secondary education. The methodology encompassed detailed data collection procedures; meticulous data preprocessing; and the application of various advanced analytical techniques, including correlation analysis, regression analysis, factor analysis, and hierarchical clustering. Each component was elaborated to ensure clarity and depth, addressing the complexities involved in the research process and responding to the reviewers' recommendations.

3.1. Data Collection Process

The data collection was conducted across multiple government schools in Sri Lanka, which involved a sample of more than 600 students from diverse urban, suburban, and rural regions (Mahanama et al. 2018). Schools were selected based on their geographic representation and willingness to participate, which ensured a broad spectrum of socioeconomic backgrounds. The student sample was balanced in terms of gender, with approximately equal numbers of male and female students, enhancing the generalizability of the findings within the Sri Lankan educational context.

By focusing on students in grades 6, 7, and 8 (ages 11 to 14 years), performance data were collected over three consecutive academic years. This longitudinal approach provided insights into student progress and developmental trends over time. Specifically, end-term examination marks for all the subjects were gathered for each student, which encompassed core academic areas such as mathematics, science, Sinhala (primary language), English (secondary language), religion, history, health, citizenship education, geography, practical and technical skills (PTS), Tamil (secondary national language), and art.

In addition to academic performance data, comprehensive information on student learning backgrounds and learning styles was collected to enrich the analysis. This included socioeconomic indicators (parents' education levels and occupations), family background details, participation in supplementary educational support such as private tuition, and engagement in extracurricular activities such as sports, clubs, and arts programs.

To ensure the reliability and validity of the data, a stratified random sampling method was used. Schools were stratified based on geographic location and type (urban, suburban, rural), and within each stratum, schools were randomly selected. Within the selected schools, students were randomly chosen from the relevant grades to participate in the study. This sampling strategy aimed to minimize selection bias and ensure that the sample was representative of the broader student population.

Potential biases, for example, non-response bias, were addressed by encouraging participation through clear communication of the study's purpose, ensuring confidentiality, and obtaining the necessary ethics approvals. However, it is acknowledged that schools that declined participation might share characteristics that influence the findings, and this limitation was considered in the interpretation of results.

Ethical considerations were paramount throughout the data collection process. Ethics approval was obtained from the institutional review board of the affiliated university, which adhered to international standards for research that involves minors. Informed consent was secured from both the students and their parents or legal guardians. Confidentiality was maintained by anonymizing personal identifiers and securely storing data, and participants were informed of their right to withdraw from the study at any point without repercussions.

3.2. Data Collection Dimensions and Techniques

The data collection encompassed three primary dimensions: student performance data, student learning background data, and student learning style data. Student performance data included examination marks and assignment scores for each subject, obtained directly from official school records. Student learning background data were collected via structured questionnaires, which captured variables such as socioeconomic status, family background, access to additional educational support, and engagement in extracurricular activities. Student learning style data were assessed by using a questionnaire modeled after the Learning Connections Inventory (LCI) model, which evaluates individual learning preferences across scales such as sequence, precision, technical reasoning, and confluence.

To accommodate the large sample size and diverse participant backgrounds, a combination of data collection techniques was used. Both paper-based and digital questionnaires were administered to collect learning background and style data. Multiple-choice and Likert-scale questions facilitated ease of response and efficient data processing. Academic performance data were collected by obtaining permission from school administrators to access official records, with data extraction conducted on-site to ensure data integrity and adherence to confidentiality protocols.

Given the prevalence of handwritten records in schools, data digitalization was a critical step. Data were manually entered into secure electronic databases, and, to enhance accuracy and efficiency, optical mark recognition technology was used for processing questionnaire responses when feasible by using tools such as scripts for data acquisition with paper-based surveys. Pilot testing of questionnaires and data collection procedures was conducted to refine instruments and ensure clarity and appropriateness for the target age group.

The choice of data collection medium was influenced by the technological infrastructure available at each school. In schools with adequate IT facilities, digital questionnaires were administered by using computers or tablets. In contrast, paper-based questionnaires were used in schools that lacked such resources. This dual approach ensured inclusivity and maximized participation rates, with the familiarity of students with the chosen medium reducing response bias and enhancing data quality.

3.3. Data Preprocessing

Before analysis, the collected data underwent meticulous preprocessing to ensure validity and reliability. Data cleaning involved scrutinizing the dataset for inconsistencies, missing values, and outliers. In cases of incomplete records, efforts were made to retrieve missing information, and, if retrieval was not possible, then missing values were handled by using mean substitution or appropriate imputation techniques.

Examination scores were standardized to account for variations in grading scales across different subjects and schools. Z-scores were calculated to normalize the data, which facilitated meaningful comparisons. Categorical data from questionnaires, such as parental occupation and learning style preferences, were encoded by using numerical representations. For nominal variables, one-hot encoding was applied, whereas ordinal variables were encoded based on their inherent order.

The internal consistency of the questionnaires was assessed by using Cronbach's alpha, with a high reliability coefficient of 0.98, which indicated strong internal consistency among the questionnaire items. This process ensured that the data were suitable for subsequent advanced analytical techniques.

3.4. Analytical Techniques

To thoroughly analyze the data and address the research objectives, several advanced analytical techniques were applied, including correlation analysis, regression analysis, factor analysis, and hierarchical clustering. These methods are widely used in educational data mining to uncover patterns and relationships within complex datasets (Romero and Ventura 2010).

3.4.1. Correlation analysis

Correlation analysis was used to examine the relationships between different academic subjects. Specifically, the Pearson correlation coefficient, Spearman rank correlation coefficient, and Kendall tau were calculated to capture both linear and monotonic associations (Khamis 2008). Using multiple correlation measures allowed for a robust understanding of the inter-subject relationships, accommodating potential non-linearities and the ordinal nature of some data. Correlation matrices and heatmaps were generated to visualize these relationships and identify significant patterns among subjects.

3.4.2. Regression analysis

Multiple linear regression models were constructed to predict student performance in key subjects based on their scores in other subjects and background variables. This approach enabled the exploration of predictive relationships and the quantification of the impact of various factors on academic outcomes (Cohen et al. 2013). Standard diagnostic tests were conducted to ensure the validity of the regression models, including checks for multicollinearity, heteroscedasticity, and normality of residuals.

3.4.3. Factor analysis

Exploratory Factor Analysis was conducted to identify latent constructs that underlie students' performance across different subjects. The suitability of the data for factor analysis was assessed by using the Kaiser-Meyer-Olkin measure and Bartlett Test of Sphericity (Kaiser 1974). Principal axis factoring with promax rotation was used to extract factors, which allowed for correlated factors, which is appropriate given the interconnected nature of academic abilities (Fabrigar et al. 1999). Factors with eigenvalues greater than 1 were retained based on the Kaiser criterion.

3.4.4. Hierarchical cluster analysis

Hierarchical clustering was applied to group subjects based on similarities in student performance patterns. Cosine similarity was used as the distance metric due to its effectiveness in high-dimensional spaces (Tan et al. 2021). Agglomerative hierarchical clustering with average linkage was performed, and dendrograms were generated to visualize the clustering process and identify natural groupings among subjects.

3.4.5. Principal component analysis

Principal component analysis (PCA) was used as a dimensionality reduction technique to simplify the data while retaining most of the variance (Jolliffe and Cadima 2016). Standardized data were used for PCA to ensure each variable contributed equally. The principal components that explain significant variance were retained and used as inputs for clustering algorithms to enhance computational efficiency and mitigate the curse of dimensionality.

3.5. Software and Tools

A web application was developed to collect data, manage the dataset, and visualize the results of the study. The front end of the application was built by using Angular, which provides an interactive interface for data entry and real-time visualization. The back end was developed by using Node.js with the Express.js framework, handling data processing and ensuring secure and efficient management of the collected information. MongoDB was used as the database to store and manage the data due to its flexibility and scalability in handling large datasets.

For data analysis, Python was used due to its comprehensive libraries suitable for statistical analysis and machine learning. Libraries such as pandas were used for data manipulation and cleaning, NumPy for numerical computations, SciPy for statistical functions, scikit-learn for implementing machine learning algorithms, and statsmodels for advanced statistical modeling. Visualization was performed by using Matplotlib and Seaborn libraries, which were instrumental in generating plots, heatmaps, dendrograms, and other visual representations essential for interpreting the data.

This integrated software environment facilitated efficient data handling from collection to analysis, ensuring the integrity and accessibility of data throughout the research process.

4. Analysis

The analysis consists of two main parts: correlation analysis of subjects and hierarchical cluster dendrogram analysis. The correlation analysis shows a year-by-year analysis of subjects, whereas the hierarchical clustering provides an overall analysis of subjects.

4.1. Correlation Analysis

We used three fundamental correlation techniques, Pearson, Spearman, and Kendall, to examine relationships among subjects across different grades. Although Pearson and Spearman indicated higher correlation scores, Kendall produced moderate scores and provided complementary insights into monotonic (rather than strictly linear) relationships. Overall, all three methods revealed a consistent pattern of correlations, indicating that a multi-method approach offers a more comprehensive understanding of the varying strengths and nature of subject interrelationships.

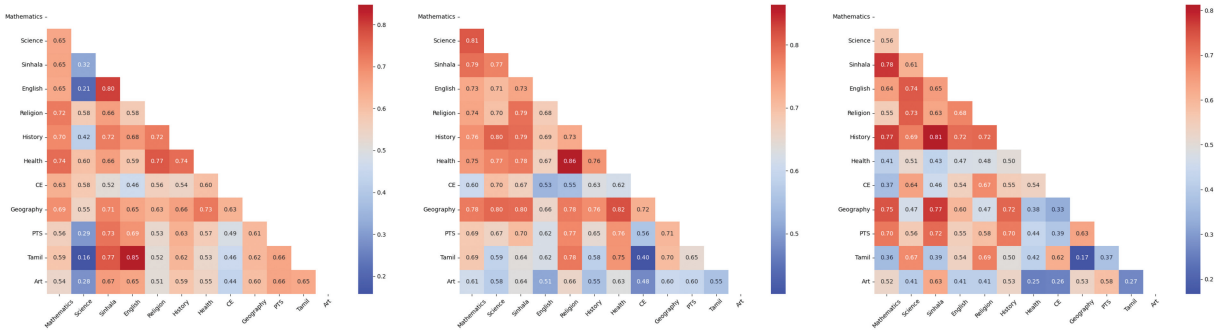
4.1.1. Correlation analysis on subjects

Shown in Figures 1, 2, and 3 are that Pearson and Spearman uncovered more high-correlation pairs among subjects compared with Kendall, which identified fewer but still meaningful, high correlations. Notably, Pearson and Spearman highlighted several strongly correlated pairs, as reflected in Tables 1, 2, and 3. Whereas Kendall coefficients were generally smaller, they confirmed the same overall distribution of subject relationships and underscored the importance of analyzing different types of associations.

The Spearman and Pearson methods identified more relationships between the subjects, which the Kendall method did not, as shown in Tables 1, 2, and 3. This suggests that the subjects do not have a completely linear relationship but rather a monotonic relationship.

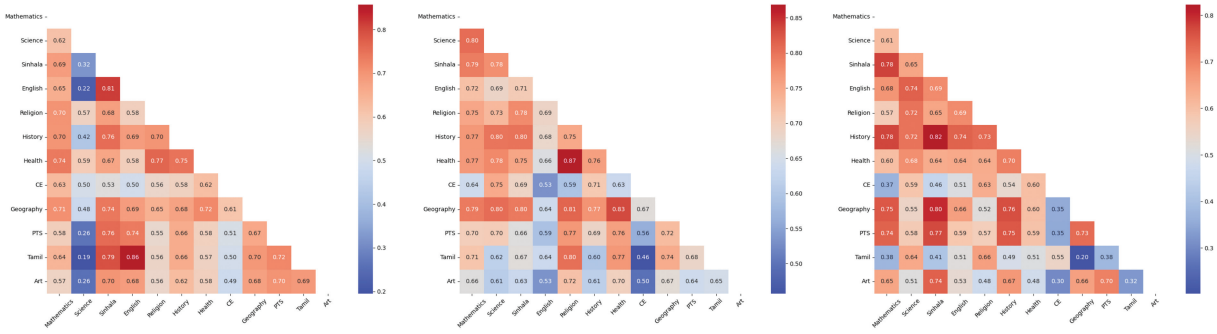
4.1.2. Correlation analysis on all grades

In this analysis, we included data from all grades (grade 6, grade 7, and grade 8). Figure 4 shows the correlation matrix heatmap, in which we identified several key patterns of relationships among different subjects.



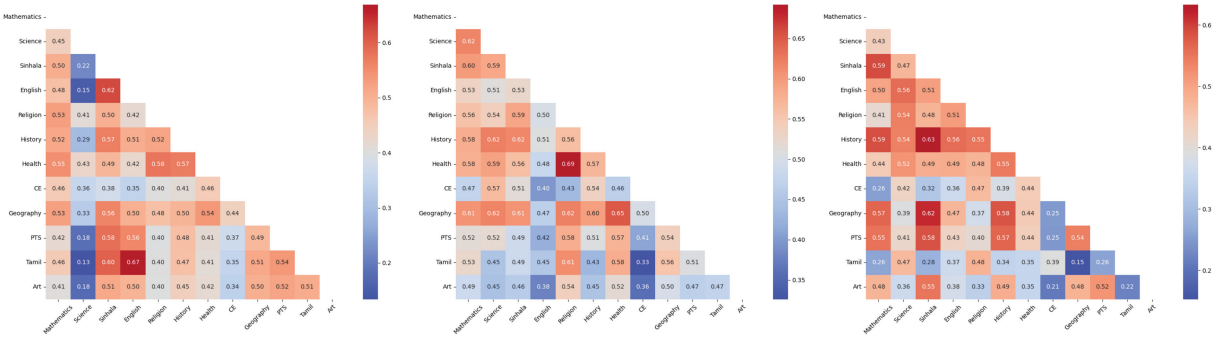
(a) Grade 6 (b) Grade 7 (c) Grade 8

Figure 1: Pearson correlation: (a) grade 6, (b) grade 7, (c) grade 8.



(a) Grade 6 (b) Grade 7 (c) Grade 8

Figure 2: Spearman correlation: (a) grade 6, (b) grade 7, (c) grade 8.



(a) Grade 6 (b) Grade 7 (c) Grade 8

Figure 3: Kendall correlation: (a) grade 6, (b) grade 7, (c) grade 8.

Table 1: Highly correlated subjects in grade 6.

Subject 1	Subject 2	Pearson	Spearman
Sinhala	English	0.80	0.81
English	Tamil	0.85	0.86

Table 2: Highly correlated subjects in grade 7.

Subject 1	Subject 2	Pearson	Spearman
Mathematics	Science	0.81	0.80
Science	History	0.80	0.80
Science	Geography	0.80	0.81
Sinhala	History	0.79	0.80
Sinhala	Geography	0.80	0.80
Religion	Health	0.86	0.87
Religion	Geography	0.78	0.81
Religion	Tamil	0.78	0.80
Health	Geography	0.82	0.83

Table 3: Highly correlated subjects in grade 8.

Subject 1	Subject 2	Pearson	Spearman
Sinhala	History	0.81	0.82
Sinhala	Geography		0.80

First, there were high positive correlations between mathematics and a range of other subjects. Specifically, strong positive correlations were observed between mathematics and science (0.857), Sinhala (0.882), English (0.882), religion (0.839), history (0.881), health (0.837), and geography (0.867). This suggests that students who excel in mathematics tend to perform well in these subjects as well. Similarly, science also exhibited strong positive correlations with mathematics (0.857), Sinhala (0.866), religion (0.833), history (0.883), health (0.873), and geography (0.863). Furthermore, Sinhala showed high correlations with religion (0.920), history (0.934), and health (0.893).

Moderate positive correlations were observed between citizenship-education and several subjects, including mathematics (0.739), science (0.848), Sinhala (0.809), religion (0.781), history (0.821), and geography (0.777). In addition, English demonstrates moderate correlations with mathematics (0.882), Sinhala (0.868), religion (0.849), and history (0.847).

Lower positive correlations were noted between academic subjects and extracurricular activities, such as club activities, racquet sports, and monitor roles. These lower correlations indicate a weaker relationship between academic performance and participation in these extracurricular activities.

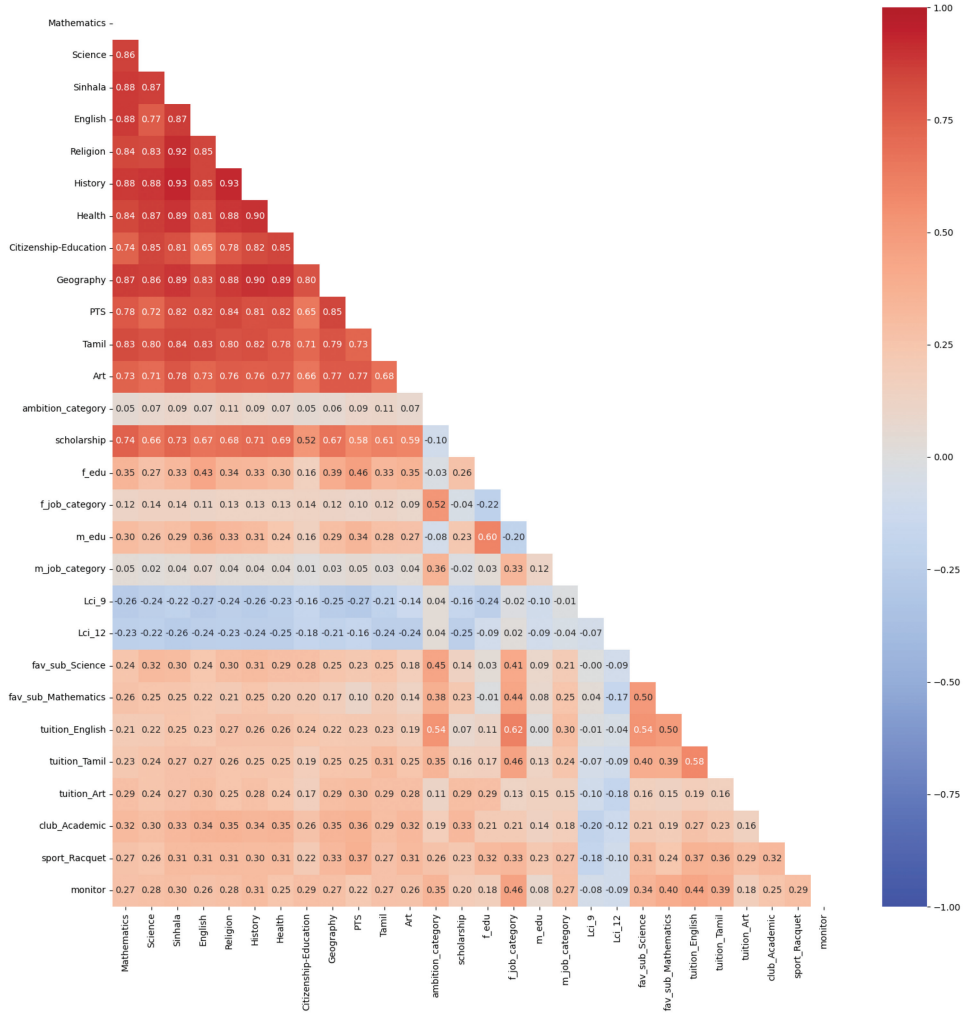


Figure 4: Correlation matrix of all grades.

Negative correlations were found with the variables LCI_9 and LCI_12, which show negative correlations with most subjects. This suggests that these variables may represent factors that inversely affect academic performance.

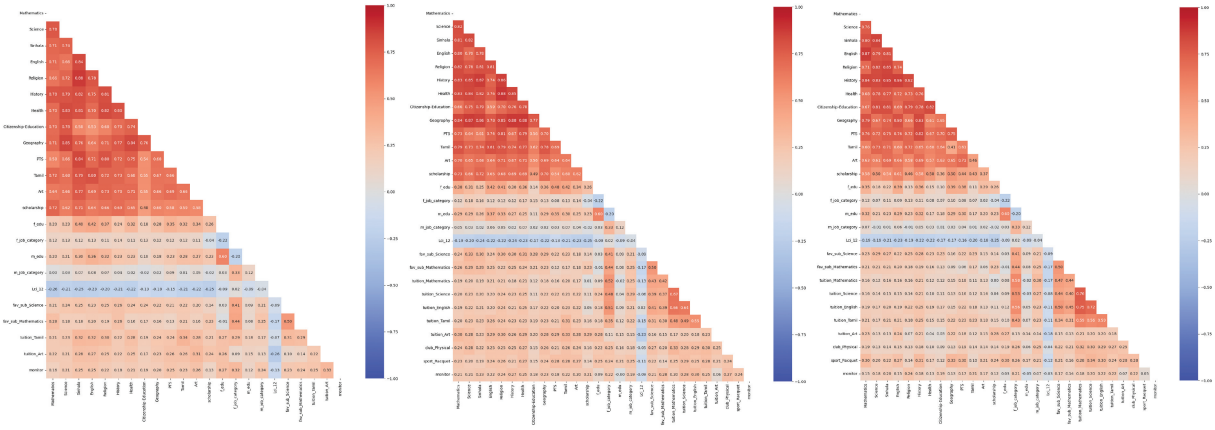
Lastly, interesting observations include small-to-moderate positive correlations between students' favorite subjects (science and mathematics) and their overall academic performance. This indicates that a student's favorite subjects can have a positive impact on his or her performance across other subjects.

4.1.3. Correlation analysis for grade 6

The correlation matrix for grade 6 (Figure 5(a)) revealed significant insights into the relationships among various subjects and other factors. Mathematics exhibited strong positive correlations with science (0.78), history (0.78), and Sinhala (0.71). Science showed high correlations with mathematics (0.78), history (0.79), and health (0.83). Similarly, Sinhala had strong correlations with English (0.84), religion (0.88), and health (0.81). English showed notable correlations with Sinhala (0.84), religion (0.78), and mathematics (0.71). Religion had high correlations with Sinhala (0.88), science (0.72), and health (0.82). In addition, extracurricular activities and parental education levels exhibited moderate correlations with academic performance. For instance, parental education (*f_edu*, *m_edu*) showed a moderate positive correlation with academic subjects, which indicated the influence of parental background on student performance.

4.1.4. Correlation analysis for grade 7

In grade 7, the correlation matrix (Figure 5(b)) indicated a pattern similar to grade 6 but with varying strengths. Mathematics demonstrated high positive correlations with science (0.82), history (0.83), and Sinhala (0.81). Science



(a) Correlation Matrix Grade 6 (b) Correlation Matrix Grade 7 (c) Correlation Matrix Grade 8

Figure 5: Correlation matrices for grades 6, 7, and 8: (a) correlation matrix, grade 6; (b) correlation matrix, grade 7; (c) correlation matrix, grade 8.

showed strong correlations with mathematics (0.82), history (0.85), and geography (0.79). Sinhala had high correlations with religion (0.83), history (0.82), and mathematics (0.81). English exhibited significant correlations with science (0.70), Sinhala (0.76), and religion (0.75). Religion had high correlations with Sinhala (0.83), mathematics (0.82), and history (0.82). Additional factors, such as extracurricular activities (tuition, clubs) and parental occupation categories, showed moderate correlations, which reflects their impact on students' academic achievements.

4.1.5. Correlation analysis for grade 8

The correlation matrix for grade 8 (Figure 5(c)) reveals that mathematics had strong positive correlations with English (0.87), history (0.84), and Sinhala (0.80). Science showed high correlations with Sinhala (0.84), mathematics (0.76), and religion (0.82). Sinhala had notable correlations with religion (0.85), history (0.85), and science (0.84). English demonstrated significant correlations with mathematics (0.87), history (0.86), and Sinhala (0.80). Religion showed high correlations with Sinhala (0.85), history (0.82), and science (0.82). Furthermore, analysis of the data indicated that extracurricular activities and parental education continue to have moderate correlations with students' performance across various subjects.

These analyses underscore the interconnectedness of different subjects and the influence of external factors, providing a comprehensive understanding of the academic dynamics for each grade.

4.2. Regression Analysis

In this study, we performed regression analyses to predict the scores of mathematics, science, and Sinhala for grades 6, 7, and 8 by using the marks of other subjects as predictors. The analyses were conducted separately for each grade. The results of these regression analyses are summarized in Table 4.

Table 4: Summary of regression analysis results.

Grade	Target	Mean Squared Error	R^2
6	Science	219.01	0.55
6	Sinhala	146.76	0.66
6	Mathematics	92.75	0.79
7	Science	97.37	0.81
7	Sinhala	46.15	0.87
7	Mathematics	113.78	0.79
8	Science	116.52	0.72
8	Sinhala	40.82	0.87
8	Mathematics	120.13	0.78

The results indicate that the regression models for predicting mathematics, science, and Sinhala scores in grades 6, 7, and 8 have varying degrees of success. The R^2 values for the models ranged from 0.55 to 0.87, which suggests that a significant proportion of the variance in the target scores can be explained by the predictor variables. The models for predicting Sinhala scores generally performed better, with R^2 values above 0.85 for grades 7 and 8.

For grade 6, the regression model for predicting mathematics scores had an R^2 value of 0.79, which indicates that approximately 79% of the variance in mathematics scores is explained by the scores in other subjects. The models for science and Sinhala had R^2 values of 0.55 and 0.66, respectively.

In grade 7, the models showed improved performance with R^2 values of 0.81 for science, 0.87 for Sinhala, and 0.79 for mathematics. This suggests that the predictor variables are more effective in explaining the variance in target scores for this grade level.

The regression models for grade 8 also demonstrated a strong performance, particularly for Sinhala, with an R^2 value of 0.87. The models for science and mathematics had R^2 values of 0.72 and 0.78, respectively.

Overall, these regression analyses provide valuable insights into the relationships among different subjects' scores and highlight the effectiveness of using other subjects' marks as predictors for the target scores in each grade.

4.3. Factor Analysis

To identify the underlying relationships among the different subjects, a factor analysis was conducted. A factor analysis was performed by using the following steps:

1. **Data Standardization:** The marks for each subject were standardized to ensure that all variables were on the same scale.
2. **Correlation Matrix Calculation:** The correlation matrix of the subjects was calculated to understand the relationships between them.
3. **Determining the Number of Factors:** Eigenvalues were calculated, and a scree plot was used to determine the number of factors. It was found that one factor had an eigenvalue greater than 1.
4. **Factor Extraction:** Factor analysis was performed with one factor, and the factor loadings were obtained.
5. **Handling Missing Values:** Missing values in the dataset were handled by filling them with the mean of each column.

The following table, Table 5, summarizes the factor loadings for each subject across three factors:

The factor analysis reveals the following insights:

- **Factor 1:** This factor has strong negative loadings for all the subjects, which indicates a general academic performance factor. The highest loadings are observed for Sinhala (–0.961), religion (–0.943), and history (–0.962).
- **Factor 2:** This factor shows very low positive loadings across all the subjects, which suggests that it might not significantly contribute to the variance in academic performance.
- **Factor 3:** This factor has a moderate positive loading for English (0.292), which indicates a specific factor that might be related to language skills or preferences.

These results highlight the underlying structure of the academic performance data, with factor 1 representing a general academic ability and factors 2 and 3 capturing more-specific dimensions of student performance. The factor loadings plot (Figure 6) and the scree plot (Figure 7) further illustrate the contribution of each factor to the overall variance in the dataset.

This factor analysis provides a comprehensive understanding of the underlying dimensions of academic performance, aiding in the identification of key areas for targeted interventions and support.

4.4. Hierarchical Cluster Analysis

Hierarchical clustering was applied to identify similar subjects in the syllabus. For this purpose, each subject was represented by a vector of 12 dimensions. Each dimension's value was calculated based on the Pearson correlation coefficient of each subject against other subjects. Cosine similarity was used to calculate the similarity between

Table 5: Factor loadings for various subjects and variables.

Subject/Variable	Factor 1	Factor 2	Factor 3
Mathematics	–0.917	0.075	0.105
Science	–0.911	0.049	–0.185
Sinhala	–0.961	0.032	0.017
English	–0.895	0.048	0.292
Religion	–0.943	0.025	0.036
History	–0.962	0.033	–0.034
Health	–0.937	0.043	–0.108
Citizenship-education	–0.848	0.041	–0.379
Geography	–0.937	0.074	–0.002
Practical and technical skills	–0.858	0.050	0.188
Tamil	–0.863	0.032	0.096
Art	–0.803	0.056	0.080
Ambition category	–0.117	–0.625	0.014
Scholarship	–0.588	0.159	0.107
Father's job category	–0.171	–0.702	–0.034
Mother's job category	–0.060	–0.379	0.090
Learning Connections Inventory 9	0.197	–0.081	–0.119
Learning Connections Inventory 12	0.196	–0.009	–0.043
Favorite subject: Sinhala	0.144	–0.433	–0.203
Favorite subject: mathematics	–0.264	–0.540	–0.043
Favorite subject: science	–0.327	–0.541	–0.125
Tuition: mathematics	–0.236	–0.878	0.029
Tuition: science	–0.251	–0.811	0.034
Tuition: English	–0.301	–0.816	–0.012
Tuition: Tamil	–0.300	–0.581	0.115

subjects. The dendrograms obtained by this approach provide an overview of the grouping of subjects. Due to the use of cosine similarity, the subjects in the same cluster require similar skill sets.

In addition to PTS, art, and citizenship education, which have a larger distance from other subjects, the hierarchical cluster dendrogram identified two clear clusters among the subjects. Mathematics and English language are in the same cluster, as shown in Figure 8, which have the highest failure rates in the ordinary level examination (Department of Examinations, Sri Lanka 2017). Moreover, Cruz Neri et al. (2021) state that mathematics and foreign languages are associated with each other. Furthermore, Bergen (2017) explains that mathematics can be thought of as a foreign language, with its unique terminology and symbol system. Therefore, the results shown in Figure 8 confirm that students in Sri Lanka also showed a similar attitude toward secondary languages and mathematics.

Religion, geography, history, and Sinhala have low distances from each other and are related to reading and memorizing. The next closest distances to these subjects are science and health, which are also related to reading skills, as described in the literature. The cluster dendrogram confirms that science has the closest distance to the above-identified reading cluster.

4.5. Spatial Cluster Analysis

Cluster analysis was performed by using various clustering algorithms to identify distinct groups within the student data. The following algorithms were used: DBSCAN, OPTICS, Mean Shift, HDBSCAN, GMM, and Spectral Clustering. This section summarizes the findings and provides an overview of the clusters identified. Among these, DBSCAN provided better clustering results in this case, identifying distinct clusters and noise points effectively.

DBSCAN (Figure 9) was applied to the PCA-reduced data with an epsilon value of 0.6 and a minimum sample size of 2. The clustering process resulted in the identification of several clusters, including noise points (denoted by cluster label –1). The characteristics of each identified cluster were analyzed, and the results are summarized below.

The analysis of DBSCAN clusters (Table 6) revealed distinct patterns in the academic performance and aspirations of students. Clusters 0, 1, and 2 exhibited higher average mathematics scores and a strong inclination toward

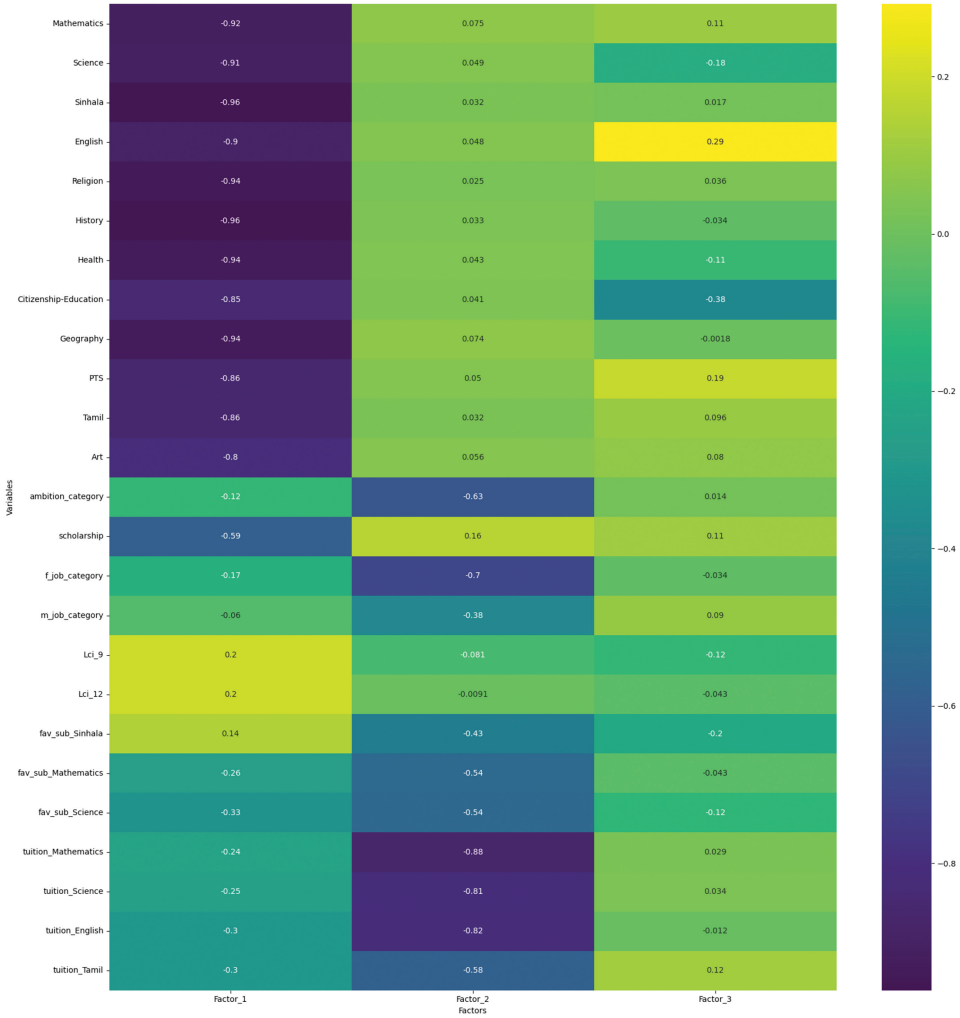


Figure 6: Factor loadings.

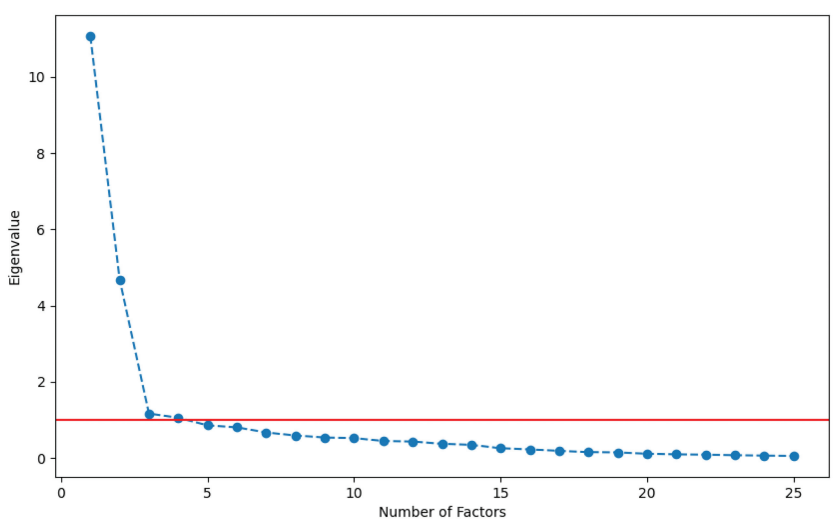
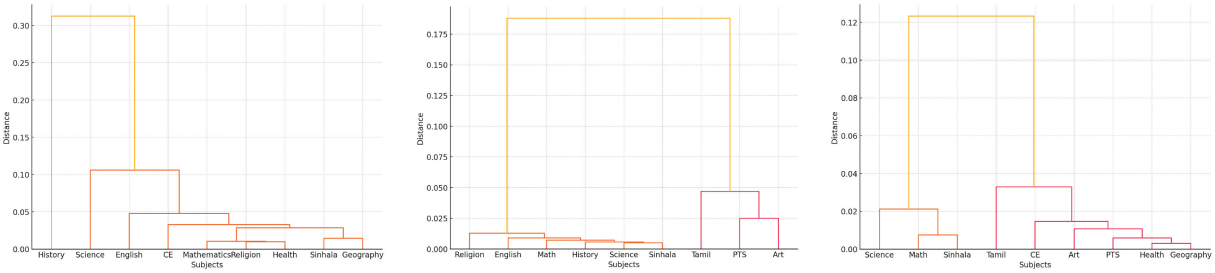


Figure 7: Scree plot.



(a) Cluster Dendrogram for Grade 6 (b) Cluster Dendrogram for Grade 7 (c) Cluster Dendrogram for Grade 8

Figure 8: Hierarchical clustering dendrograms for grades 6, 7, and 8: (a) cluster dendrogram for grade 6, (b) cluster dendrogram for grade 7, (c) cluster dendrogram for grade 8.

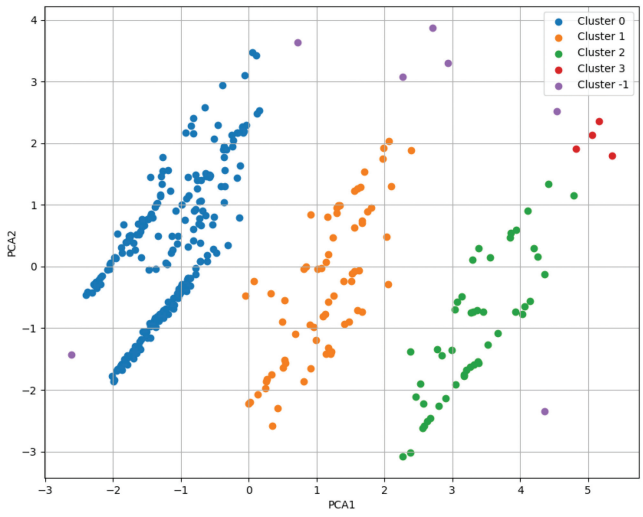


Figure 9: DBSCAN clusters visualization in two-dimensional plane.

health-care–related ambitions, with parents commonly engaged in sales or educational professions. Conversely, cluster 3, which had the lowest average mathematics score, also showed a divergence in parental occupations, which indicated a possible correlation between parental profession and student academic performance.

The noise points identified by DBSCAN (cluster –1) highlight students whose characteristics did not align closely with any other cluster, which suggests unique or outlier profiles.

5. Conclusion

This study presents a comprehensive analysis of subject correlations in secondary education through a holistic approach, encompassing a wide range of academic disciplines. By analyzing performance data from more than 600 Sri Lankan students across grades 6 to 8, we used advanced data mining techniques, including correlation analysis, regression, factor analysis, and hierarchical clustering, to uncover significant patterns in subject interrelationships.

Our findings reveal strong associations between reading and science achievement, consistent with existing literature that emphasizes the critical role of language skills in understanding scientific concepts (O'Reilly and McNamara 2007; Barnard-Brak et al. 2017). Notably, the correlation between reading and science was found to be stronger than that between science and mathematics, which highlights the importance of literacy in science education (Beylik et al. 2021; Jindra et al. 2022; Ünal et al. 2023). This suggests that interventions aimed at improving reading skills may have a substantial impact on students' performance in science.

The moderate association between science and mathematics observed in our analysis indicates that, although these subjects are related, they may require different cognitive skills or learning approaches at the lower secondary level.

Table 6: Characteristics of identified clusters by using DBSCAN.

Characteristic	Cluster –1 (Noise Points)	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Average mathematics score	66.52	70.97	72.07	69.81	57.50
Predominant ambition category	Educational instruction and library occupations	Health-care practitioners and technical occupations	Health-care practitioners and technical occupations	Health-care practitioners and technical occupations	Health-care practitioners and technical occupations
Common father's job category	Farming, fishing, and forestry occupations	Sales and related occupations	Sales and related occupations	Sales and related occupations	Installation, maintenance, and repair occupations
Common mother's job category	Management occupations	Educational instruction and library occupations	Educational instruction and library occupations	Educational instruction and library occupations	Educational instruction and library occupations
Father's education level	High school	Bachelor's degree	Bachelor's degree	Bachelor's degree	High school
Mother's education level	Bachelor's degree	Bachelor's degree	Bachelor's degree	Bachelor's degree	High school
Summary	Students with diverse ambitions and lower academic performance	Students with higher academic performance and ambition toward health-care professions	Students with the highest academic performance, aiming for health-care professions	Students with moderate academic performance, also inclined toward health-care professions	Students with the lowest academic performance, differing parental professions

This insight underscores the need for educators to tailor instructional strategies to address the specific demands of each subject. Given that our analysis did not include data beyond the eighth grade, future research could explore whether the association between science and mathematics strengthens in higher grades, as supported by another study (Wang 2005).

When comparing different correlation techniques, the Spearman correlation was more suitable for our dataset, identifying more relationships between the subjects than did the Pearson correlation coefficient. This suggests that the relationships among the subjects are more monotonic rather than strictly linear, which emphasizes the importance of selecting appropriate statistical methods to accurately capture the nuances in educational data.

The hierarchical clustering analysis revealed that subjects such as citizenship education, art, and PTS had greater distances from other subjects. This indicates that these subjects may assess unique skill sets or inherent abilities not directly linked to performance in other academic areas. Recognizing these distinctions can help educators design curricula that acknowledge the diverse talents and interests of students, potentially enhancing engagement and learning outcomes.

Factor analysis identified a general academic performance factor with strong negative loadings across all subjects, particularly Sinhala, religion, and history. This underscores the significance of a broad academic ability that spans multiple disciplines, reinforcing the idea that foundational skills in language and humanities are integral to overall academic success. Factors 2 and 3, although contributing less to the variance, highlighted specific dimensions related to language skills and other attributes, providing deeper insights into the components of student performance.

Cluster analysis by using various algorithms identified distinct groups of students with varying academic performance levels and parental education backgrounds. These clusters offer valuable insights into student profiles, enabling educators and policymakers to develop targeted interventions and support mechanisms that address the specific needs of different student groups. For instance, understanding that certain clusters of students may benefit from additional support in mathematics or language subjects can inform resource allocation and instructional strategies.

Overall, our study contributes to the field of educational data mining by demonstrating the value of a holistic approach in uncovering complex interrelationships among academic subjects. By integrating multiple analytical techniques, we provide a nuanced understanding of how various disciplines interact, which can inform the development of more effective educational strategies and policies. These findings highlight the interconnectedness of academic disciplines and the necessity for interdisciplinary approaches in education.

Future research should aim to expand the dataset to include a larger and more diverse sample, encompassing different regions and educational contexts. In addition, longitudinal studies could provide insights into how subject correlations evolve over time and across different educational stages. Exploring causal relationships by using advanced machine learning techniques would further enhance our understanding of the factors that influence student performance, ultimately contributing to improved educational outcomes and student success across various contexts.

References

Barnard-Brak, L., T. Stevens, and W. Ritter. 2017. "Reading and Mathematics Equally Important to Science Achievement: Results from Nationally-Representative Data." *Learning and Individual Differences* **58**, no. 1–9. doi:10.1016/j.lindif.2017.07.001.

Bergen, S.L. 2017. "Mathematics and Foreign Language: Authentic Texts in Mathematics." Accessed February 25, 2025. https://api.semanticscholar.org/CorpusID:189339966.

Beylik, A., and E. Genç Kumtepe. 2021. "Examining Transactional Distance in Synchronous Online Learning Environments." In *Motivation, Volition, and Engagement in Online Distance Learning*, edited by Hasan Uçar and Alper T. Kumtepe, 147–167. Hershey, PA: IGI Global. doi:10.4018/978-1-7998-7681-6.ch007.

Cohen, J., P. Cohen, S.G. West, and L. Aiken. 2013. "Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences." 3rd edition. New York, NY: Routledge. doi: 10.4324/9780203774441.

Cruz Neri, N., K. Guill, and J. Retelsdorf. 2021. "Language in Science Performance: Do Good Readers Perform Better?" *European Journal of Psychology of Education* **36**, no. 1: 45–61. doi:10.1007/s10212-019-00453-5.

Department of Examinations, Sri Lanka. 2017. G.C.E. (O/L) Examination 2017 – Performance of Candidates. Accessed February 25, 2025. https://doenets.lk/documents/statistics/G.C.E.(OL)%20%20Examination%202017%20Perfomance%20of%20Candidates.pdf.

Fabrigar, L., D. Wegener, R.C. MacCallum, and E. J. Strahan. 1999. "Evaluating the Use of Exploratory Factor Analysis in Psychological Research." *Psychological Methods* **4**, no. 3: 272–299. doi:10.1037//1082-989X.4.3.272.

Jindra, C., K. Sachse, and M. Hecht. 2022. "Dynamics between Reading and Math Proficiency over Time in Secondary Education – Observational Evidence from Continuous Time Models." *Large-Scale Assessments in Education* **10**, no. 1: 12. doi:10.1186/s40536-022-00136-6.

Jolliffe, I.T., and J. Cadima. 2016. "Principal Component Analysis: A Review and Recent Developments." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**, no. 2065: 20150202.

Kaiser, H.F. 1974. "An Index of Factorial Simplicity." *Psychometrika* **39**, no. 1: 31–36. doi:10.1007/BF02291575.

Khamis, H. 2008. "Measures of Association: How to Choose?" *Journal of Diagnostic Medical Sonography* **24**, no. 3: 155–162. doi:10.1177/8756479308317006.

Maerten-Rivera, J., N. Myers, O. Lee, and R. Penfield. 2010. "Student and School Predictors of High–Stakes Assessment in Science." *Science Education* **94**, no. 6: 937–962. doi:10.1002/sce.20408.

Mahanama, B., W. Mendis, A. Jayasooriya, V. Malaka, U. Thayasivam, and U. Thayasivam. 2018. "Educational Data Mining: A Review on Data Collection Process." In *Proceedings of the 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka, September 27–28. doi: 10.1109/ICTER.2018.8615532.

Mahmoudi, S., E. Jafari, H. Nasrabadi, and M. Liaghatdar. 2012. "Holistic Education: An Approach for 21 Century." *International Education Studies* **5**, no. 3: 178–186. doi:10.5539/ies.v5n3p178.

Marín-Marín, J., A. Moreno Guerrero, P. Dúo Terrón, and J. López-Belmonte. 2021. "Steam in Education: A Bibliometric Analysis of Performance and Co-Words in Web of Science." *International Journal of STEM Education* **8**, no. 1: 41. doi:10.1186/s40594-021-00296-x.

Miseliunaite, B., I. Kliziene, and G. Cibulskas. 2022. "Can Holistic Education Solve the World's Problems: A Systematic Literature Review." *Sustainability* **14**, no. 15: 9737. doi: doi:10.3390/su14159737.

O'Reilly, T., and D. McNamara. 2007. "The Impact of Science Knowledge, Reading Skill, and Reading Strategy Knowledge on More Traditional" High-Stakes" Measures of High School Students' Science Achievement." *American Educational Research Journal* **44**, no. 1: 161–196. doi:10.3102/0002831206298171.

Romero, C., and S. Ventura. 2010. "Educational Data Mining: A Review of the State of the Art." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **40**, no. 6: 601–618. doi:10.1109/TSMCC.2010.2053532.

Tan, P.-N., M. Steinbach, A. Karpatne, and V. Kumar. 2021. *Introduction to Data Mining*. 2nd edition. Pearson: Boston, MA, USA. ISBN 9780137506286.

Ünal, Z., N. Greene, X. Lin, and D. Geary. 2023. "What is the Source of the Correlation between Reading and Mathematics Achievement? Two Meta-Analytic Studies." *Educational Psychology Review* **35**, no. 1: 4. doi:10.1007/s10648-023-09717-5.

Wang, J. 2005. "Relationship between Mathematics and Science Achievement at the 8th Grade." *Online Submission* **5**: 1–17.

Yağcı, M. 2022. "Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms." *Smart Learning Environments* **9**, no. 1: 11. doi:10.1186/s40561-022-00192-z

Yakman, G., and H. Lee. 2012. "Exploring the Exemplary STEAM Education in the U.S. as a Practical Educational Framework for Korea." *Journal of the Korean Association For Research in Science Education* **32**, no. 6: 1072–1086. doi:10.14697/jkase.2012.32.6.1072.

Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 3, No. 1, 2025 DOI: 10.54116/jbdai.v3i1.50

Improving Large Language Model (LLM) Performance with Retrieval Augmented Generation (RAG): Development of a Transparent Generative Artificial Intelligence (Gen AI) University Support System for Educational Purposes

Nishitha Chidipothu
Rutgers University
nc795@scarletmail.rutgers.edu

Rick Anderson
Rutgers University
rick.anderson@rutgers.edu

Jim Samuel
Rutgers University
jim.samuel@rutgers.edu

Alexander Pelaez
Hofstra University
alexander.pelaez@hofstra.edu

Julia Esguerra
Rutgers University
jue5@scarletmail.rutgers.edu

Md Nurul Hoque
Rutgers University
nurul.hoque@ejb.rutgers.edu

Abstract

This study works on the development of a generative artificial intelligence (AI) university support system (GenAI-USS) by improvising retrieval augmented generation (RAG) architecture to improve the performance of large language models (LLM) in a way that supports stepwise transparency. We aim to achieve better transparency and flexibility, and improved accuracy of responses to queries based on university data assimilated from university webpages and knowledge sources. We use RAG to develop a plug-and-play mechanism, along with prompt selection to boost LLM accuracy. One of the key components in our GenAI-USS is the capture and integration of real-time information via live retrieval into the generative AI process. This domain-specific knowledge assimilation with real-time updates to capture changes and new information serves as a specialized dynamic expert knowledge database for RAG. Our RAG mechanism pulls in relevant, up-to-date information from the dynamic database, which pulls real-time data from targeted predetermined knowledge sources. The other key component in our GenAI-USS design is the deliberately

designed information processing visibility at each stage of the process to ensure full transparency, and this includes the following: overview, data collection, storage encoding, testing, chatbot interaction, and search. The testing module allows for interactive viewing of generated responses and their sources. Our strategy is expected to lead to higher-quality AI-generated output via targeted information retrieval, hallucination mitigation, accuracy improvement, and timely data updates. Essentially, on the submission of a query, the RAG-dependent GenAI-USS first identifies the most relevant information from the specialized expert knowledge database and then factors this into the generative AI response development process. This results in a successful implementation of our primary objectives of a transparent and flexible user-choice–driven RAG-based generative AI system, which also provided heuristically notable improvements in the quality of output produced.

Keywords *generative artificial intelligence, large language model, retrieval augmented generation, NLP, NLU, NLG, AI, DSS, Chatbot, transparency, ethics, education.*

1. Introduction

Textual data presents us with an array of opportunities for creating artificially intelligent innovations (Garvey et al. 2021). In recent years, there has been a significant increase in the amount of unstructured textual data being leveraged to glean business insights from textual data distributions and models (Samuel et al. 2020a; 2022). Substantial advancements in natural language processing (NLP) and natural language understanding have led to the development and adaptation of large language models (LLM) for a broad range of applications (Samuel 2023). These models are trained on vast amounts of data to model probabilistic associations between tokens to serve as a machine "brain," which can then be used for actions such as interpreting and answering human questions (IBM 2023). LLMs depend on the "transformer" model architecture that features self-attention mechanisms, which enables them to learn at a significantly accelerated pace compared with conventional models (CloudFare 2024). LLMs are a significant part of the basis for generative artificial intelligence (Gen AI) for text and language generation capabilities, leading to a new wave of AI technologies and applications (Samuel et al. 2024a). Human queries or questions serve as unstructured input, and the LLMs have the capacity to generate output via probabilistic associations of word sequences. The quality of LLM responses depends on the training data, occasionally resulting in "hallucinations," in which the model produces spurious answers, leading to wrong and often wild responses to human users (Xu et al. 2024).

The introduction of OpenAI's ChatGPT (https://openai.com/) and Google DeepMind's Gemini (https://deepmind.google/technologies/gemini/) has given an opportunity to the general public to derive answers to queries that may not be readily available on traditional search tools such as Google (OpenAI 2024; Google Deepmind 2024). Transformers are used in building LLMs, which can deal with human language and various tasks related to NLP, such as machine translation (Anderson et al. 2024). LLMs have shown unparalleled prospects and capacity for innovations in AI applications, leading to the development of numerous enhanced NLP methods and tools. As an illustration, it has been shown that LLMs can improve the scope of NLP functions such as emotion classification and sentiment analysis for public sentiment perception (Alan et al. 2024; Samuel et al. 2020b, 2024b). Popular LLMs are generally referred to as foundation models and have outstanding "emergence" and "homogenization" capabilities (Samuel 2023; Tam 2023). Here, emergence refers to the potential for spontaneous and organic discovery of surprising features and possibilities with LLMs, and homogenization refers to the capability of LLMs to increasingly serve as a common platform for various AI applications. However, these Gen AI tools also have their respective drawbacks, including the potential for misinformation, hallucination, and harmful content generation. Next, we discuss some known challenges and risks associated with these technologies.

A significant challenge of LLMs is their tendency to generate factually inaccurate, meaningless, and absurd information, although seemingly credible, also known as hallucination (Li et al. 2023). This issue is particularly concerning in contexts in which factual accuracy is critical, such as for news reporting and academic research. LLMs can subtly alter factual information while maintaining a seemingly credible narrative, as demonstrated in one of our experiments. In this case, an LLM-based generative system provided a fairly believable narrative about a historical event that it specified as occurring on Monday, October 21, 2014. Although all other facts were correct, the date was inaccurate: October 21, 2014, is actually a Tuesday, not a Monday, as specified by the LLM. This subtle distortion highlights the reliability challenges of LLMs in factual applications.

Another challenge is in keeping the knowledge base of the LLMs updated in real-time. Without continuous updates, these models rely on outdated information that leads to erroneous answers to simple questions such as "Who is the current president of the United States?" LLMs trained on largely historical data may not sufficiently incorporate

contemporary events or real-time developments and fail to adapt to the rapid changes in real-world information (Yu and Ji 2023; Duan et al. 2023). Furthermore, overleveraging could lead to dependency on LLM-based tools and associated harms, such as diminished intellectual capabilities for students in academic environments, resulting in ethical dilemmas for educators. Necessary guidelines and strategies must be implemented by policymakers to foster genuine learning and mitigate the disadvantages that lead to ethical and practical challenges (Das 2024).

The challenge of LLM-generated inaccuracies, hallucinations, and other errors can be significantly mitigated by the use of retrieval augmented generation (RAG) methods, which have become popular in the recent past. A key advantage of RAG is its ability to maintain current knowledge by directly encoding website content, which can be refreshed as source material changes. Because the LLM summarizes data directly from the vector store, responses are grounded in the exact source material, which ensures relatively better accuracy and traceability. We propose a novel RAG-based architecture for the organization of a plug-and-play mechanism for improved transparency and flexibility. One of the key components in our proposed Gen AI university support system (GenAI-USS) is the capture and integration of real-time information. This is achieved via a live retrieval process embedded into the Gen AI system, which updates the RAG database with near real-time information and can also be updated at will. This serves as a specialized expert knowledge database for our GenAI-USS system. Our RAG mechanism pulls in relevant, up-to-date information from a predetermined dynamic database, which pulls real-time data from targeted knowledge sources.

Furthermore, our GenAI-USS design is deliberately designed to maximize process visibility of each stage of the Gen AI process to ensure full transparency. This includes all the major phases, including the following: overview, data collection, storage encoding, testing, chatbot interaction, and search. In addition, the testing module allows for interactive viewing of generated responses and their sources. The GenAI-USS design is centered on testing against relevant question datasets and answers evaluated to be accurate. The combined optimization of embedding and language models ensures that responses remain contextually anchored to website data while enabling efficient updates and source attribution. This approach reveals content gaps and structural barriers in the current site architecture that may impede effective knowledge extraction. Scalability can then be achieved by optimizing the embedding and language models for speed, cost, and energy consumption.

The rest of this article is organized as follows: we perform a literature review on topics critical to our research, with a focus on identifying key concepts and facts on areas such as RAG, best practices on the use of tools such as Hugging Face (https://huggingface.co/) and Streamlit (https://streamlit.io/), and recent advancements in LLMs and prompt engineering. We also identify potential gaps and scope for improvement. Under "Domain and Data," we discuss the various aspects of how data are being used, interpreted, and leveraged for domain-specific considerations and challenges. Next, we discuss the design and development of our GenAI-USS application, explaining each stage with details about areas such as the LLMs used, embedding models, data sources, prompts, and output. Finally, we discuss the implications of our research and GenAI-USS application, the scope for future research, and conclude with thoughts on the way ahead for linguistic Gen AI.

2. Literature Review

We examine several key themes for our research, covering RAG, best practices on the use of tools such as Hugging Face and Streamlit, and prompt engineering. We also cover prompt-design, zero-shot and few-shot learning, embeddings, chunking parameters, temperature, and associated implications. We first try to understand the kind of models and tech stack required for our GenAI-USS application to be successful. We then explore the prompt engineering techniques, including understanding how to design the prompts, and zero-shot and few-shot learning in prompt engineering. Despite significant advancements, there are gaps that remain in the literature. Thus, this review provides us with a foundation for understanding and exploring concepts and usage of these tools. This will set the stage for our current application as well as for future research. We start with RAG, because this mechanism anchors our motivation and our framework; we then introduce Hugging Face as our resource hub for operating purposes; this is followed by elaborations on the user interface, prompt engineering, designing prompts, and other critical elements of our framework.

2.1. RAG

Over time, we have encountered challenges with LLMs that produce inaccurate text-based generative responses, primarily due to the absence of precise and up-to-date datasets attached or linked to the queries referred to as "hallucinations." This led to the emergence of RAG. RAG uses LLM to generate responses with the help of a database attached so that when a query is presented, the RAG system first identifies relevant information from external sources and then integrates this information into the response generation process (Jiang et al. 2023). The database

attached will be divided with the help of a splitter because RAG works effectively with smaller text segments stored as document snippets. These document snippets will be fed into an embedding machine to convert the text into vector embeddings. This step helps in retrieving more factual information, reducing the occurrence of hallucinations, thereby improving the quality of outputs (Gao et al. 2024; Aquino 2024).

The forward-looking active RAG uses iterative retrieval-augmented generation to anticipate future content in sentences. It retrieves relevant documents and regenerates sentences that contain low-confidence tokens. This approach enhances the efficiency of retrieving information multiple times (Jiang et al. 2023). Given its proven effectiveness, RAG has become the most cost-effective, straightforward, and low-risk solution to achieving superior performance for GenAI applications, which leaves many companies with little choice but to adopt it (Proser 2023). Furthermore, recent research highlights the importance of using RAG in contexts that require high accuracy, such as university-level textbooks, in which ChatGPT-4 was unsuccessful in delivering accurate information (Wang et al. 2024).

2.2. Hugging Face

Hugging Face stands out as a rapidly expanding hub for hosting open-source projects centered on machine learning and AI (Hugging Face 2024) and will serve as a data store for all the LLMs that we require for GenAI-USS. This platform's primary focus is on transformers that streamline the process for individuals and small business startups to develop extensive LLMs (Vasilis 2024). Hugging Face's transformers are a groundbreaking advancement in NLP, combining transfer learning methods with large-scale transformer language models. It offers state-of-the-art transformer architectures, along with a collection of pre-trained models, which makes it a cornerstone in modern NLP research and provides powerful tools for various tasks (Wolf et al. 2020).

2.3. User Interface

We use Streamlit, an open-source Python framework (Streamlit 2024), to develop our framework. Streamlit offers a chance for individuals less experienced in developing applications to perform well by providing resources that enhance the appearance and interactivity of projects. We are hosting and deploying our project on the Streamlit community cloud. The deployment process involves configuring the Streamlit app for web integration, cloning our GitHub repository, and using the platform's infrastructure to ensure smooth integration with other machine learning models (Imanuelyosi 2022).

2.4. Prompt Engineering

Prompt engineering optimizes and refines input queries so that the LLMs can generate more enhanced, accurate, and coherent responses. Our focus is on using prompt engineering to ensure that we guide the users in providing accurate inputs to achieve relevant outputs. GPT- and DALL-E-like foundation models use natural language prompts to facilitate interaction with AI models, which have introduced us to newer tools and methods. For example, Prompt Sapper facilitates building AI services based on prompt engineering (Cheng et al. 2024). Prompt engineering has been used as an important technique for designing relevant instructions or queries to improve the performance of the LLMs, which are used in AI, ML, and NLP tasks such as sentiment analysis, summarizing, questions answering, and arithmetic reasoning (Shi et al. 2023). Shi et al. (2023) also emphasized the use of chain of thought, zero-chain of thought, and in-context learning in effective prompt engineering. Furthermore, Ekin (2023) pointed out that clarity, explicit constraints, and leveraging various types of questions are some important factors for prompt engineering. Meanwhile, Lo (2023) emphasized conciseness, logic, explicitness, adaptability, and reflectiveness, also known as the CLEAR framework, in dealing with the prompt engineering for AI, LLM, and NLP models. Prompt engineering uses instruction-basis, information-basis, reformulation, and metaphoric prompt techniques, along with effective evaluation processes for instructing LLMs to improve the performance of the AI-NLP tasks (Rathod 2024). In addition, Ye et al. (2024) emphasized removing unwanted elements from the prompts, thus improving capabilities of reasoning and optimizing communication between tasks to build a meta-prompt so that we can lead the LLMs in enhancing their performance.

2.5. Designing Prompts

In designing effective prompts to interact with the AI, LLM, and NLP models, it is essential to have a comprehensive understanding of what the users are looking for and the capabilities of the models. There are some key principles in prompt design that can improve the effectiveness of prompting across various models and tasks (Herrmann and Nierhoff 2017). Herrmann and Nierhoff (2017) emphasized making the prompts optional and comprehensible so that the users can interact with the models without any difficulties and pressure. They suggested that each

prompt should have a clear purpose that leads to specific actions and desired outcomes. In addition, the prompts should be structured according to the respective user contexts and users should have control over the prompt design so that they can achieve improved acceptance and effectiveness of prompting (Herrmann and Nierhoff 2017).

Desmond and Brachman (2024) suggested a trial-and-error process in prompt designing. This iterative testing system allows the users to refine the prompts based on the responses from the models. Furthermore, Liu and Chilton (2021) emphasized adding structured elements, such as subject and style, to improve the coherent outputs in generative tasks. Although these techniques and strategies are useful for creating a framework for designing prompts effectively, there are some inherent complexities and challenges for the users to interact with the AI-NLP models, such as performing consistently across diverse tasks and applications (Dang et al. 2022).

2.6. Zero-Shot and Few-Shot Learning

Zero-shot and few-shot learning are instrumental in dealing with AI models, particularly in limited labeled data settings. These techniques allow the AI models to leverage the previous knowledge efficiently and generalize the outputs from none to very little labeled data. If there are no direct training examples in the datasets, zero-shot learning helps AI models recognize the unrevealed classes (Deng et al. 2024). Deng et al. (2024) demonstrated the speech-to-text alignment effectively by using Wave2Prompt, which combines spoken units with LLMs to execute zero-shot activities. In addition, Liu et al. (2021) showed how the Micro framework improves the capacity of the LLMs within contexts and enhances the performance of extracting zero-shot relations by using varied datasets without updating parameters. However, few-shot learning can handle the AI models with limited examples. Few-shot learning helps in cross-domain fault diagnosis by using embedding optimization and solves the challenges and problems effectively in industrial tasks (Qiu et al. 2024). Liu et al. (2021) introduced a combined representation learner for classifying few-shot images, which can remarkably enhance performance across diverse datasets. Although these methods have great prospects, there are some challenges in generalizing the outputs across various domains and tasks, which requires future research initiatives in this arena.

2.7. Context and Prompt Engineering

Contextualization is instrumental in improving the performance and effectiveness of AI, LLM, and NLP models as well as user experience by leading the prompts to specific contexts. Jasmine (2024) pointed out that AI models and systems can achieve improved and accurate outputs from personalized, contextualized, and relevant prompts. Muktadir (2023) stated that contextualized prompts can achieve enhanced controllability and adaptability in AI-NLP tasks by using transfer learning and attention mechanisms used for context-aware language models. In addition, de Fonseca et al. (2023) demonstrated that the prompts that are context-aware showed improved performance, more accuracy, and cost efficiency compared with the traditional supervised methods and techniques. Although contextualization is beneficial for prompting AI, LLM, and NLP models, this requires extensive data to build effective prompts. There is also the possibility of overfitting to a certain context, limiting the capabilities of generalizability of the models across diverse topics.

2.8. Embeddings

Embeddings are considered the languages of LLMs and GenAI (QuantumBlack 2023). This approach uses linear algebra to convert real-world objects into numerical representations in a high-dimensional space, which allows machine learning and AI systems to comprehend complex knowledge. These embeddings are continuous values. Embeddings provide a simplified representation of real-world data while preserving semantic and syntactic relationships (AWS 2024). In this project, we are working with two different types of embedding models to choose from: sentence-transformers/all-MiniLM-L6-v2 and thenlper/gte-smalls. In the future, we can add more embedding models to work with. Here, the main purpose of using embedding models is to convert the JSON documents, which are in the form of vector stores, for the LLM model to interpret and answer our queries in the next steps.

2.9. Chunking Parameters

Chunking is the process of breaking down large files into smaller segments for better semantic understanding. Before embedding, our primary objective will be to provide as little noise as possible to the embedding model for it to stay semantically relevant. “Your goal is not to chunk for chunking sake, our goal is to get our data in a format where it can be retrieved for value later” (Mishra 2024). There are multiple chunking strategies: fixed-size chunking, recursive chunking, document-based chunking, and semantic chunking. In our project GenAI-USS, we are

focusing on recursive character text split, which is a method that takes up large texts and splits them into smaller chunks based on the parameters given, including chunk size and chunk overlap.

2.10. Temperature

The "temperature" of an LLM is a key control parameter that governs the randomness of the responses generated by the LLM (Peeperkorn et al. 2024). Usually, the default is set to 1, which represents a balanced position. When set closer to 0, a lower level of randomness is expected, and a higher temperature significantly greater than 1 will result in hallucinations. In our design, we choose a low temperature so that the LLM prioritizes factuality during the response process rather than randomness. This parameter can be used to control an LLM's entropy, and, when the temperature is set lower, then the outputs tend to be more factual, predictable, often repetitive, and more closely and more likely aligned with source information. In contrast, higher temperature settings spur LLMs to generate more random, less likely aligned with source information, and relatively more unique responses. This is an important control parameter and needs to be set to be aligned with the main objectives of a Gen AI system.

3. Conceptualization of GenAI-USS

Our current project develops and significantly improves the scope and performance of our academic bot's capabilities, leading to the creation of the Rutgers University GenAI-USS. GenAI-USS is a project developed by the Office of University Online Education Services (UOES TLT n.d.) and the Master of Public Informatics Program at the Bloustein School, both at Rutgers University (Rutgers University 2001). The RAG ideation for GenAI-USS proceeded from the Public Informatics NLP Studio, including the Spring 2024 and Fall 2024 classes in the Master of Public Informatics program. GenAI-USS, also known as the "Chatbot Online Resource (COR) Navigator," is a project under the UOES. COR Navigator was initially designed to assist academic units across Rutgers University in crafting and delivering hybrid and fully online courses. The platform aims to enhance effectiveness, streamline processes, and elevate satisfaction within the Rutgers Online Learning community, which includes prospective students, current students, faculty, and emerging online programs (UOES Rutgers 2023). The GenAI-USS project focuses on developing an interactive question-and-answer interface, including frequently asked questions, and enabling knowledge synthesis through a RAG module. By allowing the incorporation of a broad range of URLs, this system enhances relative accuracy, depth of knowledge, options for expanding breadth of knowledge, and contextual relevance of Gen AI outputs.

GenAI-USS uses a local embedded selection approach, curating information from diverse sources such as blogs, documents, PDFs, and web pages hosted within the UOES domain, with the capability to be extended to any part of Rutgers University. The curated dataset is integrated into a vector database and processed through an LLM by using the RAG framework. This integration streamlines access to valuable, previously hidden information and facilitates seamless interactions between faculty, students, and staff, and the GenAI-USS platform. A conceptual representation of the system architecture is presented in Figure 1. The "Organize" section highlights semantic organization, data clarity, and redundancy cleanup. The framework distinguishes between the human input, referred to as the "query," and the adaptable and testable prompt provided to the LLM. This adapted prompt, informed by local embeddings and the LLM's pretrained knowledge, enhances output quality and relevance compared with raw queries without prompt-engineered support. Through this adaptive framework, GenAI-USS improves system transparency and flexibility, and also supports LLM performance and accuracy, providing an innovative tool for augmenting human performance. This initiative reflects UOES's commitment to advancing online education at Rutgers University by integrating cutting-edge technology and addressing the specific needs of its diverse academic community (Samuel et al. 2022).

Our primary objective is to provide a platform for faculty, students, researchers, staff, and other stakeholders at Rutgers University to access information from the GenAI-USS. This will be achieved by implementing the RAG-LLM application, Hugging Face, and StreamLit interface. The front-end application will resemble Figure 2, which features a chatbot interface that is provided for asking questions and receiving answers.

3.1. Domain and Data

Similar to many other organizations, academic institutions are interested in implementing LLMs to enrich the learning experience for both faculty and students (ElementX 2024). It is clear that LLMs contain inaccurate information, and RAG addition will solve this problem; in our case, in which academic institutions need up-to-date and accurate information, it is ideal to implement RAG-LLM. The academic institution that we are focusing on is Rutgers University. It is ranked as the number 1 public university in the state of New Jersey and holds a history of more than 250 years with 67,200 undergraduate and graduate students, and more than 17,450 full-time and part-time staff

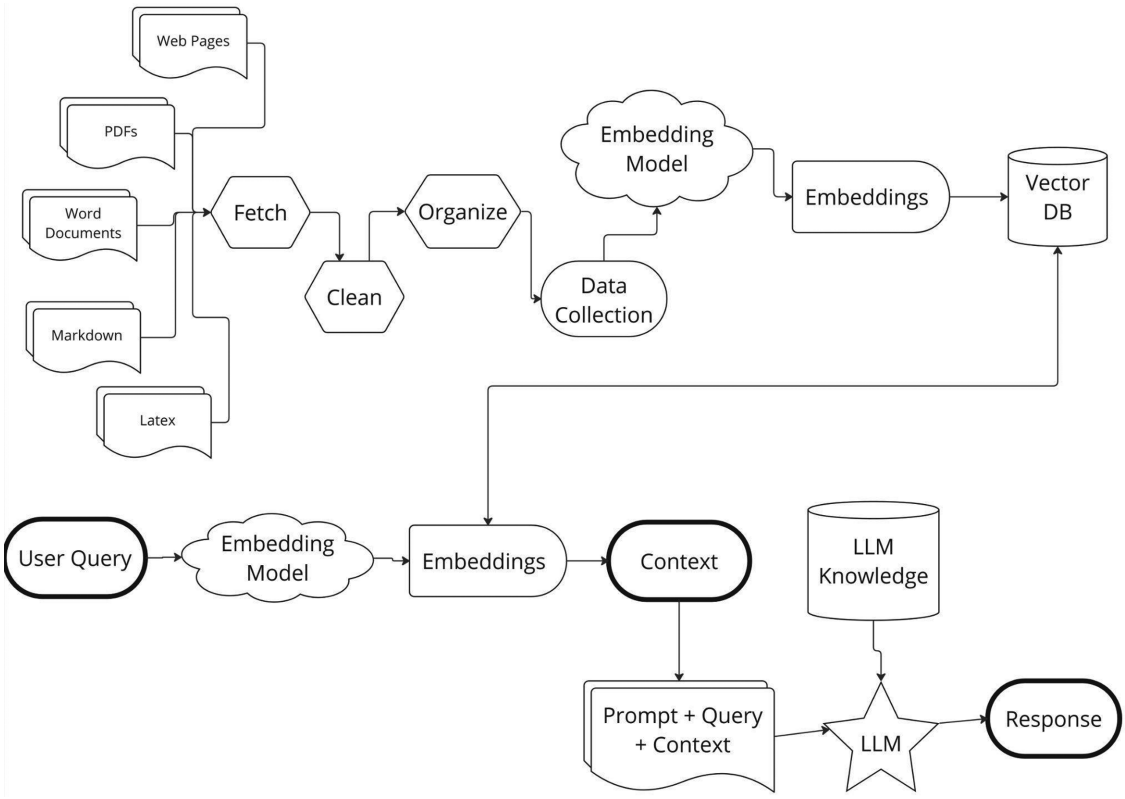


Figure 1: Generative artificial intelligence (AI) university support system (GenAI-USS) architecture.

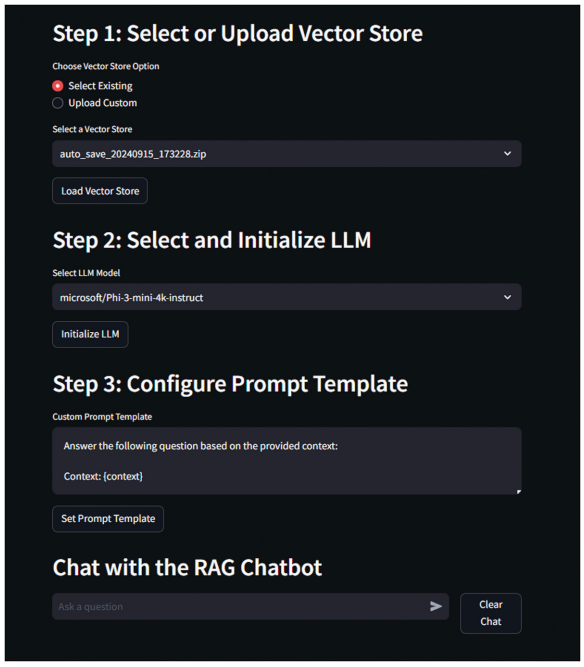


Figure 2: Screenshot of chatbot retrieval augmented generation (RAG).

(Rutgers 2020). A university of this scale has an extensive number of resources for students and faculty that act as silos of information and resources. Faculty and students do not have guidance on how to navigate through these resources, even with capable search engines to index that material. To fulfill this purpose for students and faculty at Rutgers University, we will implement RAG. The first step in the process will be with Rutgers Online Degree Programs (https://www.rutgers.edu/academics/online-degree-programs) as our primary website source. Our final goal is to access academic information through application programming interfaces (APIs), databases, document repositories, websites, or PDFs but we are currently concentrating on accessing information in the form of website URLs. The source data from Rutgers Online Degree Programs | Rutgers University will be used as input to the retriever component of RAG. This component will retrieve the relevant information by using the user's query. Next, a prompt is generated based on the retrieved information and user query, which will serve as an input to the LLM, in turn, generating an answer. The final step would include the LLM parsing the generated answer and presenting it to the user in a clear and understandable format. For the purpose of maintaining high-quality prompts, we also developed a reasonably comprehensive question bank for all reasonable questions to train the model. Our primary question bank of most anticipated questions that students and faculty would pose consists of 100 questions from prospective students' perspectives and 50 questions from faculty's perspective, based on information available on the webpages.

3.2. GenAI-USS Prompt and Prompt Control

The RAG pattern contains four key configurable points that influence its performance. We use sensible defaults for prompts and configurations for quick interactive testing, which allows users to quickly evaluate the tool's performance with minimal customization. This approach enables rapid iteration and general feedback collection. Our process extends the interactive testing by supporting comprehensive validation approaches. Users can test custom prompts for specific scenarios, whereas the advanced features enable bulk testing of multiple prompts. By following our specified data format, entire collections of prompts can be systematically evaluated against test question sets, enabling thorough performance analysis.

3.3. GenAI-USS Application

A Streamlit-based user interface will be used to facilitate the mining, processing, and embedding of the data from the Rutgers Online Degree Programs | Rutgers University. The GenAI-USS system is organized into multiple pages, each dedicated to a specific part of data processing workflows. This user interface will guide users through each step of the process, ensuring a seamless and intuitive user experience. After launching the app, users can navigate between different pages by using the sidebar. Each page includes interactive elements, such as input fields, dropdown menus, and checkboxes, which allow users to customize each step of the data processing pipeline.

First Page: The first page of the chatbot application (Figure 3) serves as a dashboard that serves as an overview of the chatbot's current settings and provides an opportunity to configure the system's performance, select different language models, and initiate queries.

General Settings Panel: The first section of the page provides current settings for running the chatbot. This includes multiple settings related to how the chatbot environment is configured and is being executed. These settings include the following:

- **Running on Streamlit Cloud**: This option checks whether the application is being hosted on Streamlit Cloud. The value displayed will either be true or false, depending on whether the cloud hosting is active or not.
- **Number of CPUs:** This option helps understand the number of available CPU cores allocated to the chatbot.
- **Using Multiprocessing Session State**: This option checks whether the application is leveraging multiprocessing to enhance the performance of the state. If true, then multiprocessing is enabled to handle tasks concurrently and can improve performance and process time.
- **TOKENIZERS_PARALLELISM**: This option, if set to "True," can boost the efficiency of text processing by parallelizing the tokenization of input text across multiple threads.

Vector Store Backup Datasets: This section of the page enables the user to choose any dataset from multiple vector store backup datasets. These vector stores serve as repositories of pre-processed text data, stored in vector format. This step ensures that the user does not perform similar vector data extraction and can directly proceed to perform queries on currently available data backups.

Selecting the LLM: The LLM defines how the chatbot processes language, generates responses, and interprets user queries. This page preselects three models to compare, chosen to represent different trade-offs in performance

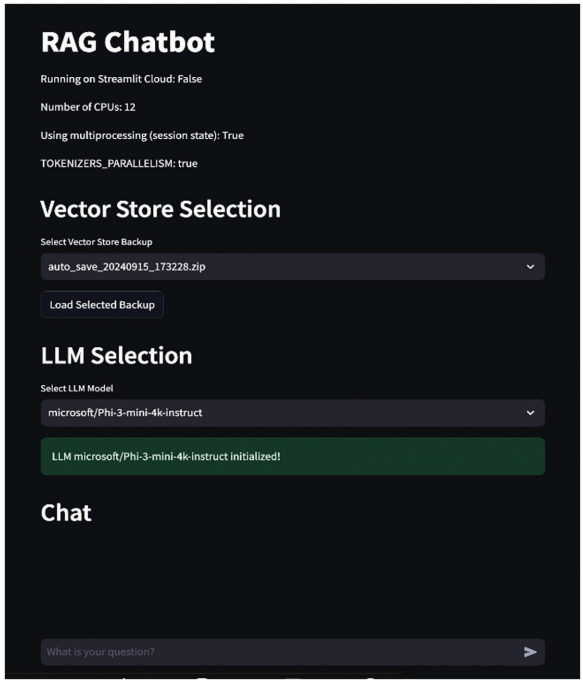


Figure 3: Screenshot of the application.

characteristics. Each model offers distinct advantages in terms of cost, computational requirements, parameters, response relevancy, and context window size. For example, Mistral-7B-Instruct provides a balance of efficient computing and good response quality, whereas GPT-3.5-turbo offers stronger performance at a higher cost (Glover 2024). Phi-3.5-mini-instruct demonstrates capabilities with lower resource requirements. For example:

- **Mistral-7B-Instruct (7B parameters, 32K context window):** Provides efficient computing and good response quality, suitable for self-hosted deployment
- **GPT-3.5-turbo (175B parameters, 16K context window):** Offers stronger performance through API access, with higher associated costs
- **Phi-3.5-mini-instruct (3.8B parameters, 2K context window):** Demonstrates capable performance with lower resource requirements, ideal for lighter workloads

This step makes sure the model is being tailored to different use cases, in which we can choose the context window, number of parameters, and file size. This includes determining whether local LLM access versus remote-hosted API access is appropriate. This step allows for the testing of LLM configuration impact on the question and answer (QA) performance and accuracy.

Chat Section: Users can interact directly with the chatbot by entering their queries. It consists of an Input Field, a Submit Button, and an Output Display. We can input the query and click the submit button to generate the output that is based on the LLM selected and the underlying vector store dataset, ensuring relevance and accuracy.

Second Page: This page (Figure 4) provides the project overview and serves as an index page, providing users with brief information about the steps and contents available in the application.

The step-by-step process is to work on the following:

- Data Collection
- Data Organization
- Encoding Vector Storage
- Testing and QA
- Chatbot Implementation

Figure 4: Screenshot of the project overview.

With each step, we will be provided with updates with regard to whether the data are defined or documents are fetched. This page is built mainly with the help of the streamlit switch_page() function, which directs to multiple pages created in the application. This page gives an overview of the key concepts, steps to get started, prerequisites, and expected outcomes. This page mainly serves the purpose of setting expectations. Users can proceed to the next page by clicking on the "Start with Data Collection" button.

Third Page: This page (Figure 5) provides information on data collection. The core functionality of this page revolves around scanning the URLs to extract information, organizing it into tabular format, and preparing it for further processing.

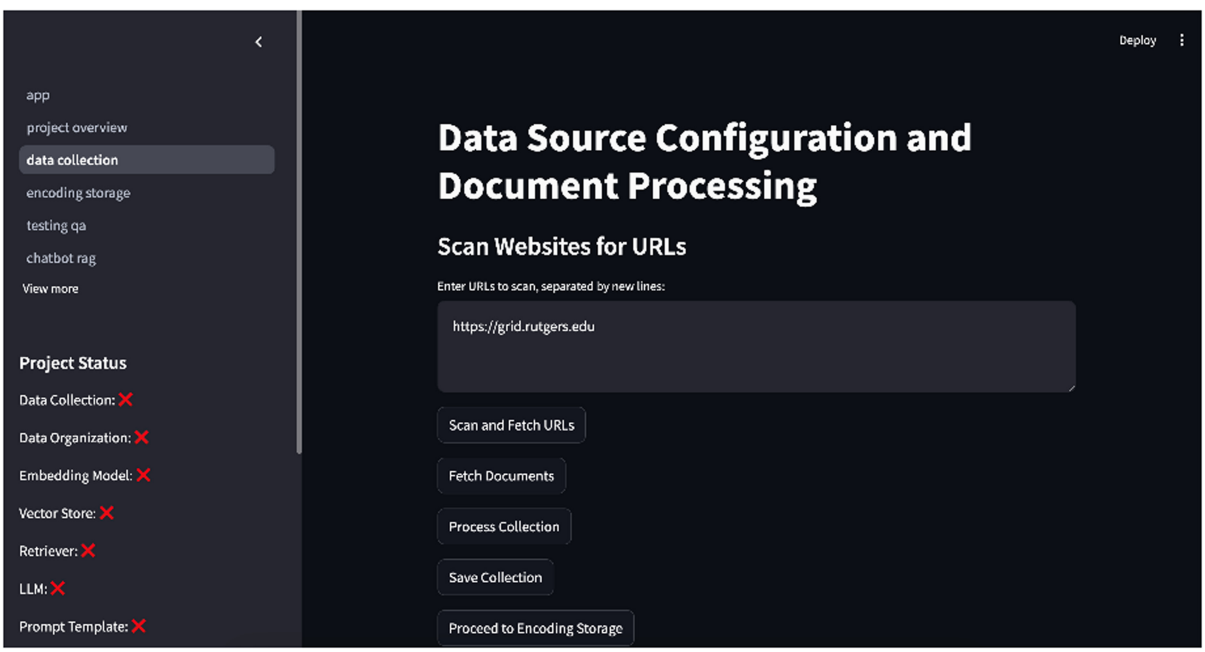


Figure 5: Screenshot of data collection.

URL Input Section: This section on the page allows the users to input the URLs that will act as the source data. The users can enter multiple URLs to scan, separated by new lines. The information is extracted from these URLs and displayed in tabular form. These tables will include columns that detail the following:

- URL link
- Type of URL: this column specifies whether it is an internal or external URL. Internal URLs refer to web pages within the same website, whereas external URLs are links to web pages outside the main website but accessible through it.
- Page name
- Scanned date and time

This step ensures that we gain access to every page related to the given website.

Document Processing and Cleanup: After extracting all the URLs associated with the main source URL, users can proceed to fetch, clean, and organize documents. Here, users can have the option to delete the unnecessary URLs before initiating the document retrieval process by using the "**Fetch**" button. The documents will be fetched in the form of a JSON file, which the users can save and download for later use, as shown in Figure 6. This page is crucial in ensuring that all the relevant data sources are accurately captured and pre-processed before advancing to the next steps. By giving the users control over the URL input and cleaning process, it is ensured that the system is robust and adaptable to various use cases.

Fourth Page: This page (Figure 7) provides information on the encoding storage. This step is critical for processing and storing the extracted information for further processing, particularly in tasks such as question-answering and conversational search. The number of documents that are fetched from the session state will be available. Before proceeding to the encoding step, users are given the option to upload a JSON file. This will provide an opportunity for users to work with a previously saved dataset or any new data in JSON format that needs to be processed. After this, users will be able to choose the embedding model for processing. "Embeddings" refer to vectors that encapsulate the connections and significance among words, thereby representing semantic relationships (Bhavsar 2024). These embeddings will serve as a foundation for question-answering, conversational search, and other functions. The embedding model has three different options from which to choose. These embedding model options are selected to represent different trade-offs in computational requirements and embedding quality as follows:

- **sentence-transformers/all-MiniLM-L6-v2 (384 dimensions, 6 layers):** Balances performance and efficiency, producing high-quality embeddings while maintaining reasonable computational costs. This model excels at semantic similarity tasks and shows strong performance in production environments.
- **thenlper/gte-small (384 dimensions, 4 layers):** Optimized for scenarios with limited computational resources, offering faster inference times while maintaining acceptable embedding quality. Its reduced architecture makes it suitable for environments in which processing speed is prioritized over maximum accuracy.
- **Other:** The framework supports the integration of custom embedding models, which allows teams to implement specialized models for specific use cases. This flexibility enables the testing of newer models or domain-specific variants as they become available.

More flexibility: After selecting an embedding model, users configure two critical parameters, chunk size and chunk overlap, which together determine how documents are segmented for embedding. These parameters are essential for maintaining semantic coherence and ensuring accurate retrieval of relevant information from the vector store.

- **Chunk Size:** The maximum number of characters each chunk can hold

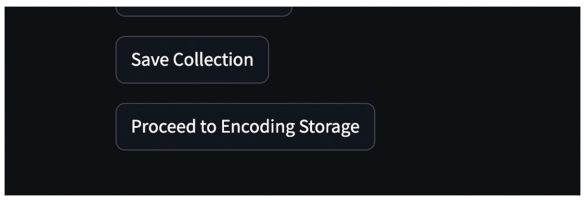


Figure 6: Save data collection.

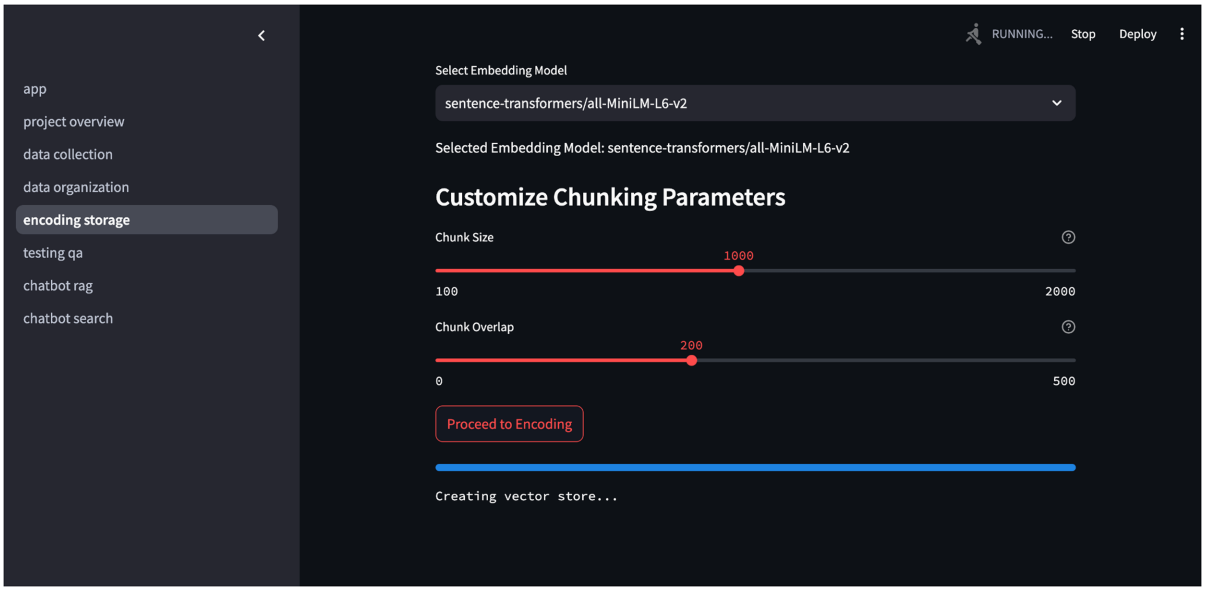


Figure 7: Screenshot of encoding storage.

- **Chunk Overlap:** The number of characters that should be shared between two consecutive chunks (Peter 2023)

We then proceed to "Create Vector Store," which gives us a progress update, along with in-detail information about the total number of documents processed, total chunks, average chunk length, splitting time, encoding time, and total processing time. Once the vector store creation is complete, we can download the vector store, which can be stored or used in the next stage: testing and QA.

The encoding and storage step is crucial because it transforms unstructured data into structured vector representations that can be effectively used in downstream tasks such as semantic search, conversational AI, and question-answering systems. The key statistics to measure the size and performance of creating the vector store from our collection are shown in Figure 8.

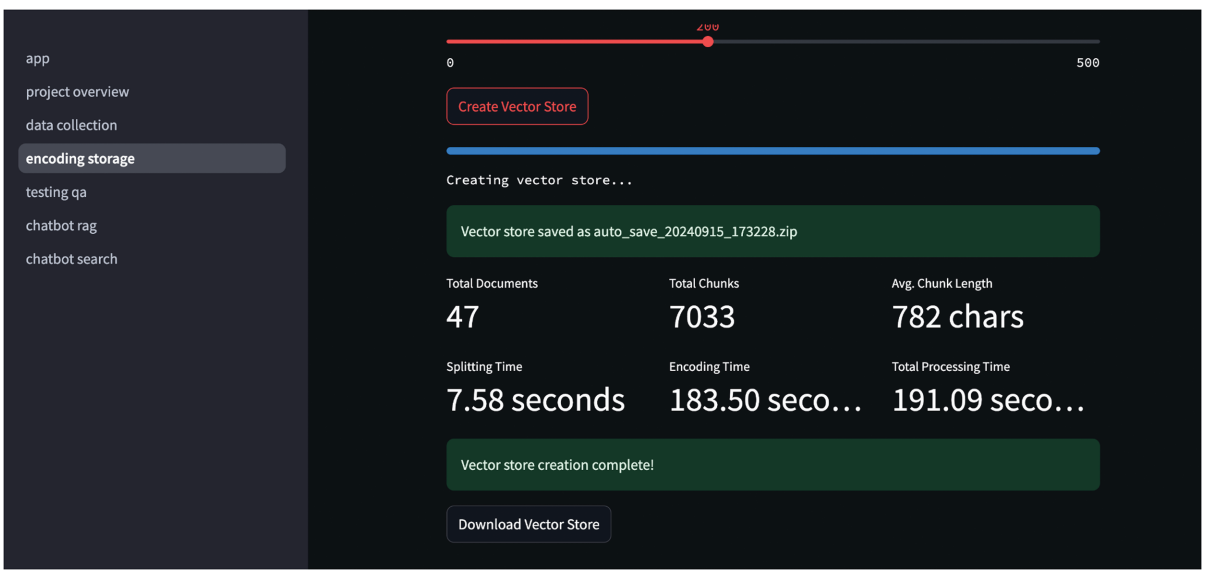


Figure 8: Screenshot of encoding process.

Fifth Page: This page is dedicated to testing the model and QA. We now have a list of vector store documents saved from the previous step, as shown in Figure 9. We can select the vector store backup from the list of backups and then proceed to load it. We also have an option to unload the current one to be sure of the ones we are trying to select.

In Figure 10, the LLM Configuration Step is shown, which involves the following:

- **Selecting the LLM** will have three different options to choose from:
  1. "mistralai/Mistral-7B-Instruct-v0.2"
  2. "Phi-3.5-mini-instruct," "gpt-3.5-turbo"
  3. "Other"

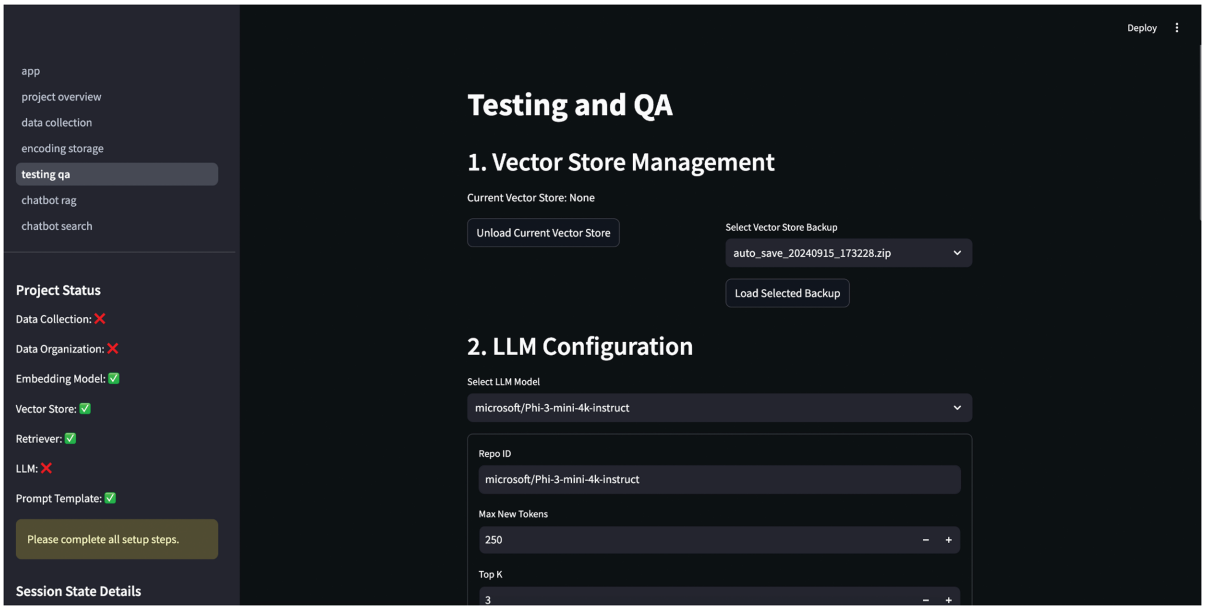


Figure 9: Screenshot of testing question and answer (QA).

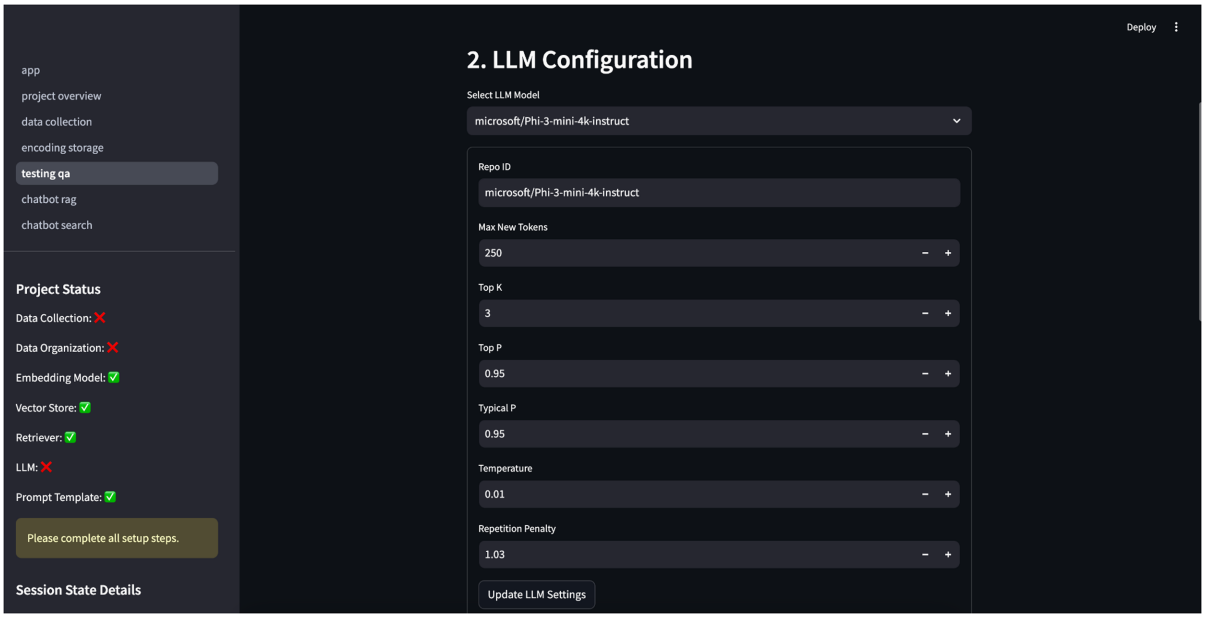


Figure 10: Screenshot of configuration details.

- **Model Configuration Parameters**: This includes the following:
  1. Repo ID
  2. Max New Tokens - the maximum number of tokens the model should generate in the response
  3. Top K - the top "k" number of responses
  4. Top P - smallest set of tokens whose cumulative probability is greater than p
  5. Typical P - a variation of top p in which the model aims to select those tokens whose probability is closest to the expected
  6. Temperature - randomness
  7. Repetition Penalty - this parameter throws a penalty at the model for generating the same token multiple times.

Once these values are configured, we can update the prompt template to proceed to "Ask Questions" (Figure 11). This is a crucial step for ensuring the LLM's output is structured and maintained. The user can fine-tune the

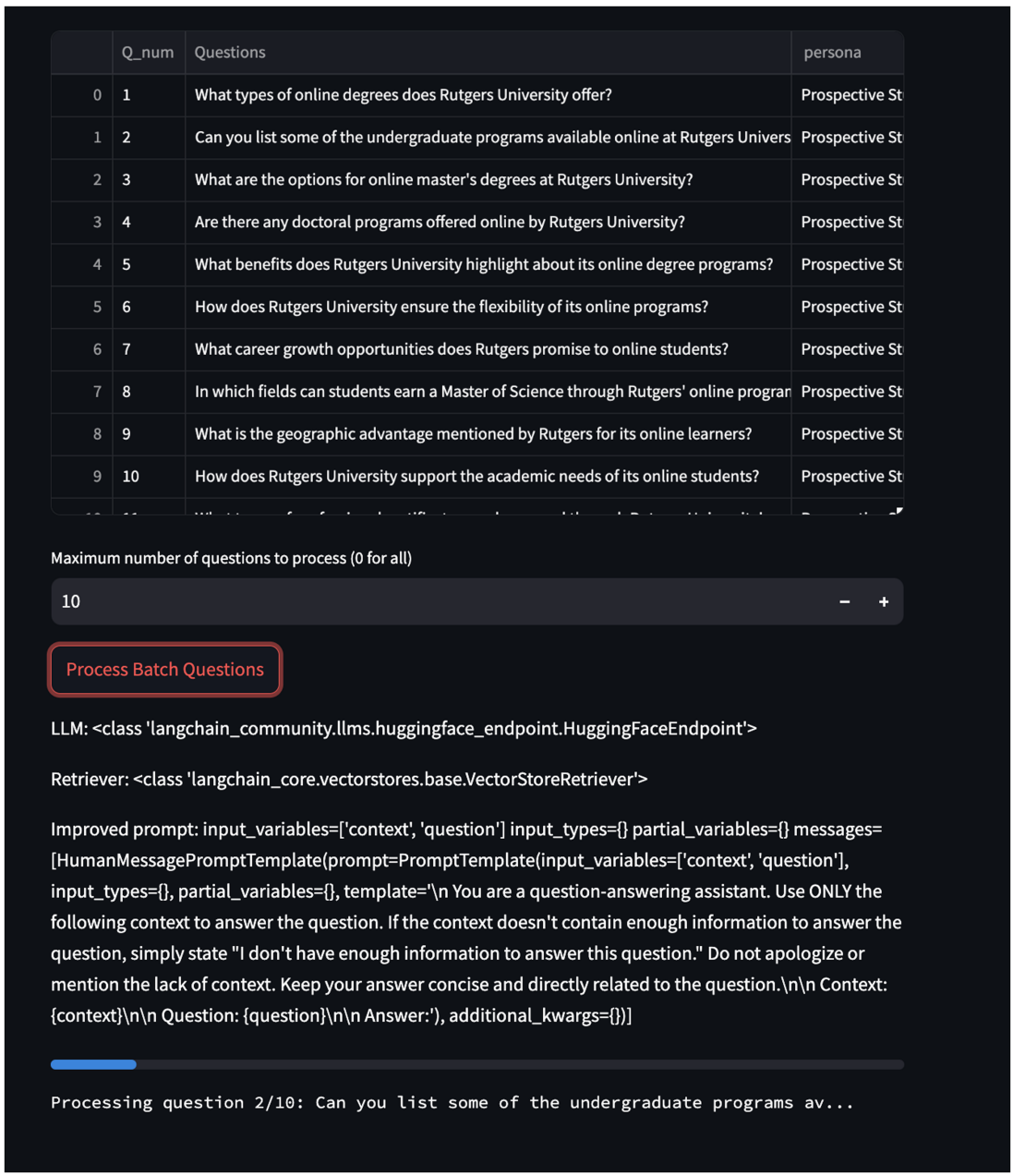


Figure 11: Screenshot of questioning the chatbot.

template based on the requirements. The next step is to enter the question and proceed to check the responses. An optional step is to pursue “Batch Processing” (Figure 12), a process that enables a computer to efficiently manage and process large volumes of data simultaneously (Runalloy 2024). In this step, we upload the Questions CSV file and, optionally, the Prompts CSV file, and proceed to process all the batch questions. This step will ensure that we have high-speed processing and highly efficient workflows with high accuracy and minimal error.

Sixth Page: This page provides information on the RAG-based chatbot interaction and can be accessed via the options (Figure 13). The final page is the chatbot interface, in which we will be conducting long-form testing of the RAG chatbot, allowing for context-aware conversations by storing the history of conversations. This interface will perform similar operations to that of the Testing QA page, except the Testing QA page is more like a question-answer interface, whereas the main purpose of the chatbot RAG page is for testing longer conversations.

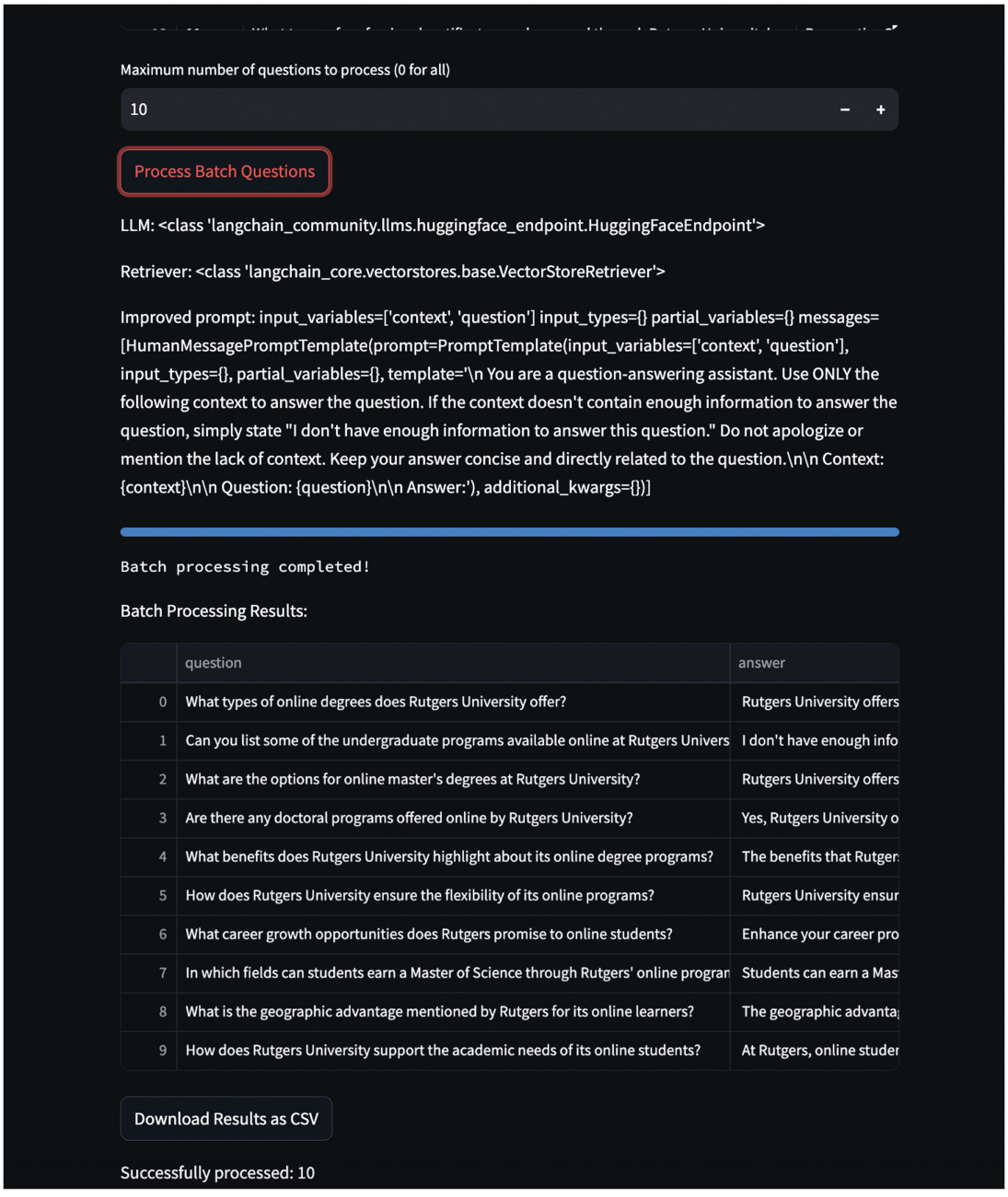


Figure 12: Screenshot of batch processing workflow.

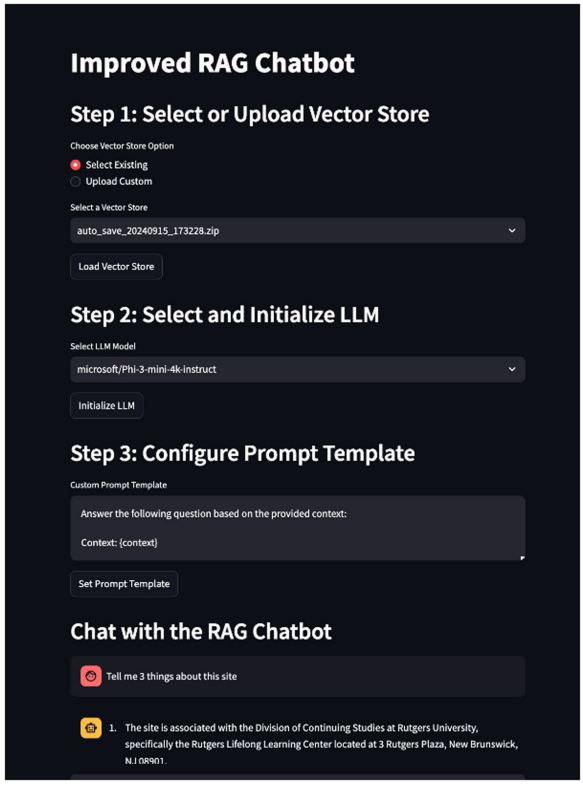


Figure 13: Screenshot of chatbot retrieval augmented generation (RAG).

- Step 1: "Select Existing"—vector stores populated with data or "Upload Custom"—upload your own dataset or vector store.
- Step 2: Select and initialize the LLM
- Step 3: Configure the prompt template
- Step 4: Chat with the RAG chatbot—this chatbot will retrieve relevant information from the vector store based on the user's input.

The chatbot search will not only give a detailed answer to the question but will also give information about the Input and JSON data related to it. This step-by-step process provides a more structured way to configure and evaluate the RAG chatbot, ensuring that it performs well in more complex and long-term conversational scenarios. An example of the output is provided in Figure 14.

3.4. Discussion—Strengths and Weaknesses

Gen AI-based applications have gained prominence in recent years, and there have been numerous calls for addressing the risks associated with Gen AIs (Golda et al. 2024; Samuel 2021). The GenAI-USS system is no exception to the systemic challenges faced by generative applications, hence, we focus our discussion here on application-specific strengths and weaknesses. The strengths include better transparency, valuable flexibility, and choices for users, along with improved accuracy of responses, the use of a plug-and-play architecture, integration of real-time information, information processing visibility, hallucination mitigation, and higher overall quality of AI-generated output. The known uncertainties and weaknesses are based on the fact that this is a prototype system and is yet to be subject to open use by students, staff, faculty, and other university stakeholders. These include a lack of clarity on the comprehensiveness of the GenAI-USS system, the potential for hallucination, the potential inability to follow prompts that lead to accurate information stated in a wrong style or tone, the potential for misuse, and the risks associated with the information contained in the LLM that the application is built on. One way to address some of these weaknesses is to run more automated and bulk processing for chat testing, with human validation on a broad range of possible input queries. It must be noted that the architecture supports the testing of new collections and related materials. Key advantages of our design are transparency, testing capabilities, and options for the use of

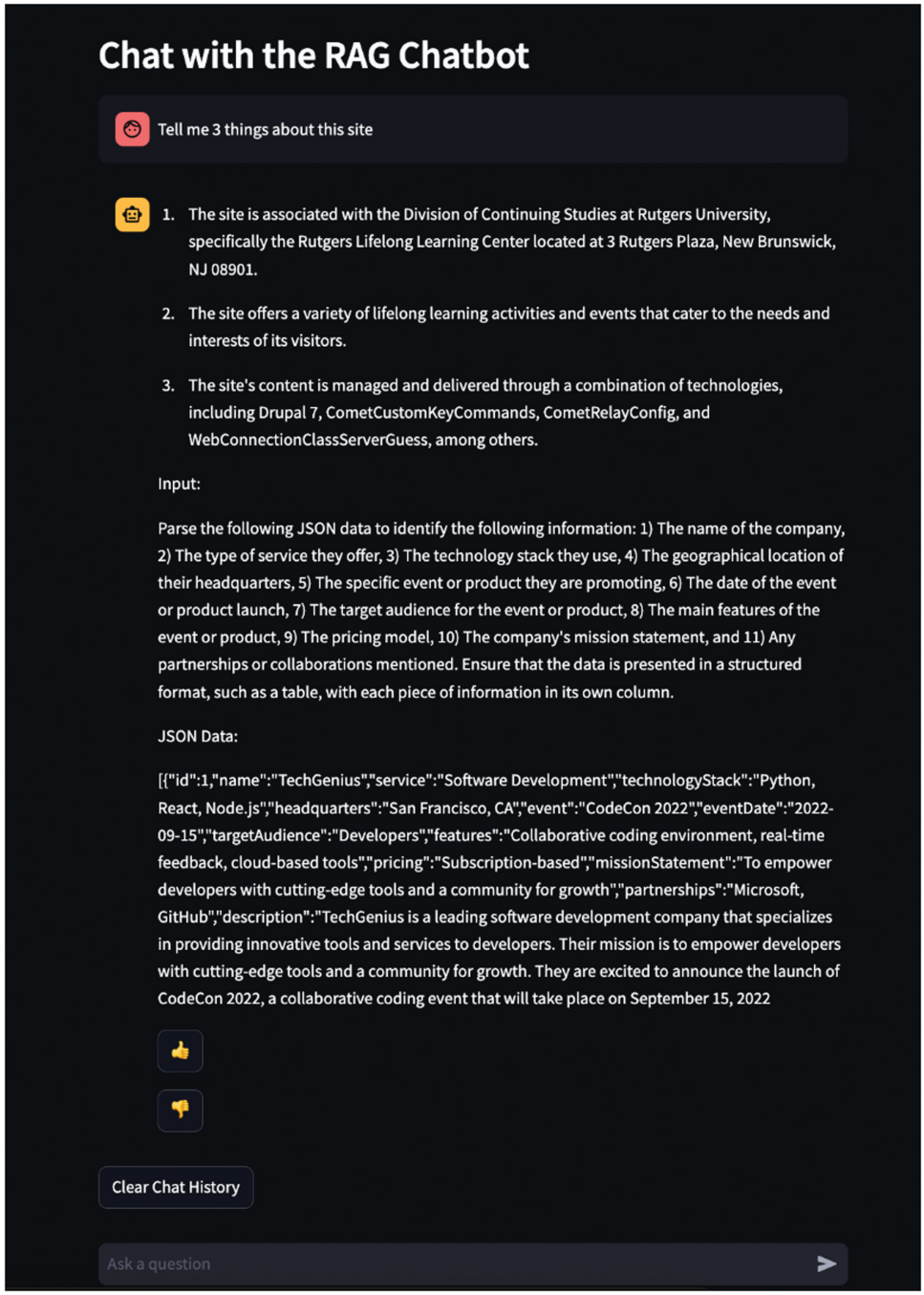


Figure 14: Screenshot of retrieval augmented generation (RAG) chatbot response.

open-weight LLMs. This implies that we can extensively test via QA source material for topical completeness, run tests on prompts to see if they guide the LLM to improved answers, test and compare LLMs for performance and quality, test and compare combinations of embedding models and LLMs, and ability to change basic embedding model properties and check with QA performance and relevancy changes. Other strengths include standard RAG benefits, including the ability of the system to pull somewhat relevant information from the LLM or be designed to have the option to say that it does not have a satisfactory answer to the query if the RAG database does have the necessary information to respond to the query. Furthermore, the testing mechanism allows us to assess if the dataset or the vector would have the information necessary for a conversation.

4. Future Research

Our current focus is on accessing information through website URLs. In the future, we plan to expand our scope by including various forms of text-based media, image-based media, audio-based media, and video-based media. Each media type serves a different purpose, and it can be chosen based on the user's preference. Future research also needs to address LLM quality evaluation and task appropriateness. Such evaluation needs to include improved measures for the accuracy of responses, coherence, consistency, and linguistic fluency. Specifically, we intend to articulate measures to indicate whether the vector embeddings capture semantic similarities, whether a correct chunking strategy has been chosen, and evaluate whether the prompts generate meaningful outputs across different contexts. Such measures can guide future research by providing ways to assess the LLM's performance in both technical and output evaluation dimensions. This work establishes a framework for scalable, cost-effective AI deployment in academic settings through optimized model selection and performance benchmarking. The system in future work will enable the generation of comprehensive testing datasets through LLM-generated synthetic QA data, expanding our evaluation capabilities. In addition, apart from quality improvement and development evaluation measures, future research needs to address incorporating new features into the user dashboard, such as sentiment scores on likely public perception, novel information classification mechanisms, and popularity of generated content (Rahman et al. 2021; Ali et al. 2021; Samuel 2018; Garvey et al. 2021). Finally, it would be valuable for future iterations of the vector store to be tested and verified for customized RAG chatbots. Validated vector stores are the source of knowledge for RAG-aware chatbots. It is anticipated that future research will lead to increased personalization and adaptivity to individual users, groups, and organizations.

5. Conclusion

Our design strategy for GenAI-USS facilitates an excellent approach to enhanced transparency and user-friendly flexibility of options, along with notable improvements to the AI-generated output via targeted information retrieval, hallucination mitigation, accuracy improvement, and timely data updates. On the submission of a query, our RAG-dependent system first identifies the most relevant information from the specialized expert knowledge database and then factors this into the Gen AI response development process with very high levels of transparency and a segment for testing, all of which cumulatively leads to vastly improved levels of effectiveness and user satisfaction. Our tests with the beta version produced highly positive results on qualitative human evaluation of generated output. Given the rapid rate of change in the field of AI and NLP technologies, we anticipate major updates to LLM and RAG models and architectures. The modular and open-weight (or open source) LLM-based approach we have used is expected to enable us to adapt to the presently foreseeable arena of upcoming technological changes and advancements. Thus GenAI-USS has the potential to serve as a transparent and flexible RAG system with scope for additional improvements of AI-generated output. As emphasized by Samuel et al. 2024, We hope that GenAI-USS will be a part of the wave of AIs that usher in the broadly anticipated "new era of artificial intelligence," leading to a better quality of life for all.

References

Alan, A. Y., E. Karaarslan, and Ö. Aydin. 2024. "A RAG-Based Question Answering System Proposal for Understanding Islam: MufassirQAS LLM." Preprint, submitted March 2025. https://doi.org/10.48550/arXiv.2401.15378

Ali, G. M. N., M. M. Rahman, M. A. Hossain, M. S. Rahman, K. C. Paul, J. C. Thill, et al. 2021. "Public Perceptions of COVID-19 Vaccines: Policy Implications from US Spatiotemporal Sentiment Analytics." *Healthcare* **9**, no. 9: 110 MDPI.

Anderson, R., C. Scala, J. Samuel, V. Kumar, and P. Jain. 2024. "Are Emotions Conveyed Across Machine Translations? Establishing an Analytical Process for the Effectiveness of Multilingual Sentiment Analysis with Italian Text." *Journal of Big Data and Artificial Intelligence*, **2**, no. 1: 57–73. doi: 10.54116/jbdai.v2i1.30

Aquino, S. 2024. "What is RAG: Understanding Retrieval-Augmented Generation." Qdrant. Accessed November 12, 2024. https://qdrant.tech/articles/what-is-rag-in-ai/

AWS. 2024. "Start Building on AWS Today." Cloud and Platform Services. Accessed November 12, 2024. https://aws.amazon.com/

Bhavsar, P. 2024. "Mastering RAG: How to Select an Embedding Model." Galileo. Accessed May 6, 2024. https://www.rungalileo.io/blog/mastering-rag-how-to-select-an-embedding-model#:~:text=Embeddings%20refer%20to%20dense%2C%20continuous

Cheng, Y., J. Chen, Q. Huang, Z. Xing, X. Xu, and Q. Lu. 2024. "Prompt Sapper: A LLM-Empowered Production Tool for Building AI Chains." *ACM Transactions on Software Engineering and Methodology* **33**, no. 5: 1–24. doi: 10.1145/3638247

CloudFare. 2024. "Connect, Protect, and Build Everywhere." Cloud and Cybersecurity Services. Accessed November 12, 2024. https://www.cloudflare.com/

Dang, H., L. Mecke, F. Lehmann, S. Goller, and D. Buschek. 2022. "How to prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models." Preprint, submitted September 3. https://doi.org/10.48550/arxiv.2209.01390

Das, S. 2024. "Exploring the Role of Large Language Models in Education. ELearning Industry." Accessed November 12, 2024. https://elearningindustry.com/exploring-the-role-of-large-language-models-in-education

de Fonseca, F. P. C., I. Paraboni, and L. A. Digiampietri. 2023. "Contextual Stance Classification Using Prompt Engineering." *Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology*, Belo Horizonte, Brazil, SBC, September 25–29. https://doi.org/10.5753/stil.2023.233708

Deng, K., G. Sun, and P. C. Woodland, 2024. "Wav2Prompt: End-to-End Speech Prompt Generation and Tuning for LLM in Zero and Few-shot Learning" Preprint, submitted June 1. https://doi.org/10.48550/arxiv.2406.00522

Desmond, M., and M. Brachman. 2024. "Exploring Prompt Engineering Practices in the Enterprise." Preprint, submitted March 13. https://doi.org/10.48550/arxiv.2403.08950

Duan, J., H. Cheng, S. Wang, A. Zavalny, C. Wang, R. Xu, et al. 2023. "Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models." Preprint, submitted May 2024. https://doi.org/10.48550/arXiv.2307.01379

Ekin, S. 2023. "Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices." Accessed June 20, 2023. 681648.pdf (d197for5662m48.cloudfront.net)

ElementX. 2024. "Enhancing Education with RAG: How Universities Can Benefit." Accessed November 12, 2024. https://www.elementx.ai/post/enhancing-education-with-rag

Gao, Y., Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, et al. 2024. "Retrieval-Augmented Generation for Large Language Models: A Survey." Accessed April 28, 2024. https://arxiv.org/pdf/2312.10997

Garvey, M. D., J. Samuel, and A. Pelaez. 2021. "Would You Please like my Tweet?! An Artificially Intelligent, Generative Probabilistic, and Econometric Based System Design for Popularity-Driven Tweet Content Generation." *Decision Support Systems* **144**: 113497. doi: 10.1016/j.dss.2021.113497

Glover, E. 2024. "Mistral AI: What to Know About Europe's OpenAI Rival." Built In. Accessed November 12, 2024. https://builtin.com/articles/mistral-ai

Golda, A., K. Mekonen, A. Pandey, A. Singh, V. Hassija, V. Chamola, et al. 2024. *Privacy and Security Concerns in Generative AI: A Comprehensive Survey*. IEEE Access.

Google DeepMind. 2024. "Google DeepMind." Accessed November 12, 2024. https://deepmind.google/

Herrmann, T., and J. Nierhoff. 2017. "Prompting—A Feature of General Relevance in HCI-Supported Task Workflows." *Proceedings of the 19th International Conference, HCI International 2017*, Vancouver, BC, Canada, Springer, Cham, July 9–14. https://doi.org/10.1007/978-3-319-58750-9_17

Hugging Face. 2024. "Hugging Face—On a Mission to Solve NLP, One Commit at a Time." Accessed April 29, 2024. https://huggingface.co/

IBM. 2023. "What Are Large Language Models (LLMs)?" Accessed April 29, 2024. https://www.ibm.com/topics/large-language-models

Imanuelyosi. 2022. "Deploy Your Streamlit Web App Using Hugging Face." Accessed April 29, 2024. https://medium.com/@imanuelyosi/deploy-your-streamlit-web-app-using-hugging-face-7b9cddb11148

Jasmine, K. S. 2024. "Unlocking the Power of Prompt Engineering: Diverse Applications and Case Studies." In *Transforming Education With Generative AI: Prompt Engineering and Synthetic Content Creation*, edited by Ramesh C. Sharma and Aras Bozkurt, 411–432. Hershey, PA: IGI Global. doi: 10.4018/979-8-3693-1351-0.ch020

Jiang, Z., F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, et al. 2023. "Active Retrieval Augmented Generation." Preprint, submitted Oct 23. https://doi.org/10.48550/arXiv.2305.06983

Li, J., X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen. 2023. "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models." arXiv.org, doi: 10.48550/arXiv.2305.11747

Liu, V., and L. B. Chilton. 2021. "Design Guidelines for Prompt Engineering Text-to-Image Generative Models." Preprint, submitted September 2023. https://doi.org/10.48550/arxiv.2109.06977

Lo, L. S. 2023. "The CLEAR Path: A Framework for Enhancing Information Literacy Through Prompt Engineering." *The Journal of Academic Librarianship* **49**, no. 4: 102720. doi: 10.1016/j.acalib.2023.102720

Mishra, A. 2024. "Five Levels of Chunking Strategies in RAG| Notes from Greg's Video. Medium." Accessed November 12, 2024. https://medium. com/@anuragmishra_27746/five-levels-of-chunking-strategies-in-rag-notes-from-gregs-video-7b735895694d

Muktadir, G. M. 2023. "A Brief History of Prompt: Leveraging Language Models. (Through Advanced Prompting)." Preprint, submitted November 28. https://doi.org/10.48550/arxiv.2310.04438

OpenAI. 2024. "OpenAI." Accessed November 13, 2024. https://openai.com/

Peeperkorn, M., T. Kouwenhoven, D. Brown, and A. Jordanous. 2024. "Is Temperature the Creativity Parameter of Large Language Models." Preprint, submitted May 1. https://doi.org/10.48550/arXiv.2405.00492

Peter, A. 2023. "What Chunk Size and Chunk Overlap Should You Use?" DEV Community; DEV Community. Accessed November 12, 2024. https://dev.to/peterabel/what-chunk-size-and-chunk-overlap-should-you-use-4338

Proser, Z. 2023. "Retrieval Augmented Generation (RAG)." Pinecone. Accessed April, 2024. https://www.pinecone.io/learn/retrieval- augmented-generation/.

Qiu, C., T. Tang, T. Yang, and M. J. Chen. 2024. "Learning to Generalize with atent mbedding ptimization for ew- and ero-hot ross omain ault iagnosis." *Expert Systems with Applications* **254**: 124280124280. doi:10.1016/j.eswa.2024.124280.

QuantumBlack, A. M. 2023. "Embeddings: The Language of LLMs and GenAI - QuantumBlack, AI by McKinsey - Medium. Medium; Medium." Accessed November 12, 2024. https://quantumblack.medium.com/embeddings-the-language-of-llms-and-genai-b74c2bef105a

Rahman, M. M., G. M. N. Ali, J. Samuel, X. J. Li, K. C. Paul, P. H. J. Chong, et al. 2021. "Socioeconomic Factors Analysis for COVID-19 US Reopening Sentiment with Twitter and Census Data." *Heliyon* **7**, no. 2:e06200.

Rathod, J. D., and G. V. Kale. 2024. "Systematic Study of Prompt Engineering." *International Journal for Research in Applied Science & Engineering Technology* **12**, no. VI: 597–613. doi: 10.22214/ijraset.2024.63182

Runalloy. 2024. "What Is Batch Processing? Definition, Use Cases, and Alternatives." Runalloy.com. Accessed November 12, 2024. https://runalloy.com/blog/what-is-batch-processing/

Rutgers University. 2001. Rutgers University. https://www.rutgers.edu/

Samuel, J., M. M. Rahman, G. M. N. Ali, Y. Samuel, A. Pelaez, P. H. J. Chong, et al. 2020a. "Feeling Positive About Reopening? New Normal Scenarios from COVID-19 US Reopen Sentiment Analytics." *IEEE Access*, **8**, 142173–142190.

Samuel, J., G. M. N. Ali, M. M. Rahman, E. Esawi, and Y. Samuel 2020b. "Covid-19 Public Sentiment Insights and Machine Learning for Tweets Classification." *Information*, **11**, no. 6: 314.

Samuel, J. 2018. "Information Token Driven Machine Learning for Electronic Markets: Performance Effects in Behavioral Financial Big Data Analytics." *Journal of Information Systems and Technology Management* **14**, no. 3: 371–383. doi: 10.4301/S1807-17752017000300005

Samuel, J. 2021. "A Call for Proactive Policies for Informatics and Artificial Intelligence Technologies." Scholars Strategy Network. Accessed April 23, 2024. https://scholars.org/contribution/call-proactive-policies-informatics-and

Samuel, J., R. Palle, and E. Soares. 2021. "Textual Data Distributions: Kullback Leibler Textual Distributions Contrasts on GPT-2 Generated Texts, with Supervised, Unsupervised Learning on Vaccine & Market Topics & Sentiment." *Journal of Big Data Theory & Practice* **1**, no. 1: 1–18. doi: 10.54116/jbdtp.v1i1.20

Samuel, J., R. Kashyap, Y. Samuel, and A. Pelaez. 2022. "Adaptive Cognitive Fit: Artificial Intelligence Augmented Management of Information Facets and Representations." *International Journal of Information Management* **65**: 102505. doi: 10.1016/j.ijinfomgt.2022.102505

Samuel, J. 2023. "The Critical Need for Transparency and Regulation Amidst the Rise of Powerful Artificial Intelligence Models. *Scholars Strategy Network* (SSN, 2023). URL: https://scholars.org/contribution/critical-need-transparency-and-regulation

Samuel, J., A. Tripathi, and E. Mema. 2024a. "A New Era of Artificial Intelligence Begins. . .Where Will It Lead Us?" *Journal of Big Data and Artificial Intelligence* **2**, no. 1: 1–4. doi: 10.54116/jbdai.v2i1.40

Samuel, J., T. Khanna, and S. Sundar. 2024b. "Fear of Artificial Intelligence? NLP, ML and LLMs Based Discovery of AI-Phobia and Fear Sentiment Propagation by AI News." Available at SSRN: https://ssrn.com/abstract=4755964

Shi, F., P. Qing, D. Yang, N. Wang, Y. Lei, H. Lu, et al. 2023. Prompt Space Optimizing Few-shot Reasoning Success with Large Language Models. doi: 10.48550/arxiv.2306.03799

Streamlit. 2024. "A Faster Way to Build and Share Data Apps." Streamlit.io. Accessed April 29, 2024. https://streamlit.io/

Tam, A. 2023. "What are large language models. MachineLearningMastery." com. Accessed November 12, 2024. https://machinelearningmastery. com/what-are-large-language-models/

UOES Rutgers. 2023. "Office of University Online Education Services." Accessed November 14, 2024. https://uoes.rutgers.edu

UOES TLT. n.d. "Our mission. Our Mission | Teaching and Learning with Technology." Accessed May 7, 2024. https://tlt.rutgers.edu/our-mission (website has since been replaced by the UOES website).

Vasilis, T. 2024. “What is Hugging Face 🤗 and why use it for NLP and LLMs? Apify Blog.” Accessed April 29, 2024. https://blog.apify. com/what-is-hugging-face/

Wang, X., Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, et al. 2024. “SCIBENCH: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models.” Accessed May 5, 2024. https://arxiv.org/pdf/2307.10635

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, et al. 2020. “HuggingFace’s Transformers: State-of-the-art Natural Language Processing.” Preprint, submitted July 14. https://doi.org/10.48550/arXiv.1910.03771

Xu, Z., S. Jain, and M. Kankanhalli. 2024. “Hallucination Is Inevitable: An Innate Limitation of Large Language Models.” Preprint, submitted February 2025. https://doi.org/10.48550/arXiv.2401.11817

Ye, Q., M. Axmed, R. Pryzant, and F. Khani. 2024. “Prompt Engineering a Prompt Engineer.” Preprint, submitted July 3. https://doi.org/10.48550/arXiv.2311.05661

Yu, P., and H. Ji. 2023. “Information Association for Language Model Updating by Mitigating LM-Logical Discrepancy.” Preprint, submitted February 2024. https://doi.org/10.48550/arXiv.2305.18582

Journal of Big Data and Artificial Intelligence

Journal of Big Data and Artificial Intelligence

https://JBDAI.org

The *Journal of Big Data and Artificial Intelligence* publishes one volume of high quality scholarly and practitioner articles on artificial intelligence (AI), informatics, data science, computer science and information science, and related topics annually, along with special issues on a rolling basis. Accepted articles are made available online for early access—to submit articles, please visit:
https://jbdai.org/index.php/JBDAI/announcement/view/5

CALL FOR MANUSCRIPTS—2025

If you would like to contribute to or join the JBDAI Reviewer or Editorial teams, Contact us: editor@jbdai.org