**Detecting Government Fraud Using Semi-Supervised Machine Learning**

Dissertation Proposal

Submitted to National University

School of Technology and Engineering

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OFPHILOSOPHY

by

WILLIAM ISAAC SAPP

San Diego, California

December 2024

**Table of Contents**

[Chapter 1: Introduction 4](#_Toc184834432)

[Statement of the Problem 6](#_Toc184834433)

[Purpose of the Study 7](#_Toc184834434)

[Introduction to Theoretical Framework 8](#_Toc184834435)

[Introduction to Research Methodology and Design (Nature of the Study) 9](#_Toc184834436)

[Research Questions 10](#_Toc184834437)

[Hypotheses 11](#_Toc184834438)

[Significance of the Study 11](#_Toc184834439)

[Definitions of Key Terms 12](#_Toc184834440)

[Summary 13](#_Toc184834441)

[Chapter 2: Literature Review 14](#_Toc184834442)

[Databases and Search Strategy 16](#_Toc184834443)

[Theoretical Framework 19](#_Toc184834444)

[Data Ethics and Legal Frameworks 25](#_Toc184834445)

[COVID-19 and The Paycheck Protection Program 28](#_Toc184834446)

[Fraud in Government Programs 32](#_Toc184834447)

[Machine Learning for Fraud Detection 36](#_Toc184834448)

[Summary 48](#_Toc184834449)

[Chapter 3: Research Method 53](#_Toc184834450)

[Research Methodology and Design Process Diagram 54](#_Toc184834451)

[Research Methodology and Design (Nature of the Study) 56](#_Toc184834452)

[Population and Sample 63](#_Toc184834453)

[Instrumentation 67](#_Toc184834454)

[Operational Definitions of Variables 73](#_Toc184834455)

[Study Procedures 78](#_Toc184834456)

[Assumptions 93](#_Toc184834457)

[Limitations 94](#_Toc184834458)

[Delimitations 96](#_Toc184834459)

[Summary 99](#_Toc184834460)

[References 101](#_Toc184834461)

**List of Tables**

**List of Figures**

[**Figure 1** *Classification Through Clustering Framework* 22](#_Toc198886049)

[**Figure 2** *COVID-19 Deaths, by Week, in The United States* 30](#_Toc198886050)

[**Figure 3** *Labor Market Policy Responses to Recessions the United States and Europe* 33](#_Toc198886051)

[**Figure 4** *PPP Loan Application Process Flowchart* 36](#_Toc198886052)

[**Figure 5** *Logistic Regression Response Function* 38](#_Toc198886053)

[**Figure 6** *Graphic Representation of a Random Forest Decision Tree* 39](#_Toc198886054)

[**Figure 7** *SVM Hyperplane* 40](#_Toc198886055)

[**Figure 8** *Illustration of K-Means Clustering Method* 41](#_Toc198886056)

[**Figure 9** *Graphic Depiction of DBSCAN Clustering* 42](#_Toc198886057)

[**Figure 10** *Hierarchical Clustering Dendrogram* 43](#_Toc198886058)

[**Figure 11** 55](#_Toc198886059)

[**Figure 12** 80](#_Toc198886060)

# Chapter 1: Introduction

The Paycheck Protection Program (PPP), signed into law in April 2020, was designed to provide financial relief to American small businesses during the early stages of the COVID-19 Pandemic (Demko et al., 2021). The PPP enabled small businesses, including sole proprietorships and self-employed persons, to take low-interest business loans to continue paying their employees and covering other business operating costs during the widespread lockdowns put in place at the time. The PPP facilitated nearly $660 Billion in uncollateralized loans, approximately 90% of which were ultimately forgiven (Emmons & Dahl, 2022).

While many government subsidy or financial assistance programs are vulnerable to fraudulent activity, initial reporting suggests that at least $200 Billion in forgiven PPP loans were likely fraudulent (USSBA OIG, 2023). Government fraud investigations are primarily informed by whistleblowers and conducted by attorney investigators (Council of the Inspectors General on Integrity and Efficiency, 2011). Government agencies are increasingly interested in employing machine learning within their oversight framework (West, 2021). Although the use of machine learning tools in fraud detection is well-established within the financial sector, published government applications outside Medicare fraud are limited.

Two machine learning tasks are supervised and unsupervised learning. Within the fraud detection domain, supervised machine learning typically involves using a training set of previously identified or labeled data that is flagged fraudulent or not fraudulent. This labeled training dataset is used to train a fraud detection model, which will ultimately be used against unlabeled data to determine whether it is fraudulent or not. The quality of the training data is largely deterministic of the quality of the trained algorithm. For example, the bias introduced in the training data (e.g., the training data only contains fraudulent loan applications from a particular minority group) will be present in the model and applied to any future data analyzed by the algorithm (Dridi, 2022a).

Unsupervised machine learning does not require previously labeled training data for model development. Unlike supervised machine learning, unsupervised machine learning tools can identify relationships and discover insights about a dataset without the supervision of training or prior knowledge of independent variables. In fact, one subset of unsupervised machine learning, principal components analysis (PCA), is used to identify the most influential independent variables within a dataset. Other common forms of unsupervised machine learning include clustering, which groups similar records together, and anomaly detection, which identifies outliers (Dridi, 2022b).

Supervised, semi-supervised, and unsupervised ML techniques hold promise for detecting PPP fraud. Machine learning is used extensively in the financial fraud detection sector (Ashtiani & Raahemi, 2022). Financial institutions and other private sector organizations regularly employ supervised machine learning techniques to detect fraudulent credit card |transactions, fraudulent loan applications, and other electronic transactions (A. Ali et al., 2022). Similarly, the healthcare sector uses unsupervised learning to identify anomalies which aid in diagnosis and treatment plans (Nassif et al., 2021). While typical fraud detection using ML employs supervised machine learning which rely on training datasets, unsupervised machine learning does not require labeled data. Given the hundreds of billions in PPP funds and the lack of previously labeled fraudulent training data (Bailey et al., 2021), there is a growing need for fraud detection procedures that are still effective when employed using previously unstudied PPP loan application data. In this case, a semi-supervised approach using unsupervised machine learning can be employed to cluster loan applications and identify anomalous loan applications which could then be flagged for further investigation.

## Statement of the Problem

The problem to be addressed in this study is the lack of intelligent unsupervised fraud identification in the government domain, specifically in the Paycheck Protection Program. As outlined by Ali et al. (2022) and Minastireanu & Mesnita (2019), a substantive number of studies focus on supervised machine-learning techniques, resulting in a notable gap in the literature on unsupervised learning in fraud detection. Similarly, Ali et al. (2022) and Dridi (2022a) note that much of the recently published work on fraud detection using machine learning focuses on the credit card or healthcare fraud domains.

Specifically within the government fraud domain, fraud within the United States PPP is a high-profile issue that affects the integrity of the program and the welfare of the American economy. According to Bailey et al. (2021) and Ma & McKinnon (2020), over five hundred billion dollars in PPP loans were disbursed with little regulatory or oversight controls put in place. Initial law enforcement investigations revealed pervasive fraud in the program. Bailey et al. ( 2021) suggest that completed PPP fraud investigations and data from previous relief programs (e.g., transfer learning) could be used to develop machine learning models aimed at detecting PPP fraud. However, fraud detection is typically conducted using supervised machine learning techniques that rely on labeled data, which depends on Department of Justice (DOJ) indictments or court filings for PPP fraud. Reliance on the limited number of confirmed fraud cases will introduce models trained using only the high-profile cases DOJ investigated. This may introduce significant bias, potentially resulting in a high false negative rate (Benala & Tantati, 2022). Short of an effective and unbiased fraud detection scheme, PPP loan fraudsters will likely continue to benefit at the expense of the United States taxpayer. Additionally, without precise tools to detect PPP fraud, opponents of the program are free to make baseless claims against its general effectiveness and use them as ammunition to oppose future disaster relief funding.

## Purpose of the Study

The purpose of this study is to develop intelligent semi-supervised fraud identification methodologies in the government domain, specifically in the Paycheck Protection Program. The practical application of semi-supervised learning for fraud detection within the PPP will increase program integrity and public trust in disaster relief efforts Additionally, there is an opportunity to apply models developed using PPP data to similar government program datasets via transfer learning. Unsupervised machine learning algorithms do not rely on labeled training data; therefore, the lack of reliably identified fraudulent data within a real-world dataset is not a limiting factor, as identified by Benala & Tantati (2022).

While previous work incorporating Unsupervised ML with PPP loan fraud detection focused on risk mitigation for financial institutions using proprietary methods and non-public financial records such as clustering using existing bank records for businesses with similar profiles (Crowe, n.d.), this study presents a novel approach to PPP fraud detection using a classification through clustering methodology intended to support government-led fraud detection. This study will use the publicly available, open-access PPP Loan dataset to evaluate the effectiveness of unsupervised learning techniques such as clustering, anomaly detection, and PCA. While the population for this study is the complete record of PPP loan applications, this study will limit analysis to an (albeit large) sample of PPP loan applications over $150k. This sample reflects the data population, including all instances over the predetermined threshold, resulting in nearly one million records. This study leverages secondary data, is location agnostic, and will be conducted at the researcher's home location.

Data gathering and collection will consist of downloading the open-access dataset from its government data publisher, in this case the Small Business Administration (SBA). Once data is loaded and requisite transformations and normalizations are conducted, it will be analyzed using a series of unsupervised machine learning algorithms encompassing PCA, clustering, and anomaly detection. Additionally, the limited number of labeled fraudulent applications will be used as part of a training dataset to conduct supervised machine learning and binary classification for cross-validation.

While the dependent variable in this study is the binary classification likely fraudulent vs not likely fraudulent, independent variables will include loan application features such as dollar amount, credit history, or employment status. The finalized list of independent variables will be determined as a result of PCA. For example, unsupervised clustering will group fraudulent and non-fraudulent activity. The measures of effectiveness for unsupervised learning models will occur in two methods; clustering effectiveness will be measured with common unsupervised evaluation metrics such as cohesion and separation, while the small number of previously labeled data (e.g., from guilty pleas and verdicts on alleged fraud) enable measuring of typical supervised machine learning metrics such as precision and recall (Dridi, 2022b)

## Introduction to Theoretical Framework

This study will utilize the classification through clustering framework proposed by (López et al., 2012). In their work, the authors conducted classification via unsupervised machine learning using clustering algorithms to sort records into a finite number of clusters, enabling labeling or classification of the data. In addition to clustering, the authors also conducted PCA to reduce the dimensionality of the dataset.

Through the classification through clustering framework, previously unlabeled PPP loan data can be analyzed for both key independent variables (i.e., via PCA) and sorted into clusters based on likely fraudulent and likely non-fraudulent applications. The fraud identification problem is ultimately a binary classification problem (i.e., fraudulent vs non-fraudulent), which is typically performed using a training dataset and supervised machine learning algorithms such as logistic regression. However, the noted lack of previously labeled data, as identified Bailey et al. (2021), precludes the use of effective binary classification via supervised machine learning alone.

## Introduction to Research Methodology and Design (Nature of the Study)

This study will use a quantitative experimental design to identify key features and methodologies to identify PPP loan fraud using unsupervised and supervised machine learning. According to Barroga & Matanguihan (2022), quantitative research is the appropriate methodology for comparing relationships among variables. Similarly, experimentation is the most efficient research design to establish cause and effect between independent and dependent variables (Stoner et al., 2023). By systematically manipulating independent variables, such as clustering techniques or preprocessing methods, while measuring their effects on fraud detection accuracy, experimentation provides a structured approach to validate hypotheses. This design enables the evaluation of specific methodologies, such as the integration of PCA and semi-supervised learning, under controlled conditions to assess their impact on model performance. Furthermore, experimentation ensures reproducibility, allowing this study to contribute practical insights to both the theoretical development of semi-supervised learning and its application in government fraud detection.

In this study, existing unsupervised learning algorithms and methodologies such as clustering, anomaly detection, and PCA will be the independent variables, while the dependent variables are the resulting performance metrics when applied to the PPP loan dataset. Additionally, unsupervised machine learning model effectiveness will be compared to previously identified PPP loan fraud using supervised learning performance metrics. The quantitative research methodology is appropriate as the goal is to evaluate the relationship among variables, in this case, the relationship between the various unsupervised machine learning models and the performance evaluation metrics when applied to the PPP loan dataset. Experimentation will enable the control and manipulation of the independent variable (the choice of algorithm) to achieve the highest performance evaluation metrics.   
 This quantitative experimental study will use publicly available PPP loan data and existing unsupervised and supervised machine learning algorithms to detect fraudulent PPP loan applications. Collection of this data will be performed electronically via direct download from the United States Small Business Administration website. All experimentation will be conducted within a Google Colab environment using Python.

## Research Questions

### RQ1

What are the key features or variables associated with fraudulent loan applications within the PPP?

### RQ2

What novel combination of existing unsupervised and supervised learning models can effectively identify fraudulent activity within the PPP?

## Hypotheses

### H10

PPP Loan applications are best clustered and further classified given the complete list of values of each feature or variable in the dataset.

### H1a

PPP Loan applications are best clustered and further classified given the values of specific features or variables in the dataset.

### H20

All combinations of unsupervised and supervised models perform identically in detecting fraud: Model1=Model2=Modelk.

### H2a

Not all unsupervised learning models used in conjunction with supervised learning models perform identically in detecting fraudulent activity. At least two model combinations differ.

## Significance of the Study

Much of the published research on ML-driven fraud detection relies on supervised ML given a robust training dataset (Dridi, 2022a). When there is a significant imbalance in the dataset (i.e., when there is a disproportionate number of records labeled not fraudulent versus fraudulent) several imbalance compensation techniques are typically applied to reduce their effects on algorithm development. However, when presented with a unique dataset in an underrepresented fraud detection domain (e.g., government fraud) solely supervised ML techniques include either incorporating unrelated training data from previous fraudulent activity investigations (Bailey et al., 2021), whereas solely unsupervised ML often require significant third party data such as bank records to increase dimensionality and variability (Crowe, n.d.). Incorporating unsupervised ML for anomaly detection, PCA, and clustering for fraud detection using solely the government published PPP loan dataset in conjunction with supervised learning techniques, and then comparing findings to the limited known fraudulent activity (as determined by fully prosecuted cases) will provide the framework for a novel classification through clustering fraud detection methodology not limited by the traditional requirement of training datasets.

## Definitions of Key Terms

### 18 U.S.C. §§ 1001, 1342

Title 18, United States Code (U.S.C.) contains the laws of the United States pertaining to crimes. § 1001, Statements or Entries Generally, covers false statements made in matters involving each branch of government (*18 USC 1001: Statements or Entries Generally*, 2004). § 1342, Fraud by Wire Radio or Television, covers the use of electronic means to defraud or obtain money under false pretense (*18 USC 1343: Fraud by Wire, Radio, or Television*, 2008).

### Paycheck Protection Program

Funded by congress in the spring of 2020, the Paycheck Protection Program (PPP) made available nearly $670 billion to eligible small business via government backed loans which could eventually be partially or completely forgiven (Humphries et al., 2020).

### Unsupervised Machine Learning

Unsupervised machine learning is a form of machine learning used to identify patterns and anomalies in data without the need for previously labelled data (Dridi, 2022b).

### Semi-Supervised Machine Learning

Semi-supervised learning combines elements of both supervised and unsupervised learning, which is useful when a small set of labeled data is available, but most of the data remains unlabeled (Dridi, 2022a).

### Supervised Machine Learning

Supervised machine learning is a form of machine learning used to identify patterns and anomalies in a dataset. Supervised machine learning encompasses a series of algorithms which are trained against previously labelled data (Dridi, 2022a).

## Summary

There is extensive research exploring the application of supervised machine learning techniques to aid in fraud identification in the private and financial sectors. However, since supervised machine learning requires previously labelled data, these techniques will not be effective against a novel dataset or fraud identification problem. To address this gap in literature this study will investigate the application of semi-supervised machine learning techniques to aid in fraud detection using the PPP loan dataset. Comparing results of various supervised adnd unsupervised machine learning algorithms using established measures of effectiveness, this study aims to develop a novel methodology for fraud identification when presented with an unlabeled dataset.

# Chapter 2: Literature Review

The purpose of this literature review is to establish a foundational understanding of machine learning methodologies, ethical considerations, and the unique operational context required for effective fraud detection within the Paycheck Protection Program (PPP). The rapid deployment of PPP funds during the COVID-19 pandemic exposed significant vulnerabilities within public sector programs to fraud and misuse. This study specifically addresses these challenges by exploring semi-supervised machine learning models tailored to the imbalanced datasets characteristic of government-administered financial aid programs, which frequently feature scarce instances of fraudulent activity relative to the vast number of legitimate claims (Itri et al., 2019; Zhao et al., 2024).

This chapter explores relevant literature across key areas to address the complexities involved in this type of fraud detection. These areas include foundational data science methodologies, theoretical frameworks that guide the interpretation and design of fraud detection models, ethical and legal considerations necessary to ensure responsible AI deployment, and the specific operational context of the PPP. Each section builds upon a synthesis of studies, integrating key findings and highlighting critical research gaps that inform the methodology choices outlined in Chapter 3 (Debener et al., 2023; Gui et al., 2024; Rixom et al., 2021).

The structure of this literature review is organized as follows:

1. **Theoretical Framework**: This section examines clustering techniques and dimensionality reduction methods as applied to fraud detection, particularly within imbalanced datasets. The Fraud Triangle Theory and its expanded versions offer insights into the behavioral drivers of fraud, connecting technical approaches to behavioral risk indicators and guiding the study's design (Awang et al., 2020).
2. **Data Ethics and Legal Frameworks**: Ethical considerations, such as fairness, accountability, and privacy, are critical when deploying machine learning in public sector contexts. This section discusses the implications of frameworks like GDPR and CCPA, ensuring the study aligns with legal standards and ethical expectations in government applications (Emilio Ferrara, 2023; Koreff et al., 2023).
3. **COVID-19 and the Paycheck Protection Program**: The PPP’s rapid deployment during the pandemic provides a unique case study of fraud vulnerabilities in government-administered relief programs. This section addresses the inherent risks and operational challenges within the PPP, using comparative studies to highlight the need for fraud detection strategies adaptable to high-volume, high-urgency scenarios (Bozza, 2024; Miller & Bertozzi, 2024).
4. **Fraud in Government Programs**: Broadening the context, this section reviews challenges across other public sector programs, discussing macro-level governance and meso-level organizational controls that influence fraud detection. This comparison underscores the importance of adapting fraud detection methods to meet the regulatory and operational needs specific to public sector programs (A. Ali et al., 2022; King et al., 2023).
5. **Machine Learning for Fraud Detection**: This section examines the application of supervised, unsupervised, and semi-supervised learning techniques for fraud detection, with an emphasis on semi-supervised models that effectively handle the imbalanced data typical in fraud cases. Evaluation metrics are reviewed to clarify performance considerations specific to imbalanced datasets in fraud detection (Debener et al., 2023; Gui et al., 2024).
6. **Summary**: The chapter concludes with a synthesis of research gaps, setting up a direct lead-in to Chapter 3, where these insights inform the selection of methodologies. This includes the rationale for focusing on semi-supervised models and clustering techniques, which align with the identified challenges and operational demands of fraud detection within the PPP.

To ensure a rigorous approach, this literature review relied on peer-reviewed studies and comprehensive database searches that included EBSCOhost, ProQuest, Google Scholar, and arXiv. This multi-source approach supports a balanced review, incorporating established methodologies and emerging trends to ground this study’s design choices in theoretical and practical insights.

## Databases and Search Strategy

To ensure a comprehensive and academically rigorous literature review, this study employed a multi-database search strategy, drawing on diverse sources to cover machine learning methodologies, fraud detection in government programs, and data ethics frameworks. Primary databases accessed include the Northcentral University (NCU) Library with EBSCOhost and ProQuest platforms, Google Scholar, and arXiv for preprints and cutting-edge research. Each database contributed unique insights, with academic journals providing validated studies and arXiv supporting recent developments in machine learning. This multi-source approach ensured a well-rounded review that includes both foundational theories and emerging trends in fraud detection.

### Search Terms and Keywords

The following search terms and combinations were used to ensure a thorough review across each main section of this chapter. Specific terms were selected based on relevance to fraud detection, machine learning, and government program oversight, with search parameters customized by section and sub-section to target the most pertinent literature.

**1. Data Science Methodologies for Fraud Detection**

* **Supervised Learning Techniques**:
  + Search terms: “supervised learning for fraud detection,” “classification in fraud detection,” “decision trees fraud detection,” “random forests in fraud detection,” “logistic regression for fraud”
* **Semi-supervised Learning Techniques**:
  + Search terms: “semi-supervised learning in fraud detection,” “imbalanced data semi-supervised,” “semi-supervised machine learning fraud,” “credit card fraud semi-supervised learning,” “PPP fraud semi-supervised models”
* **Unsupervised Learning Techniques**:
  + Search terms: “unsupervised learning fraud detection,” “clustering for fraud detection,” “anomaly detection unsupervised learning,” “autoencoders fraud detection,” “PCA in fraud detection”

**2. Theoretical Frameworks**

* **Clustering and Dimensionality Reduction**:
  + Search terms: “clustering for fraud detection,” “K-means fraud detection,” “hierarchical clustering in finance,” “dimensionality reduction techniques fraud,” “principal component analysis (PCA) fraud detection”
* **Fraud Triangle and Related Theories**:
  + Search terms: “Fraud Triangle theory,” “expanded Fraud Triangle,” “behavioral theories fraud detection,” “opportunity pressure rationalization fraud,” “capability as a factor in fraud”

**3. Data Ethics and Legal Frameworks**

* **Ethical Implications of Machine Learning**:
  + Search terms: “data ethics in machine learning,” “fairness and bias in AI,” “AI accountability in fraud detection,” “transparency in machine learning fraud detection,” “privacy concerns machine learning government programs”
* **Legal Standards and Regulations**:
  + Search terms: “GDPR machine learning,” “CCPA and fraud detection,” “regulatory standards in machine learning,” “data protection and government programs,” “AI regulations and compliance”

**4. COVID-19 and the Paycheck Protection Program (PPP)**

* **PPP Loan Effectiveness**:
  + Search terms: “COVID-19 PPP program,” “fraud in Paycheck Protection Program,” “PPP loan misuse,” “government relief fraud,” “financial aid fraud detection”

**5. Fraud in Government Programs**

* **Macro-Level Governance and Regulatory Challenges**:
  + Search terms: “government fraud prevention strategies,” “public sector fraud detection,” “macro-level fraud governance,” “regulatory challenges in fraud detection,” “public policy on fraud”
* **Meso-Level Organizational Controls**:
  + Search terms: “organizational controls fraud prevention,” “internal fraud detection controls,” “auditing practices public sector,” “PPP fraud detection organizational practices”

**6. Machine Learning for Fraud Detection**

* **Evaluation Metrics for Imbalanced Data**:
  + Search terms: “fraud detection metrics,” “evaluation metrics imbalanced data,” “precision-recall imbalance fraud,” “AUC-ROC fraud detection,” “confusion matrix fraud detection”

These search terms guided a systematic review process that ensured the inclusion of both foundational and emerging literature across key topics. Boolean operators (AND, OR) were applied to expand or narrow searches as needed, and filters for peer-reviewed publications and recency (primarily 2018-2024) were used to maintain relevance and academic rigor. Each selected source contributes to a layered understanding of fraud detection methods, ethical considerations, and government program oversight, grounding this study in both theoretical and applied research.

## Theoretical Framework

Fraud detection within the PPP presents distinct challenges due to highly imbalanced datasets and limited labeled instances of fraudulent activity. To address these challenges, classification through clustering serves as a primary methodological approach, leveraging unsupervised and semi-supervised learning to identify outliers that may indicate fraud. This approach clusters similar data points, enabling the detection of anomalies as potential fraud cases without the extensive labeling required by traditional supervised learning models (Gui et al., 2024; Miller & Bertozzi, 2024). Supported by dimensionality reduction techniques and the Fraud Triangle behavioral model, classification through clustering provides a robust and interpretable framework that addresses both technical and behavioral dimensions of fraud risk.

### Classification through Clustering in Fraud Detection

Classification through clustering focuses on grouping data points based on similarity, creating clusters that represent "normal" behavior patterns. In the context of fraud detection, these clusters establish baseline behaviors, while outliers—data points that deviate significantly from cluster norms—are flagged as potential fraud indicators. This technique is particularly effective for highly imbalanced datasets, where labeled fraud cases are scarce relative to the vast number of legitimate transactions (Debener et al., 2023; Itri et al., 2019).

Incorporating insights from López et al. (2012), classification through clustering achieves accurate predictive performance in semi-supervised environments, using clustering to effectively categorize outliers even with minimal labeled data. By forming clusters of similar transactions, this method can classify previously unseen cases based on their proximity to identified clusters, enhancing the model’s ability to generalize in scenarios with minimal labeled data. In the PPP, classification through clustering addresses the need for rapid and scalable fraud detection in a high-volume, low-fraud setting, effectively leveraging unsupervised learning to detect anomalies.

**Figure 1**   
*Classification Through Clustering Framework*

A diagram of a clustering process

Description automatically generated

***Note.*** *Source* (López et al., 2012)

### Clustering Techniques for Fraud Detection

K-Means Clustering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and Hierarchical Clustering are some of the most used algorithms for detecting outliers and anomalies within large datasets. These methods enable the grouping of loan applications with similar features, and any significant deviations from these clusters may suggest fraudulent behavior.

* **K-Means Clustering**: This algorithm partitions data into a predefined number of clusters based on feature similarity. In the context of the PPP, businesses with comparable payroll sizes, loan amounts, and industry classifications would be grouped together. Outliers within these clusters—such as businesses with inflated payrolls relative to their peers—would be flagged for further investigation (López et al., 2012).
* **DBSCAN**: Unlike K-Means, DBSCAN does not require the number of clusters to be specified beforehand. It detects clusters based on the density of data points, making it particularly useful for identifying smaller, densely packed clusters of fraudulent activity that might otherwise be overlooked in larger datasets (Carcillo et al., 2021).
* **Hierarchical Clustering**: This method creates a tree-like structure of nested clusters. It is useful for detecting fraud at multiple levels of granularity, helping to identify not only individual anomalies but also broader patterns of suspicious behavior within certain segments of the data (Zhao et al., 2024).

### Clustering and Principal Component Analysis (PCA) in Fraud Detection

Clustering is a fundamental technique in fraud detection, offering a means to group similar data points and identify anomalies indicative of fraudulent activities. Unlike supervised methods, clustering operates without labeled data, making it particularly valuable for applications where labeled instances of fraud are scarce. Techniques such as k-means, DBSCAN, and hierarchical clustering have been widely used in the literature.

Carcillo et al. (2021) ilustrated how clustering can be integrated into hybrid supervised-unsupervised frameworks. For example, clustering algorithms can segment loan applications into groups based on shared attributes, flagging outliers as potential fraudulent cases. These flagged cases can then be passed to a supervised classifier for further evaluation, improving overall fraud detection rates.

Dimensionality reduction, particularly PCA, enhances clustering by simplifying high-dimensional data while preserving the most significant variance in the dataset. This is especially important in datasets like those associated with the PPP, where numerous variables such as loan amount, applicant history, and lender behavior contribute to data complexity. Miller & Bertozzi (2024) demonstrated the combined use of PCA and graph-based clustering to identify anomalous patterns in high-volume datasets. By projecting data into a lower-dimensional space, PCA not only reduces computational costs but also enhances clustering performance.

**Advantages of Clustering and PCA in Fraud Detection**:

1. **Scalability**: Clustering and PCA can efficiently handle large datasets, making them suitable for high-volume programs like the PPP.
2. **Flexibility**: These methods are adaptable to different fraud scenarios, from small-scale fraud networks to systemic patterns.

**Challenges and Mitigations**:

1. **Data Quality**: Clustering accuracy depends on clean and well-preprocessed data. Feature engineering and outlier handling are critical steps.
2. **Model Interpretability**: PCA reduces interpretability by transforming original features. However, explainable AI techniques can help bridge this gap.

### Supporting Framework: The Fraud Triangle and Its Relevance

While clustering offers a data-driven basis for detecting anomalies, the **Fraud Triangle** provides a behavioral context for interpreting these outliers. The Fraud Triangle, which identifies pressure, opportunity, and rationalization as key factors driving fraud, complements classification through clustering by linking data patterns to potential motivations for fraudulent activity. In the PPP, where financial pressures from the COVID-19 pandemic and the rapid availability of funds created new fraud risks, the Fraud Triangle aids in contextualizing anomalies by associating certain deviations with underlying behavioral factors (Awang et al., 2020; Bozza, 2024).

Integrating the Fraud Triangle with classification through clustering enhances model interpretability by providing a behavioral rationale for data points that diverge from cluster norms. For instance, anomalies flagged in the clustering process that align with high-risk indicators from the Fraud Triangle may represent cases where financial pressure or opportunity contributed to potentially fraudulent behavior. This integration supports a more adaptive fraud detection model, responsive to the unique characteristics of the PPP.

### Expanded Fraud Triangle: Capability as a Factor

An expanded version of the Fraud Triangle introduces capability as a fourth factor, addressing the role of specialized knowledge or access in enabling individuals to exploit program vulnerabilities. Capability is especially relevant in the PPP context, where individuals with insider knowledge or technical skill may engage in complex fraud schemes, concealing their actions within legitimate-looking clusters. Including capability as a factor informs the interpretation of clustered outliers, recognizing that some anomalies may reflect intentional, concealed fraud strategies rather than random deviations (King et al., 2023; Zhao et al., 2024).

Incorporating capability into classification through clustering enhances the model's predictive power by connecting technical anomalies with behavioral insights, allowing for a nuanced understanding of outliers. This framework supports semi-supervised learning by adding a behavioral layer that aids in categorizing unusual data points, effectively balancing technical accuracy with contextual relevance.

### Conclusion of the Theoretical Framework

The theoretical frameworks outlined in this section—clustering techniques, dimensionality reduction, and behavioral models such as the expanded Fraud Triangle—provide a comprehensive basis for the methodological choices in this study. By combining data-driven approaches with behavioral insights, these frameworks support a nuanced model capable of identifying fraud even within highly imbalanced datasets. This integration of technical methods with behavioral context allows for a robust fraud detection approach that aligns with the ethical, operational, and regulatory considerations specific to the PPP.

## Data Ethics and Legal Frameworks

Deploying machine learning for fraud detection within government programs such as the Paycheck Protection Program (PPP) necessitates adherence to stringent ethical and legal standards. This section examines critical aspects of data ethics, including fairness, accountability, and privacy, alongside legal frameworks like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), which govern data use and model transparency.

### Ethical Considerations in Machine Learning

Ethical considerations in machine learning are especially important in public sector applications, where models impact not only individual privacy but also public trust. Issues such as fairness and bias mitigation are paramount, as machine learning models trained on imbalanced datasets may inadvertently reinforce existing biases, leading to unfair treatment of certain groups. Studies emphasize the importance of fairness-aware algorithms and bias detection in semi-supervised models, which help prevent unjust classification outcomes (Emilio Ferrara, 2023; Zhou et al., 2022). Transparency in model operations is another essential factor; explainable AI (XAI) methodologies enable stakeholders to understand and assess model decisions, particularly when potential fraud is flagged (Koreff et al., 2023).

For the PPP, achieving ethical AI deployment requires balancing model complexity with interpretability. High-performing models that lack transparency may undermine public confidence and accountability, especially in cases where decisions impact funding or legal repercussions. To address these challenges, recent developments in XAI provide interpretability techniques that make machine learning decisions more accessible to both administrators and the general public, supporting ethical deployment in government settings (Sarker, 2021; Zhao et al., 2024).

### Legal Frameworks for Fraud Detection

Compliance with legal standards such as GDPR and CCPA is fundamental for machine learning models used in government fraud detection. GDPR mandates strict guidelines for data protection, including the rights of individuals to access, rectify, or erase their data, and requires organizations to adopt privacy-by-design principles in their data processing activities. In the context of fraud detection, GDPR's principles of data minimization and purpose limitation are essential, guiding the use of personal data strictly for specific and justified purposes (Emilio Ferrara, 2023).

Similarly, CCPA enhances data privacy rights by granting consumers greater control over their personal information, including the right to opt out of data sales and request the deletion of their information. In government-administered programs like the PPP, adhering to these regulatory frameworks requires careful management of data sources and processes to ensure data use aligns with both ethical and legal standards. Privacy-aware machine learning techniques, which protect individual data during the fraud detection process, play a critical role in regulatory compliance, supporting model development that respects individual privacy while remaining effective for large-scale fraud detection (A. Ali et al., 2022).

### Explainability Requirements for AI

Transparency and explainability in AI are critical for fraud detection in government programs, where decisions carry significant implications for fairness, accountability, and public trust. Machine learning models used to detect fraud within the PPP must balance high accuracy with the need for clear, understandable explanations of their processes. This requirement is particularly challenging in semi-supervised learning, where clustering algorithms, such as K-means or DBSCAN, introduce opacity due to their reliance on unsupervised patterns and the lack of direct interpretability in cluster assignments (Koreff et al., 2023). For example, a cluster flagged as fraudulent might combine unexpected features (e.g., high loan-to-employee ratios or anomalies in business types), making it difficult for auditors or non-technical stakeholders to understand the rationale behind a flagged decision.

This opacity is problematic in government fraud detection because interpretability is essential to justify decisions that can affect individuals or businesses. For instance, in the PPP, the inability to explain why a loan was flagged for investigation could undermine public trust and lead to claims of bias or unfair targeting (Miller & Bertozzi, 2024). Addressing this challenge involves navigating the trade-offs between interpretability and model performance. Highly interpretable models, such as logistic regression, may sacrifice performance in complex fraud scenarios, while advanced models, like semi-supervised ensembles, require explainability techniques to bridge this gap (King et al., 2023).

Emerging explainability tools, such as SHapley Additive exPlanations (SHAP) and counterfactual explanations, offer enhanced insights compared to traditional methods like Local Interpretable Model-agnostic Explanations (LIME). SHAP provides consistent feature importance rankings, while counterfactual explanations help identify minimal changes required to alter a model’s decision (Koreff et al., 2023). These methods are particularly valuable in semi-supervised settings, where both labeled and unlabeled data contribute to decision-making, making model outputs inherently more complex. Regulatory and ethical frameworks further emphasize the importance of explainability, ensuring fairness in detecting fraud patterns and preventing discriminatory practices (Miller & Bertozzi, 2024). By integrating explainability tools with high-performing semi-supervised models, government programs can responsibly deploy AI systems that are both effective and accountable.

## COVID-19 and The Paycheck Protection Program

The PPP was launched in response to the economic crisis of the COVID-19 pandemic, aiming to provide quick financial relief to businesses affected by lockdowns. Administered rapidly, the PPP disbursed over $800 billion in loans, prioritizing speed over stringent vetting processes. This urgency exposed vulnerabilities, as highlighted in the SBA OIG Report 23-09, which estimated over $200 billion in potentially fraudulent PPP and EIDL loan disbursements due to reduced controls and expedited approvals (USSBA, 2023).

The SBA’s challenges reflect broader issues in fraud detection for government aid programs, where traditional models are often insufficient for high-volume, urgent contexts. These circumstances underscore the need for adaptive machine learning models, such as classification through clustering, which can identify fraud even within imbalanced data settings (Bozza, 2024; King et al., 2023). For the PPP, where legitimate applications vastly outnumber fraudulent ones, semi-supervised methods provide an efficient approach to identifying outliers and prioritizing cases for further investigation (A. Ali et al., 2022).

In adapting machine learning models to the PPP’s high-volume, low-fraud environment, this study aims to contribute to enhanced fraud detection frameworks applicable to future government relief initiatives. By implementing techniques capable of detecting subtle patterns associated with fraud, this approach aligns with the SBA’s focus on improving oversight and accountability in federal aid disbursements.

**Figure 2**   
*COVID-19 Deaths, by Week, in The United States*

*A graph of blue lines

Description automatically generated*

***Note.*** *Provisional COVID-19 Deaths, by Week, in The United States, Reported to* (CDC, 2020)

The CARES Act, signed into law in late March 2020, along with the Consolidated Appropriations Act, signed in late December 2020 (Department Of Treasury, 2024), authorized the PPP which has since provided 934 Billion dollars in federally backed loans and grants to small businesses throughout the country (The National Law Review, 2023). Small businesses, typically under 1500 employees, sole proprietorships, and other public interest organizations under 500 employees were considered eligible to take-out two-year loans from private financial institutions (i.e., banks) and could request loan forgiveness or conversion to grants as long as the usage requirements were met. Usage requirements for PPP loans included paying existing employees, rehiring recently released employees, and up to 40% on mortgage, rental, and utility costs (Sabasteanski et al., 2021).

### PPP Loan Effectiveness

While the PPP was created to provide financial lifelines to small businesses during the COVID-19 pandemic, its overall effectiveness has been the subject of debate. The program distributed over $934 billion in loans through various rounds of funding, with the goal of preventing job losses and business closures. According to the Small Business Administration (SBA), more than 11 million businesses received PPP loans (USSBA OIG, 2023). However, despite the program’s scale, questions have been raised about whether it efficiently met its objectives and whether it disproportionately benefited certain businesses.

Independent evaluations have found that while the PPP may have helped many businesses survive, it did so at a significant cost per job saved. Autor et al. (2022) estimate that the PPP preserved between 2 to 3 million jobs annually during the pandemic, but at an exceptionally high cost of $169,000 to $256,000 per job. Much of this inefficiency stemmed from the program's design, which incentivized banks to process as many loans as possible, as the loans were federally guaranteed, and banks earned a 1% interest fee for processing them. As a result, many loans went to larger and more stable businesses that were not at immediate risk of closure (Li & Strahan, 2020).

The program’s reliance on banks to disburse funds, coupled with the absence of robust administrative infrastructure, contributed to this inefficiency. The lack of targeted funding meant that businesses that were relatively unaffected by the pandemic could still qualify for loans, while many struggling small businesses faced challenges in accessing the funds quickly (Autor et al., 2022). Furthermore, studies by the Bureau of Economic Analysis (BEA) suggest that the PPP ultimately functioned as a government subsidy that shifted the financial burden of operating expenses—such as payroll, rent, and utilities—from businesses to the federal government (Bureau of Economic Analysis, 2021).

Comparative analyses also suggest that alternative approaches used by other countries may have been more efficient. For example, Giupponi et al. (2022) compared the PPP to direct wage subsidies implemented by European countries. These subsidies, which were paid directly to workers rather than funneled through businesses, proved to be a more cost-effective way to preserve jobs and avoid administrative complexities.

Despite these critiques, the PPP succeeded in preventing mass layoffs, especially in the early months of the pandemic. The program provided critical liquidity to businesses that were facing unprecedented uncertainty and helped to stabilize the labor market during a period of extreme volatility. However, the trade-off was that a substantial portion of the funds went to businesses that did not urgently need them, and many businesses that did not meet loan forgiveness requirements were allowed to convert their loans into grants, further straining the program’s financial efficiency (Li & Strahan, 2020).

**Figure 3***Labor Market Policy Responses to Recessions the United States and Europe*

*A graph of the united states

Description automatically generated*

***Note.*** *Unemployment insurance and short-time work take-up* (Giupponi et al., 2022)

## Fraud in Government Programs

Fraud within government programs is a longstanding challenge that has been exacerbated by the scale and urgency of relief efforts, such as those seen during the COVID-19 pandemic. The PPP, while well-intentioned, became a prime target for fraud due to its size, speed of implementation, and broad eligibility criteria. Fraudsters exploited the program’s vulnerabilities by submitting false applications, inflating payrolls, and in some cases, creating fictitious businesses to qualify for loans (USSBA OIG, 2023).

The issue of fraud in government programs is not unique to the PPP. Historically, government programs—especially those disbursing large sums of money—have been prone to fraud. According to Kadens (2023), fraud has plagued public administration for centuries. For instance, during the Great Depression, programs like the New Deal also faced widespread fraud as individuals and organizations sought to capitalize on federal assistance without meeting eligibility requirements. These historical examples illustrate that government efforts to distribute relief quickly often result in trade-offs between oversight and efficiency.

The rapid implementation of the PPP created similar vulnerabilities. The program disbursed loans through thousands of private banks, which were incentivized to process loans quickly due to the federal loan guarantees. This decentralized approach meant that many fraudulent applications went unnoticed, as banks did not conduct thorough checks on the legitimacy of the applicants (Li & Strahan, 2020). Furthermore, the urgent need to distribute funds left little time to develop robust anti-fraud measures, resulting in significant amounts of taxpayer money being lost to fraud.

### Macro-Level Governance and Regulatory Challenges

At the macro level, regulatory frameworks often struggle to keep up with the complexities of large-scale government programs, especially during crises. The rapid distribution of PPP funds created challenges for federal regulators, such as the U.S. Treasury and SBA, both of which were tasked with overseeing the program (Department Of Treasury, 2024). A key issue was the lack of existing infrastructure to monitor and verify the legitimacy of millions of loan applications in real time.

Efforts to combat fraud have varied across different countries, with some nations taking more proactive approaches to prevent abuse. Bozza (2024) highlights how the U.K. employed stricter regulatory measures to recover fraudulent pandemic loans, relying more heavily on centralized investigations and automated flagging systems to identify suspicious applications. By contrast, the U.S. approach relied more on whistleblowers and retroactive investigations, which have been less effective in preventing fraud upfront.

### Meso-Level Organizational Controls

At the organizational level, the effectiveness of fraud prevention often depends on the internal controls and governance structures within the agencies administering the programs. In the case of the PPP, banks and financial institutions were responsible for distributing the loans, but they often lacked the incentive to rigorously vet applications, given the federal loan guarantees (Li & Strahan, 2020). Ali et al. (2021) argue that robust internal controls are essential to reducing the risk of fraud, particularly in public sector programs. However, the complexity and scale of the PPP overwhelmed many of these controls, allowing fraud to proliferate.

Organizational dynamics also play a role in how fraud is managed internally. For instance, Harrington & Leslie (2023) highlight how organizations with weak internal governance and poor fraud prevention measures are more likely to fall victim to sophisticated fraud schemes. This insight is particularly relevant to large government programs where decentralized structures, such as the PPP’s reliance on private lenders, complicate the enforcement of anti-fraud policies.

In the context of the PPP, operational constraints presented additional challenges as organizations had to process an unprecedented number of loans swiftly. This environment created a strain on organizational controls, as traditional fraud detection mechanisms could not adapt to the urgency of loan disbursement and were not equipped to handle the unique data profile of PPP transactions. Studies indicate that machine learning, especially semi-supervised models like classification through clustering, can augment organizational controls by detecting subtle anomalies that may otherwise go unnoticed in high-volume datasets. This is essential for programs like the PPP, where legitimate applications vastly outnumber fraudulent ones, creating an imbalance that complicates traditional auditing approaches (Gui et al., 2024; Miller & Bertozzi, 2024).

**Figure 4***PPP Loan Application Process Flowchart*

A diagram of a computer system

Description automatically generated

***Note.*** *Source:* (USSBA, 2021)

### Comparative Insights from Government Programs

Lessons from other government programs, such as Medicaid and unemployment insurance, further illustrate the importance of adaptive fraud detection strategies. Medicaid, for example, has been an ongoing target for fraud due to its extensive, complex data ecosystem and the variety of service providers involved. The program’s reliance on claims-based transactions necessitates sophisticated fraud detection algorithms capable of distinguishing between legitimate claims and fraudulent patterns. Similarly, unemployment insurance programs have faced elevated fraud risks in recent years, particularly as pandemic relief benefits introduced expanded eligibility criteria. These programs highlight the need for fraud detection models that combine automated anomaly detection with human expertise to provide a layered defense against fraud (Emilio Ferrara, 2023; Zhao et al., 2024).

Studies suggest that for public sector fraud detection to be effective, it must integrate both macro-level policy compliance and meso-level operational controls within a flexible, technology-enabled framework. By implementing classification through clustering, this study aims to bridge these levels, providing a model that can adapt to the unique regulatory and operational demands of the PPP. The use of advanced machine learning techniques not only enhances fraud detection accuracy but also allows the program to meet transparency and accountability standards essential in public sector applications.

## Machine Learning for Fraud Detection

Fraud detection in government programs like the PPP presents significant challenges due to the sheer volume of data, complexity of transactions, and constantly evolving fraud strategies. Traditional auditing and detection methods often fall short when dealing with large, high-dimensional datasets, which makes ML an ideal solution. ML models can detect hidden patterns, continuously adapt to new fraud techniques, and process massive amounts of data in real time, helping detect fraud more efficiently.

The PPP disbursed over $934 billion in loans, creating an environment ripe for potential fraud due to the rapid distribution and lack of oversight. This makes it essential to employ advanced data-driven techniques to analyze the millions of loan applications submitted to the program.

### Supervised Learning

Supervised learning models are highly effective when labeled data is available, such as pre-identified fraudulent and non-fraudulent transactions. The model is trained on this labeled data, learning patterns, and features associated with fraud. After training, the model can classify new data points by predicting whether they are fraudulent or not based on these learned patterns.

**Logistic regression** is a simple yet widely used supervised learning algorithm for fraud detection. It operates by estimating the probability that a particular loan application is fraudulent, based on input features such as the size of the loan, payroll data, and the business’s industry. Logistic regression is particularly effective when the relationship between variables is linear (Sarker, 2021).

* **Advantages**: Logistic regression is highly interpretable, meaning that the model’s predictions can be understood by non-technical stakeholders such as auditors and policymakers. It allows for easy identification of which features contribute most to predicting fraud.
* **Limitations**: The model struggles with non-linear relationships and may underperform in complex scenarios, such as detecting sophisticated fraud schemes where features interact in a more intricate manner.

**Figure 5**   
*Logistic Regression Response Function*

A graph of a function

Description automatically generated

***Note.*** *Source:* (Maalouf, 2011)

**Random forest** is an ensemble learning method that builds multiple decision trees during training and averages their outputs to make a final classification. This approach is highly robust and effective in handling both categorical and continuous variables, making it ideal for detecting complex patterns of fraudulent behavior in PPP loan application (Itri et al., 2019).

* **Advantages**: Random forest handles large datasets with many variables and is resistant to overfitting. It can manage unbalanced datasets, which is often the case in fraud detection, where non-fraudulent cases vastly outnumber fraudulent ones.
* **Limitations**: Despite its robustness, random forest models are less interpretable than simpler models like logistic regression, making it harder for human auditors to understand the decision-making process.

**Figure 6***Graphic Representation of a Random Forest Decision Tree*

A diagram of a tree

Description automatically generated

***Note.*** *Source:* (Schonlau & Zou, 2020)

**Support Vector Machines (SVMs)** is another powerful supervised learning algorithm used in fraud detection. SVMs work by finding a hyperplane that best separates fraudulent transactions from non-fraudulent ones in a high-dimensional space. SVMs are especially useful when the data is non-linearly separable, which is common in fraud detection scenarios where fraudsters employ sophisticated methods to mimic legitimate transactions (López et al., 2012).

* **Advantages**: SVMs are effective in high-dimensional spaces and can model complex, non-linear relationships between variables. They are well-suited for identifying subtle patterns in large datasets.
* **Limitations**: SVMs can be computationally intensive, especially when applied to very large datasets, and require careful tuning of parameters to achieve optimal performance.

**Figure 7***SVM Hyperplane*

A diagram of a graph

Description automatically generated

***Note.*** *Source:* (Cervantes et al., 2020)

### Unsupervised Learning and Clustering Techniques

Unsupervised learning is critical when labeled data is sparse or unavailable. In fraud detection, unsupervised learning models are designed to identify hidden structures in the data, grouping similar data points together into clusters and flagging outliers or anomalies that may represent fraudulent activity. This is particularly useful in the PPP, where labeled fraud data is limited but suspicious patterns can still be uncovered by analyzing the inherent structure of the data.

**K-Means Clustering** is one of the most common unsupervised learning algorithms used for fraud detection. It works by grouping data points into clusters based on their similarity. For the PPP, businesses with similar characteristics—such as loan amount, industry, or number of employees—would be clustered together. Outliers within these clusters, such as businesses with inflated payrolls relative to their peers, can be flagged as potential fraud cases (López et al., 2012).

* **Advantages**: K-Means is computationally efficient and scalable, making it suitable for large datasets like the PPP. It is also easy to implement and interpret.
* **Limitations**: K-Means requires the number of clusters to be specified in advance, which can be challenging if the true structure of the data is unknown. Additionally, it is sensitive to outliers, which may skew the clustering results.

**Figure 8***Illustration of K-Means Clustering Method*

*A comparison of different colored dots

Description automatically generated with medium confidence*

***Note.*** *Source:* (Rodriguez et al., 2019)

**DBSCAN** is another popular unsupervised learning algorithm that is particularly effective at detecting fraud in large datasets. Unlike K-Means, DBSCAN does not require the number of clusters to be specified beforehand. Instead, it detects clusters based on the density of data points, identifying areas of high density (potential legitimate transactions) and separating out noise or anomalies (potential fraud) (Carcillo et al., 2021).

* **Advantages**: DBSCAN is highly effective at identifying clusters of irregular or unusual behavior and can detect smaller clusters of fraudulent activity within larger datasets. It also handles noise in the data more effectively than K-Means.
* **Limitations**: DBSCAN may struggle to identify clusters in data with varying density, and it can be sensitive to the choice of parameters for detecting outliers.

**Figure 9***Graphic Depiction of DBSCAN Clustering*

A diagram of a circle with arrows and circles

Description automatically generated

***Note.*** *Source:* (Zhang et al., 2022)

**Hierarchical clustering** creates a tree-like structure of nested clusters known as a dendrogram, providing multiple levels of grouping within the data. This technique is particularly useful for detecting fraud at different levels of granularity, helping to uncover not only individual outliers but also broader patterns of suspicious behavior across various subsets of PPP loan applicants (Zhao et al., 2024).

* **Advantages**: Hierarchical clustering does not require the number of clusters to be specified in advance, making it more flexible than K-Means. It also allows for a more detailed analysis of the relationships between clusters.
* **Limitations**: Hierarchical clustering can be computationally expensive, particularly for very large datasets, and it may not perform well in situations where the true structure of the data is complex and non-hierarchical.

**Figure 10***Hierarchical Clustering Dendrogram*

A diagram of a tree

Description automatically generated

***Note.*** *Source:* (Halkidi, 2009)

### Semi-Supervised Learning

Semi-supervised learning (SSL) combines elements of both supervised and unsupervised learning, making it especially useful in fraud detection scenarios where a small set of labeled data is available, but most of the data remains unlabeled. This approach enables the model to leverage the limited labeled data to make informed predictions about the unlabeled data.

**Graph-based SSL**, as described by Miller & Bertozzi (2024), leverages graph Laplacian spectral truncation to model complex relationships within data. By representing data points (e.g., loan applicants or transactions) as nodes and their connections as edges, this approach captures relational patterns that traditional clustering methods may overlook. This is particularly effective for datasets like the PPP where interconnected entities naturally form network structures.

* **Advantages**:
  + **Relational Insights**: By encoding both local and global relationships, graph-based methods can identify nuanced patterns of fraud. For example, loans disbursed through the same lender or within the same geographic region often share behavioral correlations, such as repayment anomalies or loan-to-employee ratios.
  + **Scalability**: Spectral truncation techniques optimize graph processing, enabling the model to scale effectively for large datasets, such as the PPP's millions of records.
  + **Label Propagation**: With limited labeled fraud cases, graph-based SSL excels by propagating labels across the graph, generalizing patterns from known fraud instances to unlabeled data points (King et al., 2023).
* **Limitations**
  + **Computational Costs**: Constructing and processing large-scale graphs can be resource-intensive, particularly for high-dimensional datasets.
  + **Sensitivity to Noise**: Graph performance depends on the quality of its edges. Erroneous or noisy connections can lead to misleading patterns and reduced accuracy.
  + **Graph Construction Complexity**: Designing the graph (e.g., choosing similarity metrics or defining edge weights) requires domain expertise and can vary significantly by dataset.

**Pseudo-labeling** is a common semi-supervised learning technique in which the model first labels the unlabeled data based on its predictions from the labeled dataset. These pseudo-labeled data points are then used to retrain the model, gradually improving its accuracy. In the PPP, pseudo-labeling can be applied to flag potentially fraudulent loan applications that share characteristics with known fraudulent cases (Zhao et al., 2024).

1. **Advantages**: Pseudo-labeling allows for continuous improvement of the model’s predictions, even when only a small fraction of the data is labeled. It is particularly useful in fraud detection when the cost of obtaining labeled data is high.
2. **Limitations**: Pseudo-labeling can introduce noise into the model if the initial predictions are incorrect, leading to a self-reinforcing cycle of errors. This can be mitigated through careful validation of the model’s outputs.

**Self-training** is another iterative semi-supervised learning process, where the model is initially trained on the labeled data and then uses its own predictions to label additional data points. With each iteration, the model refines its predictions, becoming more accurate over time (Sarker, 2021).

* **Advantages**: Self-training allows the model to improve with minimal human intervention and is highly scalable, making it suitable for large datasets like the PPP.
* **Limitations**: Similar to pseudo-labeling, self-training runs the risk of propagating errors if the model’s early predictions are inaccurate.

**Active learning** further strengthens SSL by selectively labeling the most informative samples. Zhao et al. (2024) introduced the "Maximizing Expected Model Change" framework, which identifies impactful data points for labeling, reducing computational overhead while increasing model accuracy. These methodologies align with detecting evolving fraud patterns, a significant challenge in government funding oversight.

Finally, pseudo-labeling and self-training techniques can be combined with active learning strategies to achieve iterative improvements, leveraging both structured relationships (e.g., graphs) and prioritized labeling efforts. This hybrid approach offers a robust framework for fraud detection in dynamic environments.

### Evaluation Metrics for Fraud Detection in Supervised Learning Models

Given the imbalanced nature of fraud detection datasets—where fraudulent transactions represent only a small fraction of the total data—it is critical to use appropriate metrics for evaluating the performance of machine learning models. The following metrics are commonly used to assess fraud detection models:

**Accuracy** measures the proportion of correct predictions made by the model. However, in fraud detection, accuracy alone can be misleading due to the imbalance between fraudulent and non-fraudulent cases. For example, a model that predicts all transactions as non-fraudulent may achieve high accuracy but fail to detect actual fraud cases.

**Precision and Recall.** Precision measures the proportion of predicted fraudulent cases that are fraudulent. It is important in scenarios where false positives (incorrectly identifying a transaction as fraudulent) have significant consequences, such as triggering unnecessary investigations. Recall measures the proportion of actual fraudulent cases that the model correctly identifies. It is critical in minimizing false negatives, where the model fails to detect fraud, allowing it to go unchecked (Zhao et al., 2024).

**F1 score** combines precision and recall into a single metric, providing a more balanced evaluation of the model’s performance. It is particularly useful in cases where both false positives and false negatives carry significant costs.

**AUC-ROC curve** evaluates the model’s ability to distinguish between fraudulent and non-fraudulent cases across different decision thresholds. A higher area under the curve (AUC) indicates better model performance. This metric is essential for models that need to be fine-tuned to maximize fraud detection while minimizing false positives.

### Evaluation Metrics for Fraud Detection in Clustering

In unsupervised and semi-supervised learning contexts, evaluating clustering models requires specific metrics due to the absence of labeled ground truth for most data points. Effective clustering evaluation ensures that fraudulent patterns emerge clearly within clusters.

**Internal Validation Metrics**:

* **Silhouette Score**: Assesses the compactness of clusters by comparing intra-cluster similarity with inter-cluster separation. For PPP datasets, the silhouette score can optimize the number of clusters in k-means clustering or verify cluster quality in DBSCAN.
* **Calinski-Harabasz Index**: Quantifies cluster dispersion by analyzing the ratio of inter-cluster separation to intra-cluster cohesion. It is particularly useful for comparing different clustering algorithms in fraud detection, such as hierarchical clustering versus k-means.

**External Validation Metrics**: When a small subset of labeled data is available, external validation metrics provide a benchmark for clustering accuracy:

* **Adjusted Rand Index** evaluates how well clustering results align with labeled data, accounting for chance.
* **Normalized Mutual Information (NMI)** quantifies the agreement between clustering outputs and known labels, ranging from 0 (no agreement) to 1 (perfect agreement).

**Cross-Validation for Clustering Stability**

To ensure clustering robustness, cross-validation techniques are adapted to assess stability:

* **Split-Half Reliability**: Splitting the dataset into halves and applying clustering separately evaluates whether the model produces consistent results across subsets.
* **Bootstrap Sampling**: By creating randomized subsets of the dataset and re-clustering, bootstrap techniques measure consistency in cluster assignments.
* **Cluster Stability Index**: Aggregates cluster consistency metrics across iterations to quantify model reliability, critical for datasets like PPP, where noise or outliers might affect clustering.

**Integration with PPP Fraud Detection**

For the PPP fraud detection:

* **Application**: Metrics like silhouette score and Calinski-Harabasz index guide the iterative refinement of cluster quality, ensuring that fraudulent and legitimate patterns are well-separated before transitioning to binary classification.
* **Scalability**: Cross-validation ensures clustering methods perform reliably on large-scale, high-dimensional PPP datasets, adapting to the program's unique challenges of sparse labeled fraud cases and complex transaction data.

### Conclusion to Machine Learning for Fraud Detection

The combination of supervised, unsupervised, and semi-supervised models offers a comprehensive approach to fraud detection, with each methodology contributing unique strengths for handling high-volume, low-fraud datasets. By focusing on semi-supervised learning, particularly classification through clustering, this study aims to develop an adaptable model that meets the PPP’s unique operational demands while ensuring robust fraud detection.

## Summary

This literature review explored key methodologies, theoretical frameworks, and ethical considerations central to effective fraud detection in the context of the PPP. The review focused on classification through clustering as a primary method for managing the PPP’s imbalanced data structure, where fraudulent cases are vastly outnumbered by legitimate ones. Semi-supervised techniques, particularly those integrating clustering and dimensionality reduction through PCA, provide a scalable approach to identifying subtle anomalies (Gui et al., 2024; López et al., 2012). These methods are bolstered by behavioral frameworks, including the Fraud Triangle and its expanded forms, which allow for the interpretation of anomalies in relation to known fraud motivations, such as opportunity and financial pressure (Bozza, 2024; King et al., 2023).

Ethical and regulatory considerations highlighted the need for transparent, privacy-compliant models, especially in government applications where transparency supports accountability. Adherence to standards like GDPR and CCPA ensures that fraud detection technologies align with public expectations for privacy and ethical AI, underscoring the importance of explainable XAI methods in maintaining compliance (Emilio Ferrara, 2023; Koreff et al., 2023).

### Current State of the Literature

Fraud detection research has primarily focused on supervised and unsupervised learning methods due to their maturity and effectiveness in various domains:

* **Supervised Learning**: Widely applied in financial fraud detection, supervised models such as decision trees and SVMs achieve high precision and recall when sufficient labeled data is available. For instance, Bauder & Khoshgoftaar (2017) examined supervised classifiers for Medicare fraud detection, demonstrating the importance of feature engineering in achieving robust classification results​.
* **Unsupervised Learning**: Techniques like clustering and anomaly detection excel in identifying outliers or unusual patterns without relying on labeled data. Carcillo et al. (2021) demonstrated how clustering methods can effectively detect anomalies within loan application datasets, offering valuable insights in cases where labeled data is unavailable​.
* **Semi-Supervised Learning (SSL)**: SSL, which uses a small set of labeled data alongside a large volume of unlabeled data, is gaining traction in scenarios with limited labeled datasets. Xu et al. (2022) reported the superiority of SSL techniques over traditional supervised methods, particularly in domains where labeled data is scarce​.

### Unmet Needs and Gaps

Despite these advancements, several critical gaps persist in the literature:

* **Limited Research on Government Programs**: The majority of studies focus on private-sector fraud detection (e.g., credit card fraud), with limited exploration of government-specific domains like the PPP. Research on government subsidy fraud, such as the E-Rate program, highlights the need for tailored fraud detection frameworks (USGAO, 2020).
* **Challenges with Imbalanced Datasets**: Fraudulent cases in government programs often constitute a small minority, leading to significant imbalances in datasets. Techniques such as Synthetic Minority Oversampling Technique (SMOTE) have been explored (Benala & Tantati, 2022), but their specific application to PPP datasets remains underdeveloped​​.
* **Adapting to Evolving Fraud Schemes**: Government fraud schemes are dynamic, exploiting new loopholes as regulations change. Traditional supervised methods relying on static labeled datasets struggle to generalize to these novel patterns, as noted by Larson (2020) in discussions of imbalanced learning challenges​.
* **Underexplored Integration of SSL and Imbalance Compensation**: While SSL is increasingly utilized, most studies fail to integrate advanced imbalance compensation techniques such as weighted loss functions or domain-specific oversampling strategies with SSL models. This is particularly crucial for enhancing the detection of minority classes in high-dimensional datasets​​.

### Novelty of Proposed Methodology

This research addresses these gaps through:

1. **Domain-Specific Focus**: By concentrating on PPP fraud detection, this study expands the literature on government subsidy fraud. Prior research has primarily emphasized healthcare (e.g., Medicare) or financial fraud (Bauder & Khoshgoftaar, 2017), leaving government loan programs underrepresented​​.
2. **Semi-Supervised Learning with Imbalance Compensation**: Integrating SSL with imbalance compensation techniques such as SMOTE or cost-sensitive learning is a novel approach, particularly for government datasets. Xu et al. (2022) and Benala & Tantati (2022) highlight the need for this integration to address fraud detection challenges in imbalanced datasets​​.
3. **Dynamic Fraud Pattern Detection**: By leveraging unlabeled data, SSL adapts to emerging fraud patterns, overcoming the limitations of static supervised models. This adaptability is vital for detecting novel PPP fraud schemes​​.

**4. Relevance to PPP Fraud Detection**

The PPP’s unique challenges, including high-dimensional datasets, sparse labeled fraud cases, and evolving fraud schemes, necessitate innovative approaches. SSL, combined with advanced imbalance handling, offers a scalable and adaptable solution, addressing critical gaps in current fraud detection systems while aligning with government program requirements​​.

### Future Directions and Implications for Government Programs

Given these gaps, several future research directions could enhance fraud detection efficacy in government programs:

* **Advanced Clustering Algorithms for Adaptive Fraud Detection**: Developing clustering methods that adapt in real-time to shifts in transactional patterns could significantly improve the model's ability to detect emerging fraud schemes. These adaptive algorithms would be especially beneficial in high-stakes, dynamic contexts like government aid distribution.
* **Expanded Fraud Theoretical Frameworks**: Future research could extend the Fraud Triangle by adding context-specific factors—such as external pressures or changing regulatory landscapes—that are relevant to fraud in emergency relief programs.
* **Hybrid Models to Address Ethical Compliance**: Hybrid models combining rule-based and data-driven techniques may offer a pathway to enhanced interpretability without sacrificing accuracy. This approach could be particularly valuable in public sector applications, where ethical compliance is paramount.

By advancing fraud detection methodologies through these directions, future research can support the development of resilient, transparent, and adaptable systems that protect government resources. For the PPP and similar programs, these improvements promise more secure fraud prevention while maintaining the accessibility and trust that public programs require. The methods and frameworks synthesized in this study lay the groundwork for the semi-supervised approach outlined in Chapter 3, focusing on scalable fraud detection that aligns with the operational and ethical requirements of modern government initiatives.

# Chapter 3: Research Method

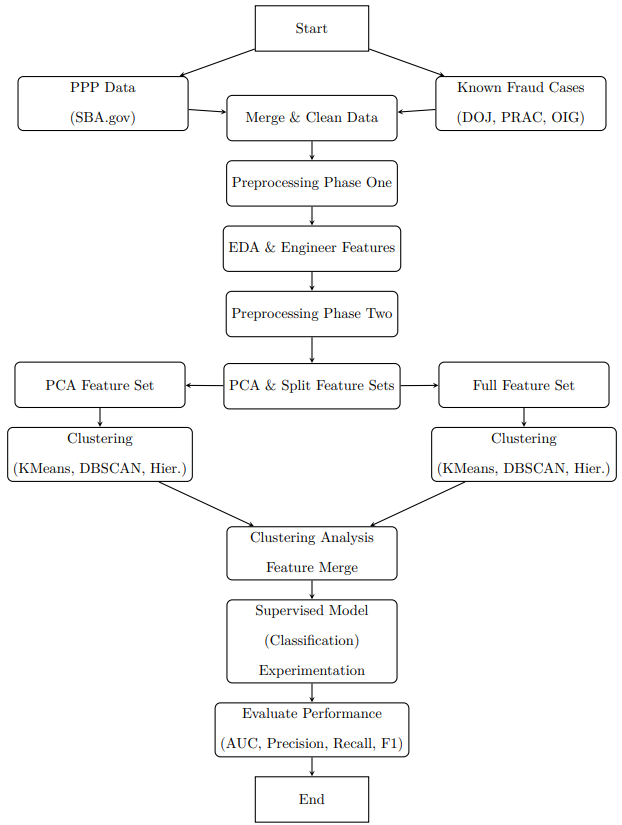
The problem addressed in this study is the lack of intelligent unsupervised fraud identification in the government domain, specifically in the Paycheck Protection Program (PPP). The purpose of this quantitative experimental study was to identify features and methodologies develop intelligent semi-supervised fraud identification methodologies in the government domain, specifically the PPP. This chapter outlines the research methodology and design of this study. It includes an overview of the population, instrumentation, data collection, and definitions of variables. It then describes the study procedure, the data analytics performed as part pf hypothesis testing, assumptions, limitations, delimitations, and ethical assurances.

## Research Methodology and Design Process Diagram

The methodological workflow used in this study is summarized in Figure 11. It outlines the major phases of the research, beginning with data collection, preprocessing, and exploratory analysis, followed by feature set partitioning and clustering using multiple unsupervised algorithms. Cluster outputs were merged back into the dataset and served as inputs to a series of semi-supervised experiments using various supervised models. Each stage in the process was aligned with the study’s hypotheses, culminating in model evaluation based on precision, recall, F1-score, and AUC-ROC.

**Figure 11**

*Research Methodology and Design Process Diagram*

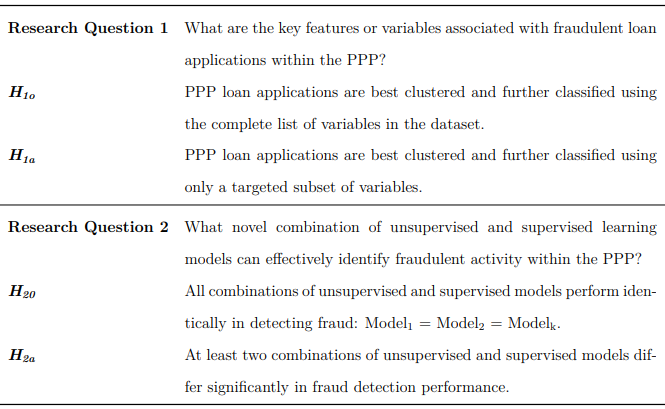


## Research Methodology and Design (Nature of the Study)

This study employed a quantitative experimental design that implemented a Classification through Clustering framework: a hybrid methodology integrating both unsupervised clustering and semi-supervised classification techniques. The design follows a systematic, hypothesis-driven structure, ensuring that each phase directly addressed the research questions and hypotheses shown in the table below:

**Table 1**

*Research Question and Hypotheses Alignment*



This design was appropriate due to the complexity of the PPP dataset, the limited number of labeled fraud cases, and the need to detect hidden or novel fraud patterns in unlabeled data. Classification through Clustering enabled the discovery of anomalies using unsupervised methods and the refinement of those findings through supervised classification Alternative approaches, such as solely supervised learning, were considered; however, such methods would have required comprehensive labeling, which is challenging in large datasets with limited labeled fraud instances. The chosen multi-phase approach allows for an adaptive, hybrid model that effectively captures anomalies, clusters data, and improves detection accuracy.

### Alternative Methodologies Considered

Three alternative approaches were considered and determined to be less appropriate for the problem context:

**1. Supervised Learning Only:** This approach was dismissed due to the scarcity of labeled fraud cases, the risk of model overfitting, and the difficulty of generalizing predictions across an imbalanced dataset.

**2. Unsupervised Learning Only:** While effective for anomaly detection, unsupervised clustering alone lacked refinement and suffered from high false-positive rates due to the absence of verified fraud labels.

**3. Manual Fraud Detection:** Manual review of loan records was infeasible due to the dataset’s size and the need for scalable, automated methods.

### Justification for the Selected Approach

The Classification through Clustering framework offered the flexibility to process high-dimensional data, incorporate both labeled and unlabeled examples, and adapt to the class imbalance inherent in fraud detection. It also enabled comparative analysis of multiple model pairings, consistent with the study’s hypothesis-driven structure. By combining anomaly detection with confirmed fraud cases, this methodology improved detection precision while preserving scalability and generalizability—critical requirements for analyzing a national-scale financial relief program.

## Population and Sample

The population for this study consisted of all businesses and entities that received loans through the PPP. The publicly available loan-level dataset released by the SBA included more than 11 million records, detailing loan characteristics, borrower location, and business classification.

However, for this study, the scope was limited to PPP loans exceeding $150,000, which represented a subset of 968,525 loans. This threshold was selected based on the elevated financial risk associated with larger loans and the greater likelihood of such loans being subject to federal investigation, which increased the availability of reliable fraud labels. Analyzing this high-value subset supported the study’s objective to identify scalable fraud detection methods capable of addressing the most impactful cases. Additionally, this subset was published by the SBA as a single CSV file, which made the data more accessible and enabled greater reproducibility of the research process across environments and systems with limited local computing power.

**Rationale for Using the Entire Population**

Using the entire population of large PPP loans allowed for greater analytical depth and accuracy in detecting fraud patterns. Larger loans are more likely to reflect complex fraud schemes and provide richer feature sets for modeling. The availability of verifiable fraud cases also improved the quality of supervised learning inputs.

Additionally, the SBA published loans exceeding $150,000 as a single, consolidated CSV file, making this high-value subset easily accessible and highly reproducible for public analysis. In contrast, the smaller loan records, though available, were split across multiple files and would have required processing over 5 GB of raw data. This would have introduced significant computational and storage overhead. Restricting the scope to the over $150K segment allowed for efficient data handling, consistent replication, and alignment with the study’s objective to identify scalable fraud detection methodologies for the most financially significant cases.

## Materials and Instrumentation

This study utilized archived public datasets, a manually labeled fraud vector, custom-developed Python scripts, and open-source machine learning libraries to build and evaluate a hybrid fraud detection pipeline. Instrumentation included both the digital tools used to acquire and process the data and the computational environment that enabled scalable modeling and evaluation.

### ****Data Sources****

**PPP Loans**: The core dataset was obtained from the SBA, which published loan-level details for all approved PPP loans. The subset used in this study consisted of loans exceeding $150,000, comprising 968,525 records and 52 variables, distributed in a single, consolidated CSV file. This format supported straightforward ingestion, consistent replication, and reproducible access to structured loan data including borrower entity name, loan amount, geographic location, NAICS codes, reported jobs, and loan forgiveness status.

**Labeled Fraud Cases**: To create the fraud-labeled subset required for supervised and semi-supervised learning, a custom web scraper was written in Python using BeautifulSoup. This tool collected approximately 2,500 press releases and case summaries from PandemicOversight.gov, which aggregates fraud reports from the DOJ, SBA OIG, and related oversight bodies. Although PRAC maintained a centralized news archive, automated scraping was necessary to extract all cases programmatically. Manual review was then conducted to filter for prosecuted PPP fraud cases, resulting in 301 confirmed fraudulent applications. These cases were matched to SBA loan records using fuzzy logic on borrower name, loan amount, and business location. The final binary label vector was stored and integrated with the broader modeling dataset.

### ****Software and Computational Environment****

**Platform**: All modeling and data engineering tasks were performed in Google Colab Pro, which provided access to both high-RAM CPU instances and A100 GPU accelerators. Exploratory data analysis (EDA) and extract-transform-load (ETL) tasks were conducted in notebook-based workflows, while unsupervised and supervised modeling was scripted in modular Python files to ensure reproducibility.

**Key Libraries and Frameworks**:

* pandas, numpy: For tabular data transformation and feature engineering
* cuml: For GPU-accelerated PCA, K-Means, and DBSCAN implementations
* scikit-learn: For logistic regression, cluster evaluation metrics (e.g., Davies-Bouldin Index), and model validation
* xgboost: For gradient boosting classification on structured data
* keras (with TensorFlow backend): For neural network construction and training
* matplotlib, seaborn: For visualizations and diagnostic plots
* BeautifulSoup: For web scraping PRAC case reports

### Operational Definitions of Variables

This study used two types of variables: raw features derived from the PPP dataset, which served as inputs to PCA and machine learning models, and experimental variables created to test the study’s hypotheses. The operational definitions below distinguish between these categories and describe how each was used in the modeling framework.

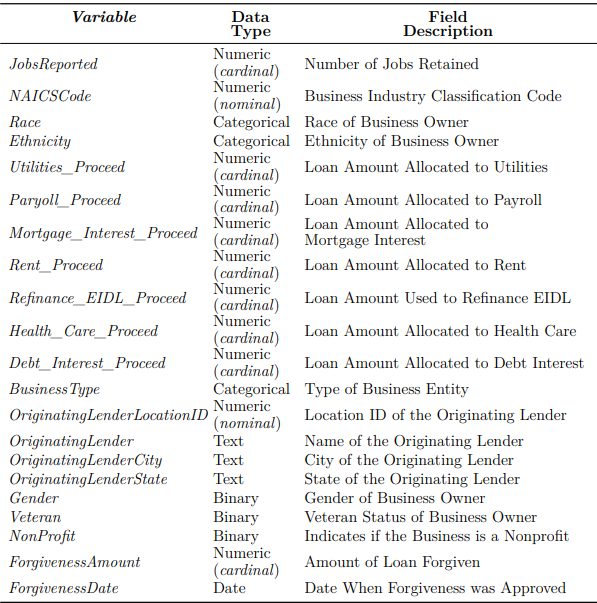
**Original PPP Dataset Features:** The table below lists the primary variables extracted from the SBA’s PPP dataset and used as input features in the modeling process. These include financial indicators, categorical attributes, and derived features engineered to capture anomalous patterns in borrower behavior.

**Table 2**

*Original Variables for The PPP Dataset*



**Table 2 – continued from the previous page**



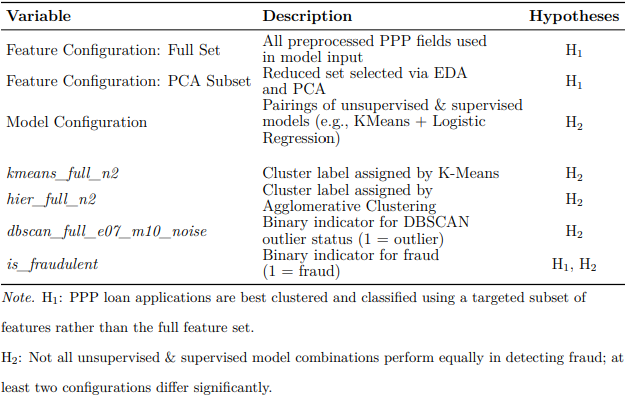
Additionally, derived features were created through feature engineering, including:

* *Loan-per-Job* Ratio: Calculated as *LoanAmount* / *JobsReported*, used to detect disproportionately large loan requests.
* *ForgivenessRatio*: Calculated as *ForgivenessAmount* / *CurrentApprovalAmount*, used to evaluate repayment behavior.

**Experimental Variables for Hypotheses Testing:** The table below presents the key variables defined for hypotheses testing within this study. These include experimental configurations such as feature sets and model pairings, as well as engineered features derived from clustering outputs. Each variable played a direct role in evaluating the study’s hypotheses regarding the impact of feature selection H1 and the effectiveness of various clustering-supervised model combinations H2.

**Table 3**

*Experimental Variables and Hypotheses Alignment*



## Study Procedures

This section outlines the procedures followed to collect, preprocess, and analyze data for the identification of fraudulent PPP loan applications. The study followed a multi-phase experimental design, incorporating both unsupervised and semi-supervised learning approaches, supported by carefully curated labeled data and scalable cloud-based infrastructure.

### Data Collection

The primary dataset was acquired from the SBA, which publicly released detailed records of approved PPP loans. This study focused on loans exceeding $150,000, comprising a subset of 968,525 records. The decision to restrict the scope to this high-value segment was based on three factors: elevated financial risk associated with larger loans, increased likelihood of investigative scrutiny and prosecution, which facilitated the identification of confirmed fraud cases, and the availability of the full over $150K dataset as a single CSV file, enhancing reproducibility and computational feasibility.

To develop a labeled dataset of fraudulent loans, a custom web scraper was built using Python’s BeautifulSoup library to extract approximately 2,500 COVID-19 relief-related articles and press releases from PandemicOversight.gov, the portal maintained by the PRAC. Each report was manually reviewed to identify instances of confirmed PPP fraud. A total of 301 unique fraudulent loan applications were matched from the primary dataset based on borrower name, loan amount, and geographic location. These cases were encoded in a binary variable *is\_fraudulent*.

### Data Preprocessing and Exploratory Analysis

Preprocessing and exploratory data analysis (EDA) were conducted in two phases to ensure the dataset was privacy-compliant, interpretable, and analytically robust for downstream modeling.

**Initial Preprocessing and EDA:** Prior to formal analysis, the dataset underwent basic transformations necessary for secure handling and visual exploration. These included hashing of PII fields (e.g., *BorrowerName*, *BorrowerAddress*, *FranchiseName*) to preserve record-level uniqueness while PII, as well as preliminary feature engineering. These transformations enabled a structured exploratory analysis. EDA techniques included univariate analysis, boxplots to detect distributional skew or outliers, and correlation matrices to identify feature redundancy and key variables. The analysis from this phase identified the reduced key feature subset which was analyzed alongside the full feature set during PCA.

**Post EDA Preprocessing and PCA:** Following this initial phase, a second round of preprocessing was conducted to prepare the dataset for dimensionality reduction and machine learning. This stage included z-score standardization of continuous variables to support distance-based clustering algorithms, as well as the encoding of categorical variables through a combination of frequency, label, and one-hot encoding methods based on feature cardinality. Missing values were addressed without imputation to avoid masking potential fraud-related patterns; categorical nulls were recoded into a “Missing” category, while missing numeric values were flagged using binary indicators. Temporal fields were further decomposed to extract discrete components such as month and day. PCA was then applied to both the full feature set and the key feature subset, with the full set achieving 95% explained variance using just three components. In contrast, the key feature subset reduced to a single principle component and was therefore discarded as it would likely introduce high levels of oversimplification during clustering. The final dataset for clustering included both PCA-reduced (3 components) and Full feature (52 variable) matrices for model comparison and evaluation.

### Clustering Methodology

Unsupervised learning was conducted to detect structure within the PPP loan dataset prior to fraud classification. Clustering was performed on two versions of the preprocessed feature matrix: a full feature set without dimensionality reduction (*X\_all\_no\_pca*), and a PCA-reduced version (*X\_all\_pca\_3*) that retained 95% of the total variance across three components. Both versions were subjected to structured hyperparameter exploration, with parameter grids explicitly defined in external JSON configuration files. These configurations enabled systematic sweeps across key clustering parameters, such as the number of clusters for K-Means and Hierarchical models, and the ε and min\_samples values for DBSCAN, allowing reproducible experimentation across the full search space. While not conducted using a formal hyperparameter search utility, this programmatic approach functionally mirrored a grid search by executing all defined combinations sequentially in a controlled manner.

Three clustering algorithms were applied to both the non-reduced and PCA-reduced feature spaces: K-Means, Agglomerative Hierarchical Clustering, and DBSCAN. K-Means was evaluated across a range of two to ten clusters to allow for the detection of both broad fraud/non-fraud distinctions and finer substructures, while Hierarchical clustering was tested over a narrower range of two to six clusters to accommodate its higher computational demands and the expectation of fewer natural groupings. DBSCAN was assessed across multiple combinations of epsilon (ε) (0.3 to 1.3) and min\_samples (3 to 15), allowing sensitivity to both compact and diffuse outliers. The final selected DBSCAN configuration (ε = 0.7, min\_samples = 10) was chosen based on qualitative alignment with known fraud cases and favorable cluster validation scores. All algorithms were executed on GPU using RAPIDS cuML to enable scalable experimentation across high-dimensional feature spaces. Each clustering result was saved along with its associated hyperparameters and evaluation metrics to facilitate downstream model integration and reproducibility.

Clustering performance was evaluated using silhouette score and Davies-Bouldin Index (DBI), computed via cuML and Scikit-learn, respectively. Across all configurations, the non-PCA version of the dataset (*X\_all\_no\_pca*) consistently produced higher silhouette and lower DBI scores, indicating better cluster cohesion and separation. As a result, clustering labels generated from the non-PCA full feature set were retained for integration into the classification pipeline.

Three clustering-derived features were extracted: *kmeans\_full\_n2*, *hier\_full\_n2*, and *dbscan\_full\_e07\_m10\_noise*. The first two represented cluster assignments from the best-performing K-Means and Hierarchical models (each with *k* = 2). The third was a binary indicator for noise points (outliers) from DBSCAN with = 0.7 and min\_samples = 10, which demonstrated strong alignment with known fraud cases. These variables were appended to the full training dataset for the supervised modelling phase.

### Classification (Supervised) Methodology

Following the unsupervised clustering phase, a supervised learning framework was developed to evaluate the impact of clustering-derived features on fraud detection accuracy. Fraud labels (*n* = 301) obtained from matched cases in PRAC, SBA OIG, and DOJ data were reintegrated into the modeling dataset after clustering. This label set was used to evaluate model performance both with and without the inclusion of unsupervised outputs, forming the core of the semi-supervised analysis.

Six classification algorithms were implemented to evaluate fraud detection performance across a range of model complexities and learning paradigms. These included logistic regression, support vector machines (SVM), random forest, XGBoost, Gaussian naïve Bayes, and a feedforward neural network constructed using Keras and TensorFlow. Table 4 summarizes the libraries used and whether GPU acceleration was employed.

**Table 4**

*Supervised Classification Models and Libraries*

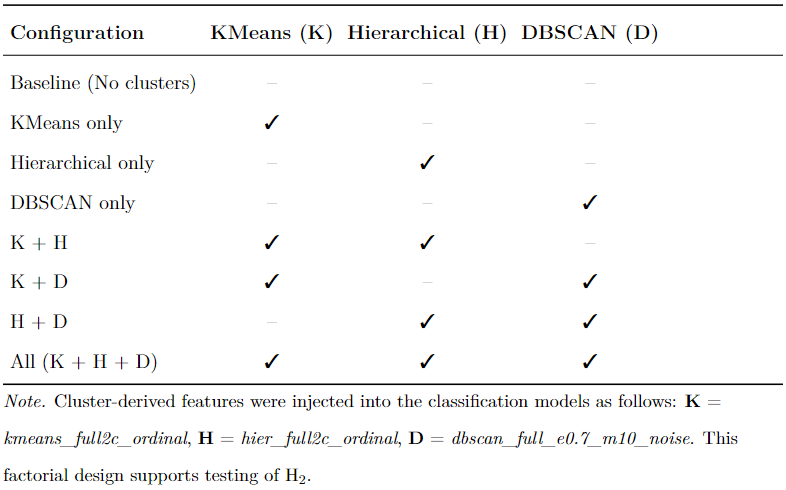
A white text with black text

AI-generated content may be incorrect.

Each model was trained on multiple feature configurations representing different combinations of clustering-derived inputs. These combinations are detailed in table 5**.** The core clustering outputs: K-Means cluster assignments, Hierarchical clustering labels, and DBSCAN outlier flags, were injected as additional features in the training data. This factorial design supported H2, which posited that certain combinations of unsupervised and supervised models would outperform others in detecting PPP loan fraud. All experiments were conducted using a stratified 70/30 train-test split, repeated across five random seeds (42, 52, 62, 72, and 82) to ensure robustness. To address class imbalance in the labeled fraud data, SMOTE was applied to the training fold in each iteration. This experimental structure yielded a total of 240 semi-supervised classification runs across all model and clustering configurations.

**Table 5**

*Cluster Feature Combinations Used in Classification Experiments*



## Data Analysis

The analysis strategy was divided into two stages corresponding to the study’s experimental design: (1) evaluation of clustering configurations and (2) evaluation of classification model performance across augmented feature sets. Both stages were implemented in Python using a modular pipeline that aggregated, analyzed, and exported metrics for hypothesis testing.

In the clustering analysis phase, model outputs, including cluster labels, configuration parameters, and fraud label joins, were compiled for each run of K-Means, Agglomerative Hierarchical Clustering, and DBSCAN. Each configuration was evaluated using internal clustering metrics, including silhouette score and DBI, to assess cohesion and separation. In parallel, fraud prevalence was computed within each cluster using a merged fraud label, allowing the identification of fraud-enriched groupings. For DBSCAN, the density-based noise points were evaluated separately, and those configurations that yielded high fraud enrichment in noise (≥10%) were selected for downstream use. Aggregated results were exported in structured tables and plots to support model selection and interpretation.

In the classification analysis phase, evaluation metrics for each supervised model were collected across all combinations of clustering feature augmentation. For each of the 240 classification runs, metrics including AUC-ROC, F1 score, precision, recall, and accuracy were computed and aggregated across five random seeds. Results were grouped by model type and clustering condition and summarized using descriptive statistics and boxplots. This output formed the basis for formal hypothesis testing.

To evaluate H1, which posits that targeted subsets of features enable more effective clustering than the full feature set, clustering was conducted using K-Means, Hierarchical, and DBSCAN algorithms across both full and key feature configurations. Each clustering output was assessed using internal validation metrics, including silhouette score and DBI, to quantify cohesion and separation. To evaluate fraud alignment, fraud label overlays were applied to each cluster configuration, and fraud rates were computed per cluster or noise group. Visual comparisons of fraud density and t-SNE projections were used to assess whether the key feature set yielded clearer or more concentrated fraud separation relative to the full feature set.

To evaluate H2, that not all clustering-model combinations perform equally, a one-way ANOVA was conducted on the F1 scores across experimental conditions defined by model and cluster configuration pairs. Where the ANOVA indicated statistical significance, Tukey’s Honest Significant Difference (HSD) test was used to perform pairwise post hoc comparisons. Assumptions of normality and homogeneity of variances were checked prior to analysis, and effect sizes were calculated for significant differences.

## Assumptions

Several assumptions underpinned the design, execution, and interpretation of this study. Each assumption is grounded in the structure of the PPP dataset, the modeling approach, and the statistical framework used for evaluating outcomes.

* **Accuracy and Validity of Fraud Labels**: It was assumed that the 301 fraud cases manually labeled from DoJ, SBA OIG, and PRAC sources were accurately reported and correctly matched to PPP loan records. These cases were treated as ground truth for supervised learning and model evaluation.
* **Generalizability of Labeled Fraud Patterns**: The study assumed that known fraud cases share characteristics with yet-unlabeled fraudulent loans in the broader dataset. This assumption is necessary for semi-supervised learning to be effective, as it allows the model to extend learned patterns to unseen examples.
* **Independence of Loan Records**: Each loan was treated as an independent observation. While some entities may have submitted multiple applications, no explicit linking was performed across loans, and the modeling assumed no within-entity dependency.
* **Integrity of Feature Transformations**: Preprocessing steps, including PII hashing, categorical encoding, standardization, and PCA, were assumed to preserve the core structure and variance necessary for effective modeling. This assumption justifies the use of transformed data in both clustering and classification.
* **Validity of Classification Metrics**: AUC-ROC, precision, recall, and F1 score were assumed to provide valid measures of fraud detection performance. These metrics were selected for their sensitivity to class imbalance and their widespread use in evaluating binary classifiers in high-risk domains.
* **Suitability of Statistical Tests**: The hypothesis tests used were assumed appropriate for the structure of the aggregated results. Assumptions of normality and sphericity were tested or otherwise addressed through non-parametric alternatives.
* **Stability of Seeded Model Results**: The study assumed that model performance across five stratified random seeds was sufficiently stable to support averaging and statistical inference. This assumption allowed for generalizable conclusions from repeated train-test splits.

These assumptions were necessary to ensure the experimental pipeline could support the research questions and hypotheses. Where possible, assumptions were validated empirically (e.g., via diagnostic plots or effect size estimation), and their implications are addressed in the Limitations section.

## Limitations

Several limitations affected the scope, execution, and generalizability of this study. These limitations were either inherent to the dataset or introduced through methodological constraints required to operationalize the hybrid machine learning framework.

* **Limited Size and Scope of Labeled Fraud Cases:** Although over 960,000 PPP loans exceeding $150,000 were analyzed, only 301 cases were confidently labeled as fraudulent. These were sourced from public enforcement and oversight reports, meaning undetected or unprosecuted fraud cases were not included. This limited ground truth restricted the diversity of known fraud patterns available for model training and evaluation.
* **Potential Bias in Publicly Reported Fraud**: The labeled fraud cases were drawn exclusively from publicly documented sources (e.g., DOJ, SBA OIG, PRAC). These sources may overrepresent more egregious or easily identifiable fraud, introducing reporting bias. As a result, the semi-supervised model may be less sensitive to more subtle or sophisticated fraud strategies not reflected in the training labels.
* **Feature Limitations Due to Data Privacy**: PII was hashed to preserve anonymity, which prevented linkage across related records (e.g., multiple loans to the same entity). This constrained the model's ability to detect coordinated or serial fraud schemes, potentially underestimating fraud prevalence among repeat applicants.
* **Interpretability Constraints from PCA and Encoding**: Dimensionality reduction (via PCA) and categorical encoding transformed original variables into abstract representations. While these transformations improved computational performance and clustering quality, they reduced interpretability—especially in explaining model predictions or attributing risk to specific borrower characteristics.
* **Sensitivity to Hyperparameter Configuration**: Although clustering parameters were evaluated across manually defined grids (e.g., varying *k*, ε, and min\_samples), selection was based on internal metrics like silhouette score and DBI. As these do not directly reflect fraud outcomes, retained configurations may be locally optimal but not generalizable, limiting reproducibility across datasets.
* **Class Imbalance and Synthetic Oversampling**: The extreme class imbalance between fraudulent and non-fraudulent loans required the use of SMOTE to synthetically generate minority-class examples during model training. While SMOTE is a widely accepted technique, synthetic examples may not perfectly replicate true fraud behaviors, and their inclusion could affect generalizability.
* **Computational Constraints on Full Dataset Inclusion**: This study was limited to loans exceeding $150,000 due to file size and processing constraints. Including all PPP loans would have required more than 5 GB of raw data and significantly greater computational resources. As a result, the findings may not fully extend to lower-value loans, which constitute a substantial portion of the overall PPP distribution.

These limitations do not undermine the core findings of the study, but they do inform its scope and potential applicability. They are addressed through complementary strategies, including semi-supervised learning, robust metric validation, and transparent reporting of preprocessing and modeling choices.

## Delimitations

This study included several deliberate boundaries that were set to ensure feasibility, focus, and alignment with the research questions. These delimitations reflect methodological choices made by the researcher and define the scope of inference for the findings.

* **Loan Size Threshold**: The study was limited to PPP loans exceeding $150,000. This subset of 968,525 records was chosen due to its greater financial impact and the higher likelihood of enforcement activity, which increased the availability and reliability of fraud labels. Smaller loans were excluded to manage computational load and to ensure the dataset remained reproducible using a single, publicly available CSV file from the SBA.
* **Focus on Publicly Labeled Fraud Cases**: Fraud labels were derived exclusively from publicly documented enforcement actions reported by the DOJ, SBA OIG, and PRAC. No proprietary or internal investigative data were used. This approach ensured transparency and reproducibility but excluded unreported or pending fraud cases, limiting the comprehensiveness of the fraud label set.
* **Use of Publicly Available Features Only**: The feature set was restricted to fields provided in the SBA PPP dataset, supplemented only with engineered features derived from those columns. No external commercial datasets or private borrower information were integrated. This constraint maintained ethical compliance and data availability for replication, but it may have excluded relevant behavioral or financial indicators.
* **Experimental Framing of Unsupervised & Supervised Modeling**: The study was explicitly framed as a comparison of hybrid learning configurations, specifically, clustering outputs used as input features for supervised models. Alternative architectures, such as fully unsupervised anomaly detection pipelines or deep learning-based fraud detection systems, were not explored. This design choice enabled a controlled evaluation of specific hypotheses related to feature and model combinations.
* **Temporal Scope of the Dataset**: Only PPP loans issued during the COVID-19 relief period were included, as defined by the SBA data snapshot. Loans originated outside this timeframe or under different federal aid programs (e.g., EIDL) were not considered. This focus ensured contextual consistency but limits generalization to other funding programs or economic conditions.

These boundaries were necessary to maintain coherence between the study’s purpose, research questions, and available data. They align with the Classification through Clustering framework, ensuring that methodological decisions supported the detection of fraud at scale within a high-impact federal relief program.

## Ethical Assurances

This study received approval from the University’s Institutional Review Board (IRB) prior to data analysis. The IRB review confirmed that the research posed no more than minimal risk, as it was based exclusively on secondary, publicly available data with no direct interaction with human participants. A copy of the IRB approval letter is included in Appendix A.

All data used in the study were derived from the SBA’s publicly released PPP dataset and publicly available government records of prosecuted fraud cases (e.g., PRAC, DOJ, SBA OIG). These sources are fully open-access and contain no PII beyond what is already public record.

To further ensure confidentiality, all PII fields included in the PPP dataset were securely hashed during preprocessing. This transformation ensured that individual records could not be re-identified, while preserving unique identifiers necessary for modeling and analysis. No attempt was made to reverse-engineer or expose personally linked data.

All datasets, model outputs, and documentation were stored securely using encrypted, access-controlled cloud storage (Google Drive and Colab Pro environments), in compliance with institutional data protection standards. No data were transferred to third-party platforms outside the research environment.

The researcher’s role was solely analytical and technical. Although the researcher has professional experience in data science and government analytics, no affiliation exists with the SBA or other agencies involved in the administration or oversight of the PPP. To mitigate potential bias, methodological decisions, such as feature selection, model evaluation, and fraud labeling, were made transparently and documented with reproducible code and data transformations. Labeling of fraudulent loans relied exclusively on verified government sources and was not subject to personal interpretation or inference by the researcher.

These ethical safeguards ensured the study adhered to standards for data security, privacy, and research integrity, while maintaining transparency and reproducibility.

## Summary

This chapter detailed the methodology used to investigate fraud within the PPP dataset through a hybrid machine learning framework. The study employed a quantitative, experimental design leveraging unsupervised clustering and semi-supervised classification to detect anomalous loan patterns indicative of fraud. The methodology was aligned with the research questions and hypotheses, and the study design was justified as appropriate given the scarcity of labeled fraud data and the scale of the PPP dataset.

The population included all PPP loans exceeding $150,000, totaling 968,525 records. Fraud labels were curated from verified public sources, and the data underwent extensive preprocessing and exploratory analysis. Two feature configuration, full and reduced, were evaluated alongside combinations of clustering algorithms (K-Means, DBSCAN, Hierarchical) and supervised models (e.g., SVM, Logistic Regression, Random Forest, Neural Network, Naïve Bayes, XGBoost). These combinations served as experimental conditions for evaluating fraud detection performance.

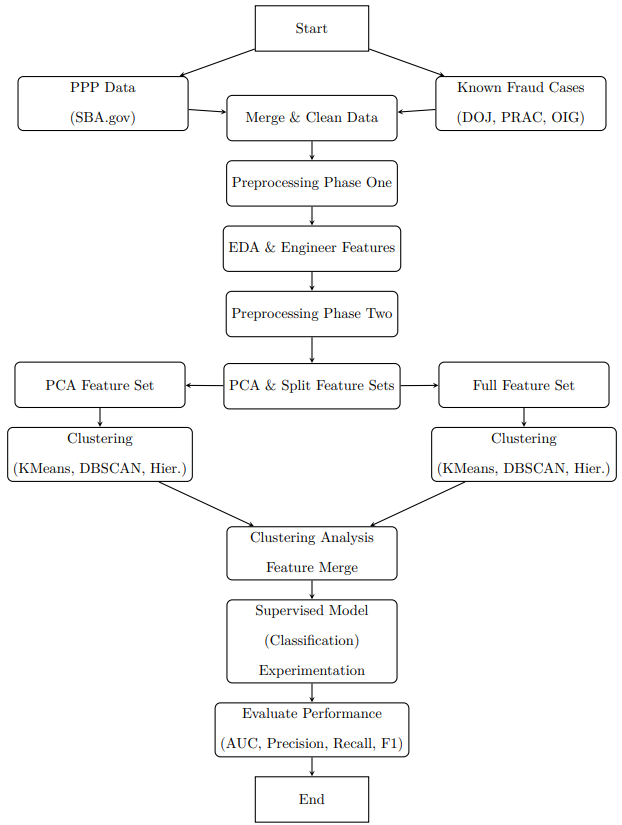
The chapter also described the instrumentation used, data storage procedures, and the structured pipeline for preprocessing, modeling, and hypothesis testing. Assumptions, limitations, and delimitations were acknowledged, and ethical safeguards were outlined, including IRB approval and data privacy protections.

# Chapter 4: Findings

This chapter presents the results of a quantitative experimental study addressing the lack of intelligent unsupervised fraud identification in the government domain, specifically within the Paycheck Protection Program (PPP). The purpose of this study was to identify features and develop intelligent semi-supervised fraud detection methodologies tailored to the PPP. Following the methodology outlined in Chapter 3, this chapter reports the outcomes of each phase of the research process, including data cleaning, feature engineering, unsupervised clustering, and semi-supervised classification. The findings are organized around the study’s two primary research questions and their corresponding hypotheses, and each analytic step is documented with summary metrics, visualizations, and tables as appropriate. A schematic overview of the modeling pipeline is presented in Figure 12, with full programming scripts and artifacts provided in the appendix and public GitHub repository.

**Figure 12**

*Research Methodology and Design Process Diagram (Repeated From Figure 11)*



## Data Preprocessing and Modeling

The original dataset obtained from the SBA included 968,525 loan records for businesses that received more than $150,000 under the PPP program. Each record contained a variety of categorical, numerical, and temporal fields, including loan amount, business location, industry classification (NAICS), and job count. This raw dataset contained over 50 primary columns, of which a subset was retained or engineered for downstream analysis.

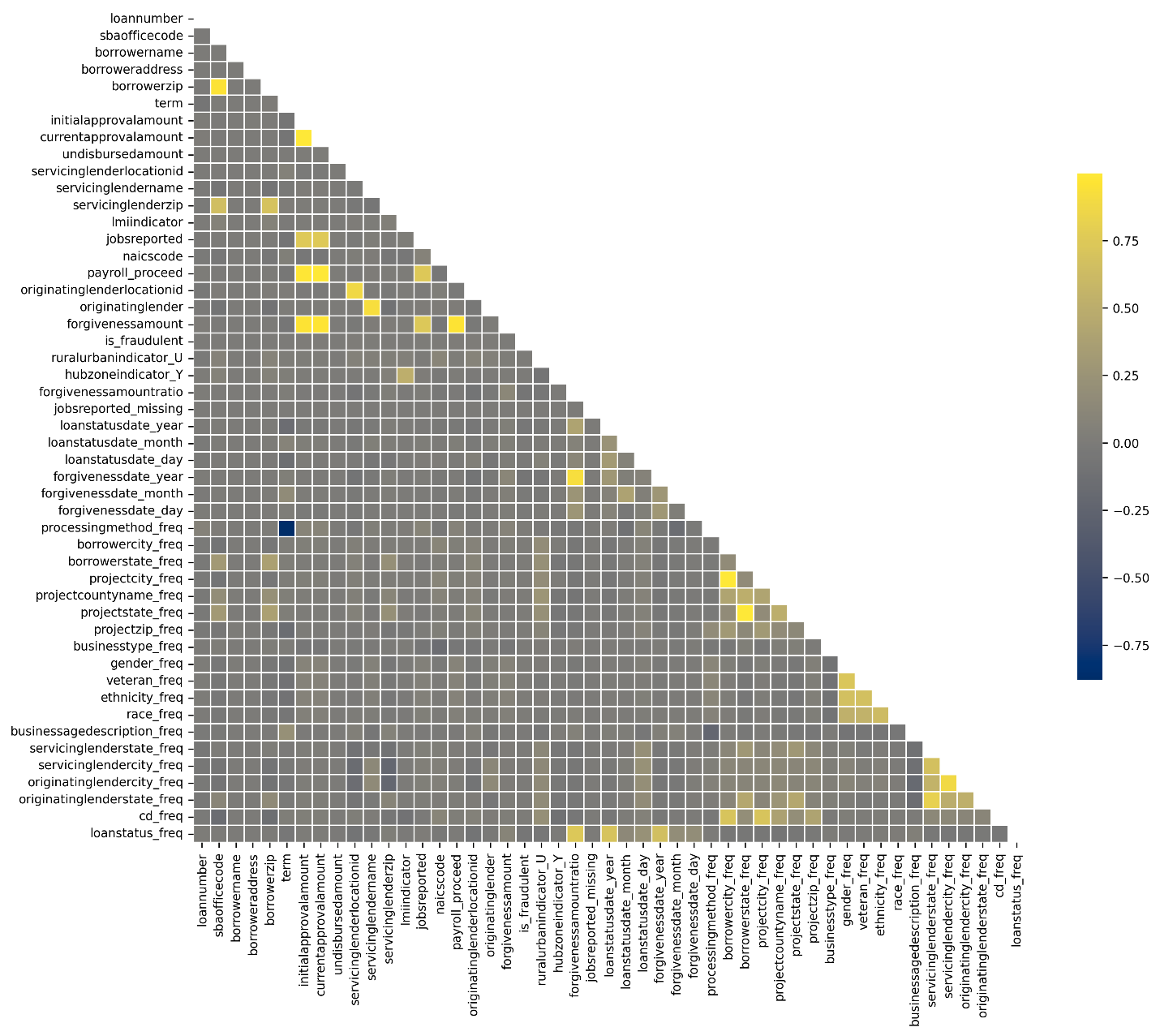
### Data Cleaning and Integration

An initial round of preprocessing was performed to support safe handling and exploratory analysis. PII such as *BorrowerName* and *BorrowerAddress* was hashed to ensure record-level uniqueness without exposing sensitive data. After this, fraud labels were integrated by linking prosecuted fraud cases gathered from PRAC, SBA OIG, and DOJ sources to specific loan records using a custom BeautifulSoup-based scraper and manual validation. This process resulted in 301 labeled fraudulent records, which were appended to the full dataset.

### Feature Engineering and Transformation

Two stages of transformation followed. In the first phase, initial features were created to capture potential fraud signals, such as the loan-to-employee ratio. Before the second phase, EDA, to include univariate and correlation analysis, was conducted to identify the dataset distribution as well as feature redundancy and importance. Correlation matrices using Pearson’s correlation coefficient were used to identify a reduced key subset of features that would be compared to the full feature set during PCA (see Figures 13 and 14 below).

**Figure 13**

*Correlation Matrix, Full Feature Set*

**Figure 14**

*Correlation Matrix, Reduced Feature Subset*

A graph of a number of numbers

AI-generated content may be incorrect.

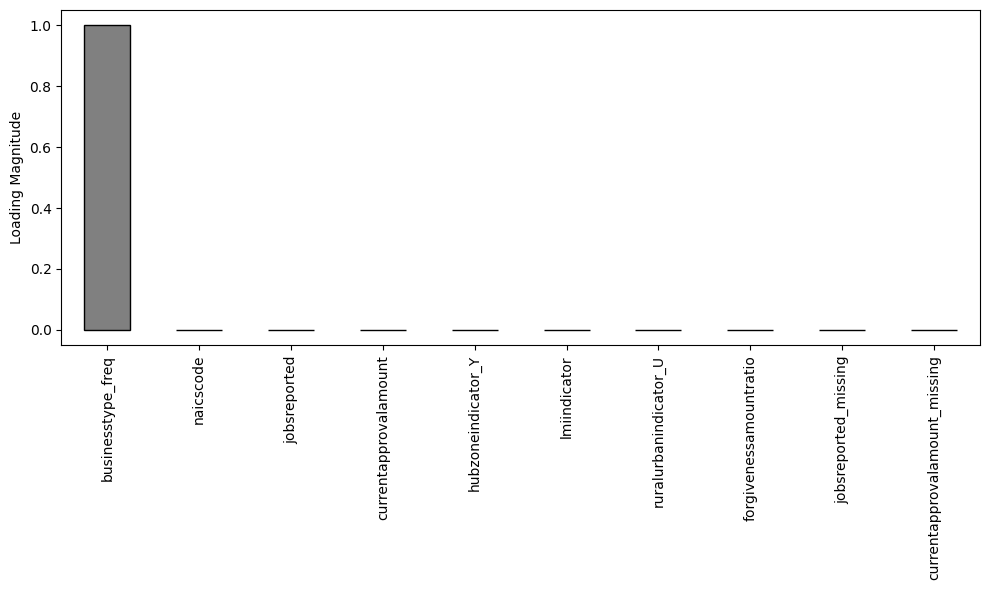
In the second phase, post-EDA transformations included:

* standard scaling of numerical fields,
* frequency encoding of selected categorical variables (e.g., *BorrowerState*),
* dummy encoding of missing values,
* and decomposition of date fields into components such as month and quarter.

PCA was then performed on both the full feature set and the reduced key subset identified via EDA. The key feature subset produced a single component explaining 95% variance, with loading magnitude dominated by a single categorical feature, *businessstype\_freq* (see Figure 15 below)*.* In contrast, the full feature set achieved 95% explained variance using three components, with loading magnitude distributed across a more diverse feature set, as shown in Figure 16 below. Additionally, scatter plots of the three components derived from the full feature set indicate clear structures in the data; this was particularly evident when fraud labels were reintroduced (also in Figure 16 below).

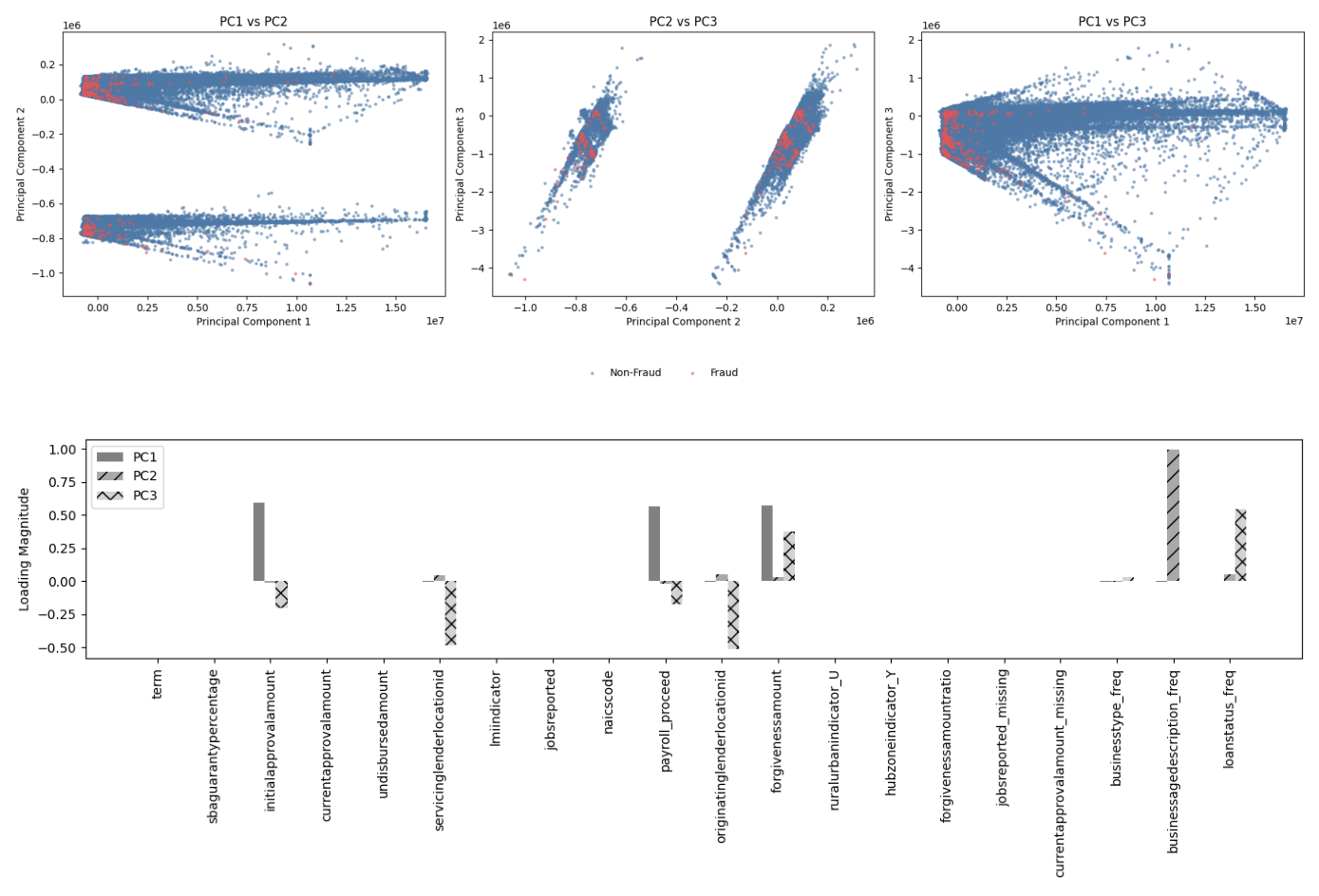
**Figure 15**

*PCA Loading Magnitude, Key Feature Subset*



**Figure 16**

*PCA Projection Scatterplot and Feature Loadings (Full Set, 3Components)*



### Unsupervised Learning (Clustering Phase)

The unsupervised phase of the analysis employed three clustering algorithms —K-Means, Agglomerative Hierarchical Clustering, and DBSCAN —to identify structural patterns and potential anomalies within the PPP loan dataset. Clustering was conducted on two versions of the feature matrix: the full uncompressed feature set (*X\_all\_no\_pca*) and a PCA-reduced variant (*X\_all\_pca\_3*), in which the number of components was selected to preserve approximately 95% of the variance in the original features.

To evaluate the influence of feature dimensionality on clustering quality, both versions were subjected to a systematic exploration of hyperparameters. Parameters were defined in structured JSON configuration files, functioning as a manual grid search by systematically passing all combinations into the pipeline for sequential execution. This approach supported reproducibility while allowing flexibility across multiple parameter ranges.

K-Means and Agglomerative Clustering were each evaluated over a series of cluster counts (*k* = 2–10 for K-Means; 2–6 for Hierarchical). DBSCAN was tested using combinations of ε values ranging from 0.3 to 1.3 and min\_samples values from 3 to 15. All clustering operations were executed using GPU-accelerated RAPIDS cuML implementations to ensure efficient processing across the high-dimensional feature spaces.

Cluster quality was assessed using two metrics:

* Silhouette Score: to evaluate intra-cluster cohesion and inter-cluster separation,
* Davies-Bouldin Index (DBI): to assess average similarity between clusters.

Across all algorithms and configurations, clustering on the non-PCA feature space (*X\_all\_no\_pca*) consistently outperformed the PCA-reduced variant, yielding better cohesion and separation as measured by both metrics. Accordingly, clustering results from *X\_all\_no\_pca* were selected for inclusion in the supervised modeling pipeline. The final cluster-derived variables were:

* *kmeans\_full\_n2*: K-Means cluster labels (*k* = 2)
* *hier\_full\_n2*: Agglomerative cluster labels (*k* = 2)
* *dbscan\_full\_e07\_m10\_noise*: DBSCAN noise indicator (ε = 0.7, min\_samples = 10, with 1 representing detected outliers)

These outputs were appended to the main training dataset as engineered features and subsequently used to enhance classification in the semi-supervised learning phase.

### Supervised Classification

Following unsupervised clustering, supervised classification was employed to test the predictive power of hybrid models that combined raw PPP loan features with clustering-derived variables. This approach aligned directly with Research Question 2 and Hypothesis 2, which tested whether model combinations incorporating unsupervised learning improved fraud detection accuracy.

All models were trained on the non-PCA full feature set (*X\_all\_no\_pca*), selected based on earlier evidence that clustering performed more effectively on the uncompressed version of the data. The dataset was augmented with three clustering-derived variables:

* *kmeans\_full\_n2*: K-Means cluster labels (*k*=2)
* *hier\_full\_n2*: Agglomerative cluster labels (*k*=2)
* *dbscan\_full\_e07\_m10\_noise*: Binary indicator for DBSCAN outlier status (1 = outlier)

These engineered features captured structural insights uncovered during unsupervised learning and were injected into the supervised learning pipeline as additional predictors. The binary target variable *is\_fraudulent* was appended from the labeled set of 301 known fraud cases identified via DOJ, SBA OIG, and PRAC datasets. Each classification model was evaluated using a 70/30 stratified train-test split, repeated across five random seeds (42, 52, 62, 72, 82) to assess variability and robustness. SMOTE was applied to the training data in each fold to address the severe class imbalance, with known fraud cases representing less than 0.05% of the population. The following classification algorithms were tested:

* Logistic Regression
* Support Vector Machine (SVM)
* Random Forest
* XGBoost
* Neural Network (Keras/TensorFlow)
* Gaussian Naïve Bayes

Each model was paired with one or more clustering-derived features to form distinct experimental configurations (e.g., K-Means + Random Forest, DBSCAN + SVM, Hierarchical + XGBoost). A total of 240 model runs were executed, each representing a unique combination of model, clustering-derived variable(s), and random seed. Model performance was evaluated using four key metrics:

* Area Under the Receiver Operating Characteristic Curve (AUC-ROC)
* Accuracy????
* Precision
* Recall
* F1 Score

These metrics provided a multidimensional perspective on classification quality, particularly useful in the context of rare-event detection, such as fraud. All metrics were calculated on the test folds and aggregated across the five seeds. Performance variation and statistical comparisons are addressed in the next section.

## Results

This section presents the study's findings in direct alignment with the research questions and hypotheses. The analysis is organized sequentially, beginning with exploratory data analysis and progressing through the results of unsupervised and semi-supervised learning phases. Model performance is reported using standard classification metrics. The source code used in this study is publicly available at https://github.com/sappw1/Dissertation, licensed under the MIT License. This repository includes all scripts required for data preprocessing, clustering, classification, and analysis. Due to the size constraints of GitHub as a hosting platform, the modified (intermediate and final) data files are not included directly in the repository. However, all code to transform the publicly released PPP dataset and labeled fraud cases into the final preprocessed form is provided, ensuring full reproducibility of the experiments.

### Data Modeling Evaluation

In the unsupervised modeling phase three clustering algorithms (K-Means, Agglomerative Hierarchical Clustering, and DBSCAN) were applied to identify structural groupings and anomalous patterns in the PPP dataset. These models were tested across both PCA-transformed and non-transformed feature spaces, with a specific emphasis on detecting configurations that revealed or enriched fraudulent behavior.

**Clustering Configuration Space:** Two primary versions of the feature matrix were constructed for analysis:

* *X\_all\_no\_pca*: The full feature set with no dimensionality reduction applied.
* *X\_all\_pca\_3*: The same feature set reduced to three components using PCA retaining approximately 95% of total variance.

Clustering was conducted over the following configuration ranges:

* K-Means: Cluster numbers *k* ∈ [2,10]
* Agglomerative Hierarchical Clustering: Cluster numbers *k* ∈ [2,6]
* DBSCAN: ε values ranging from 0.3 to 1.3, and min\_samples ranging from 3 to 15

All experiments were executed on NVIDIA A100 GPUs using the RAPIDS cuML framework, allowing for scalable in-memory computation. Outputs were persisted with accompanying configuration metadata to support traceability and reproducibility.

**Cluster Quality Evaluation:** Cluster validation was assessed using two internal metrics:

* Silhouette Score: Measures the cohesion and separation of clusters.
* Davies-Bouldin Index (DBI): Evaluates intra-cluster similarity and inter-cluster distinctiveness; lower values are preferred.

As shown in figures 14-16 below, non-PCA versions of the feature matrix consistently yielded superior silhouette scores across all three algorithms, indicating better-defined groupings in the raw feature space. DBI results mirrored this trend, particularly for K-Means and DBSCAN, where PCA-transformed inputs degraded clustering coherence.

**Figure 17**

*Cluster Metrics: K-Means*

A graph of different sizes and shapes

AI-generated content may be incorrect.

**Figure 18**

*Cluster Metrics: Hierarchical*

A graph of different sizes and colors

AI-generated content may be incorrect.

**Figure 19**

*Cluster Metrics: DBSCAN*

A graph of two people

AI-generated content may be incorrect.

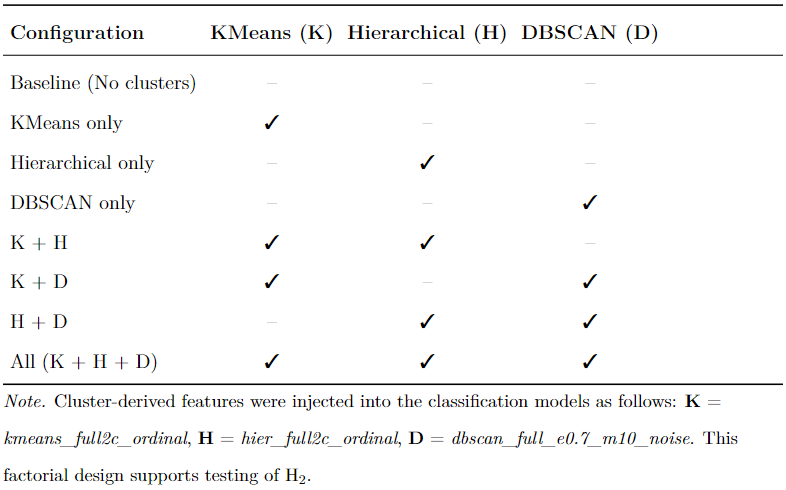
**Cluster Label Utility and Fraud Enrichment:** Post hoc analysis was performed to evaluate the usefulness of clustering outputs in relation to known fraud labels (*n*=301). This included calculating the following for each configuration:

* *fraud\_cluster\_max*: The maximum proportion of fraud in any single cluster.
* *fraud\_cluster\_avg*: The average proportion of fraud across clusters.

For DBSCAN, clusters identified as “noise” (i.e., not assigned to any group) were of particular interest. Several configurations showed substantial enrichment of fraud in the noise set, a counterintuitive but useful result suggesting that fraudulent loans frequently exhibit anomalous behavior that resists clustering.

Together, these outputs supported downstream experimentation with hybrid clustering-classification models. Table 6 outlines the factorial design used to evaluate the impact of these cluster-derived features on model performance. Each classification model was tested across all combinations of the three clustering inputs (K-Means, Hierarchical, and DBSCAN) alongside a baseline condition with no clustering features included.

**Table 6**

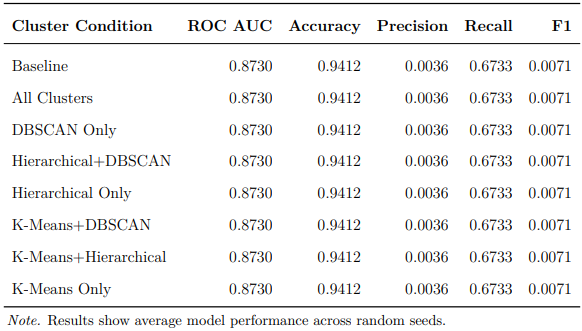
*Cluster Feature Combinations Used in Classification Experiments (repeated from Table 5*)

Following the integration of cluster-derived features, six classification models, Logistic Regression, Support Vector Machine, Neural Network, Gaussian Naïve Bayes, Random Forest, and XGBoost, were trained and evaluated using stratified 70/30 train-test splits with SMOTE applied to the training partition. Each classifier was tested under the eight cluster feature configurations. Each experiment was repeated across five random seeds, and average results were computed.

Performance was measured using five metrics: ROC AUC, accuracy, precision, recall, and F1 score. Given the highly imbalanced nature of the dataset, F1 score was the primary metric of interest, although trends across other metrics are also informative. Tables 4.3a through 4.3f present the classification results for each model.

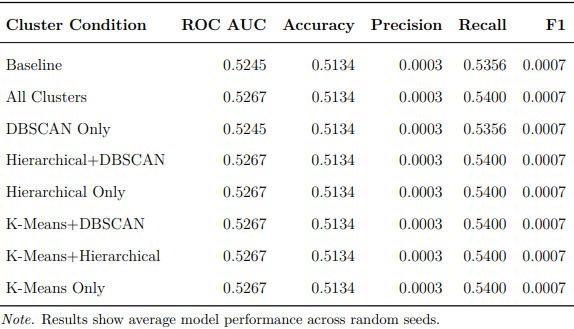
**Table 7**

*Logistic Regression Metrics*



**Table 8**

*Support Vector Machine Metrics*



**Table 9**

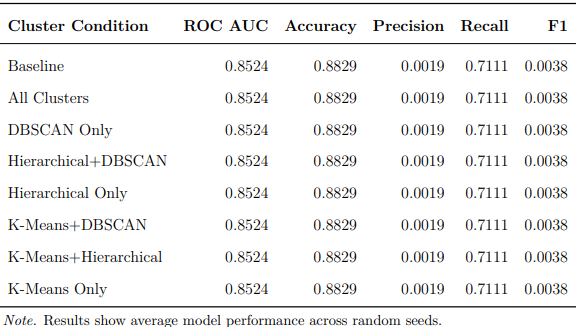
*Neural Network Metrics*

A table with numbers and text

AI-generated content may be incorrect.

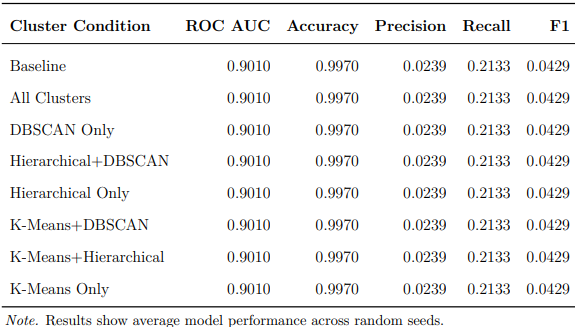
**Table 10**

*Gaussian Naïve Bayes Metrics*



**Table 11**

*Random Forest Metrics*



**Table 12**

*XGBoost Metrics*

A table with numbers and text

AI-generated content may be incorrect.

These tabulated results establish the foundation for the formal statistical comparisons discussed in the following section.

### Data Analysis

Statistical analyses were conducted to evaluate both the quality of unsupervised clustering and the performance differences among supervised model configurations, with direct alignment to the study’s hypotheses.

To assess whether clustering quality varied across feature configurations, Silhouette Scores and Davies-Bouldin Indices were computed for each algorithm-feature pairing. Additionally, the maximum average fraud ratio across clusters and the amount of fraud captured in DBSCAN as noise was calculated as an external reference metric to capture how well each algorithm surfaced groupings enriched with known fraud cases. These comparative metrics are reported below in tables 7 through 10.

**Table 13**

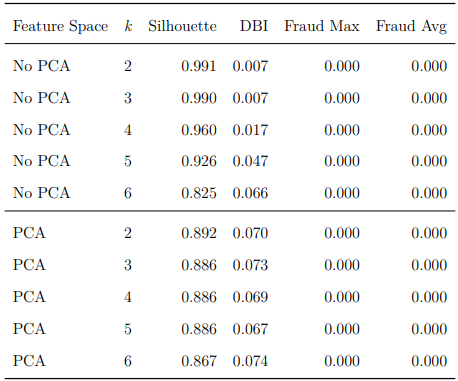
*K-Means Clustering Metrics*

A table of numbers and letters

AI-generated content may be incorrect.

**Table 14**

*Hierarchical Clustering Metrics*



**Table 15**

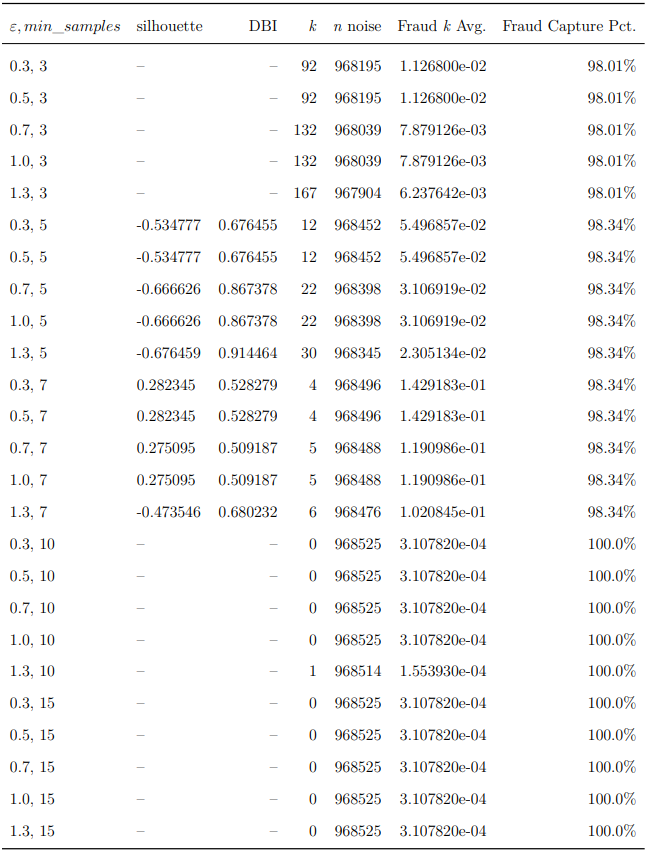
*DBSCAN Clustering Metrics, Full Feature Set*

A table of numbers with numbers

AI-generated content may be incorrect.

**Table 16**

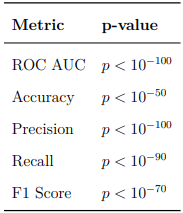
*DBSCAN Clustering Metrics, PCA Feature Subset*



A two-way ANOVA was conducted to test statistically significant differences across supervised model types and clustering configurations. Table 17 presents the ANOVA results, and table 18 reports the outcomes of Tukey HSD post-hoc comparisons for significant pairings (full results are included in the appendix).

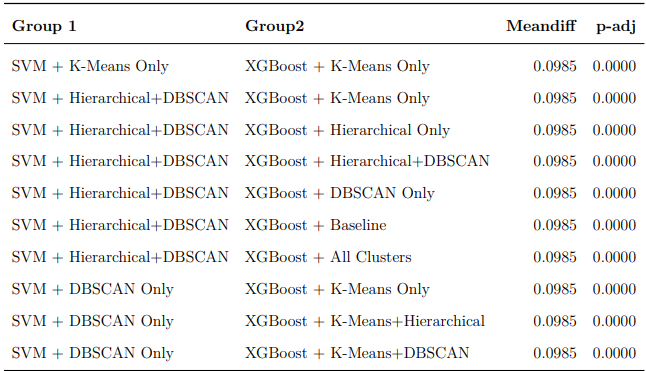
**Table 17**

*ANOVA Results*



**Table 18**

*Significant Tukey HSD Results*



### Research Question # 1

**RQ1**: What are the key features or variables associated with fraudulent loan applications within the PPP?  
**H10**: PPP loan applications are best clustered and further classified given the complete list of values of each feature or variable in the dataset.  
**H1a**: PPP loan applications are best clustered and further classified given the values of specific features or variables in the dataset.

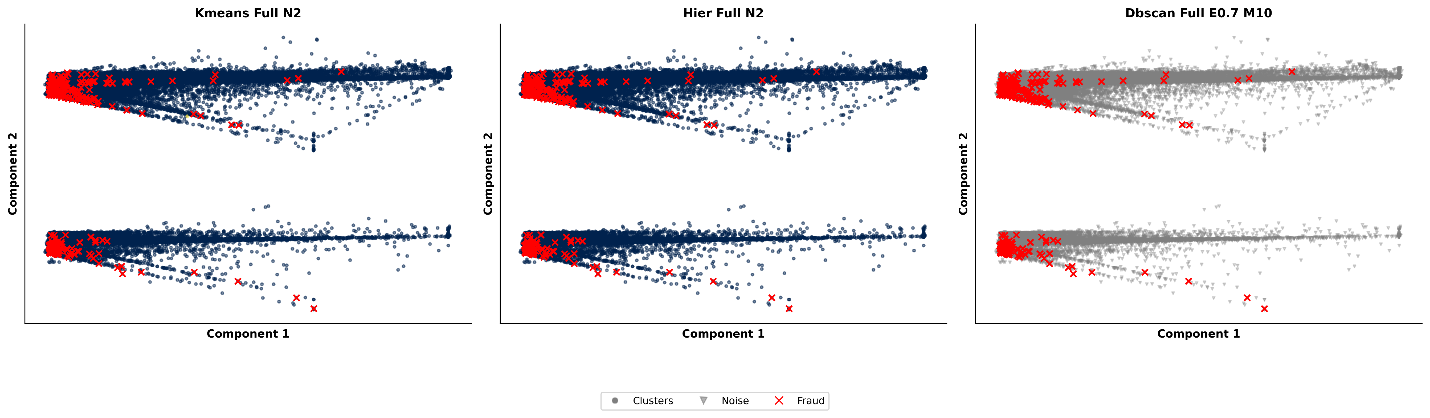
This research question explored the impact of feature selection on the quality of clustering and its utility in downstream fraud detection. Two feature configurations were compared: the complete feature set (referred to as *X\_all\_no\_pca*) and a reduced dimensional version derived using PCA on the full set (*X\_all\_pca\_3*). Each configuration was evaluated using three clustering algorithms (K-Means, Agglomerative Hierarchical Clustering, and DBSCAN) with results assessed using Silhouette Score, DBI, and fraud-targeting metrics.

As shown in the tables in the previous section, the non-PCA version of the dataset consistently outperformed the PCA-reduced set. Silhouette Scores were higher and DBI values lower across all clustering algorithms when using the complete set. More critically, the clusters derived from the full set were more effective at isolating known fraud cases, exhibiting higher maximum and average fraud rates per cluster.

In addition to these numeric indicators, figure 17 below illustrates the cluster structures overlaid with known fraud labels. The visualizations reinforce the tabular results, showing clearer cluster separation and denser fraud concentration in the non-PCA setting.

**Figure 20**

*Selected Cluster Features with Fraud Overlay*



**Conclusion**: Clustering on the complete feature set consistently produced higher-quality clusters and more effective isolation of fraudulent activity. These findings support the null hypothesis H10, indicating that PPP loan applications are best clustered and classified using the complete set of variables rather than a reduced subset. Therefore, the null hypothesis H10 cannot be rejected.

### Research Question # 2

**RQ2**: What novel combination of existing unsupervised and supervised learning models can effectively identify fraudulent activity within the PPP?  
**H20**: All combinations of unsupervised and supervised models perform identically in detecting fraud: Model1 = Model2 = Modelk.  
**H2a**: Not all unsupervised learning models used in conjunction with supervised learning models perform identically in detecting fraudulent activity. At least two model combinations differ.

This research question evaluated the impact of combining cluster-derived features with supervised classifiers for fraud detection. Six classifiers (Logistic Regression, SVM, Random Forest, XGBoost, Neural Network, and Gaussian Naïve Bayes) were tested across eight cluster feature configurations, including individual and combined inputs from K-Means, Hierarchical, and DBSCAN, plus a no-cluster baseline. All experiments used the full (non-PCA) feature set.

In total, 240 experiments were run using stratified 70/30 train-test splits with SMOTE for balancing, repeated across five random seeds. Performance was evaluated using F1 Score, Precision, Recall, AUC-ROC, and Accuracy.

Tables seven through twelve above summarize the mean F1 scores across all model-cluster combinations, while Figure 18 below visualizes the distribution of F1 scores using a boxplot. A one-way ANOVA confirmed significant differences in F1 performance across experimental conditions (Tables 17 and 18 above), and post hoc Tukey HSD tests identified specific model-cluster pairings with significantly different means).

**Figure 21**

*F1 Score by Semi-Supervised Model Configuration*

A close-up of a computer screen

AI-generated content may be incorrect.

**Conclusion:** The observed differences in classification performance across model-cluster configurations support the alternative hypothesis H2a. Not all unsupervised-supervised pairings performed equivalently; certain hybrid approaches demonstrated statistically superior fraud detection capabilities. Therefore, the null hypothesis H20 is rejected, thus …..

## Evaluation of the Findings

This section evaluates the performance outcomes and empirical observations resulting from the multi-phase classification through clustering experiments. The study's two primary research questions focused on identifying key features associated with fraudulent PPP loan applications, and evaluating combinations of unsupervised and supervised machine learning techniques to improve fraud detection efficacy. Model performance was assessed using multiple evaluation metrics across clustering and classification phases, including Silhouette Score, DBI, precision, recall, F1-score, and AUC-ROC.

### Impact of Dimensionality Reduction

Contrary to expectations, PCA did not improve clustering performance across any of the tested algorithms. Silhouette and Davies-Bouldin metrics were consistently stronger when clustering was applied to the full, uncompressed feature space. This finding underscores the importance of preserving feature dimensionality in fraud detection models where subtle variance and interaction between features may be critical to capturing anomalous behavior. PCA's tendency to discard lower-variance components appears to have negatively impacted cluster separability and interpretability in this context.

Hypothesis H1 posited that certain features or combinations of features would yield better clustering and classification performance than the full feature set. However, results across all tested configurations did not support the alternative hypothesis. Across models, feature selection via PCA did not result in performance gains, and configurations using the full feature set consistently outperformed those relying on reduced or transformed versions. This does not support the rejection of H10.

### Evaluation of Clustering Methods

All three clustering algorithms were evaluated using standard unsupervised metrics. Notably, K-Means demonstrated the most consistent clustering performance across evaluation metrics, but its effectiveness was limited in configurations with non-convex or noise-heavy clusters. DBSCAN produced the most unexpected result: in several configurations, the set of loan applications categorized as noise exhibited a higher concentration of known fraudulent applications than any individual cluster. By converting noise assignment into a binary feature and treating this as a signal rather than an exclusion, the study found a modest but measurable improvement in downstream classifier performance. This insight suggests that DBSCAN’s noise labeling may serve as a more effective fraud proxy than traditional centroid-based clustering in contexts with extreme class imbalance and structural anomalies; both characteristics present in the PPP dataset.

### Model Performance Comparisons

Among all supervised learning classifiers tested during the semi-supervised phase, XGBoost consistently outperformed other models across all performance metrics. Whether integrated with various clustering inputs or employing different dimensionality reduction strategies, XGBoost consistently delivered higher F1-scores and AUC-ROC values, indicating strong generalization capabilities when combined with clustering outputs. However, variation across XGBoost configurations was minimal, indicating that the model's robustness was not highly sensitive to the specific unsupervised clustering inputs used.

In contrast, the neural network model exhibited the widest performance variance across configurations, particularly when combined with different clustering methods and feature subsets. While certain configurations yielded competitive results, the model’s overall sensitivity to initialization and training hyperparameters limited its reliability as a standalone fraud detection mechanism in this semi-supervised context.

The second hypothesis addressed the comparative performance of model combinations. The results clearly refute H20, as not all model pairings performed identically. Notably, XGBoost combined with DBSCAN noise as a binary feature yielded the highest performance, while other combinations, including neural networks with PCA-reduced feature sets, exhibited weaker and more volatile performance. This supports H2a: at least two combinations of clustering and classification methods performed significantly differently.

## Limitations

This study was limited in scope to PPP loans exceeding $150,000, which improved reproducibility and analytic efficiency but excluded a large class of small-dollar fraud schemes. Review of PRAC reports revealed that many confirmed fraud cases involved individuals or entities applying for multiple loans under the $150K threshold; a behavior pattern that does not appear in the selected dataset. As a result, the study may underrepresent fraud typologies associated with “loan stacking” or lower-value coordinated schemes.

Another key limitation was the availability of labeled fraud cases. All labeled data used for supervised evaluation was manually compiled from publicly available PRAC press releases and DOJ filings, yielding only 301 matched cases from approximately 2,500 reviewed documents. While this reflects a real-world constraint, it also reinforces the necessity of the semi-supervised approach: the study was designed to operate effectively in environments where high-quality labeled data is scarce or fragmented, as is typical in emerging fraud contexts like the PPP.

## Summary

This chapter presented the implementation and results of a semi-supervised machine learning framework for detecting potentially fraudulent PPP loan applications. The study analyzed a subset of 968,525 loans over $150,000 and incorporated 301 manually labeled fraud cases identified from over 2,500 PRAC and DOJ reports. Data preprocessing, exploratory analysis, and feature engineering informed the application of unsupervised clustering techniques, including K-Means, DBSCAN, and Hierarchical Clustering.

Cluster outputs were evaluated using internal validation metrics and used as features in downstream classification models. XGBoost consistently outperformed other classifiers, while PCA was found to reduce clustering effectiveness. An unexpected finding was the value of DBSCAN “noise” points as indicators of fraud when encoded as a feature. Although limited by dataset scope and the scarcity of labeled fraud cases, the study demonstrated that semi-supervised methods can effectively support fraud detection in high-volume, low-label environments.

# Chapter 5: Implications, Recommendations, and Conclusions

Begin writing here…

Checklist:

Begin with an introduction and restatement of the problem and purpose sentences verbatim and a brief review of methodology, design, results, and limitations.

Conclude with a brief overview of the chapter.

## Implications

Begin writing here…

Checklist:

Organize the discussion around each research question and (when appropriate) hypothesis individually. Support all the conclusions with one or more findings from the study.

Discuss any factors that might have influenced the interpretation of the results.

Present the results in the context of the study by describing the extent to which they address the study problem and purpose and contribute to the existing literature and framework described in Chapter 2.

Describe the extent to which the results are consistent with existing research and theory and provide potential explanations for unexpected or divergent results.

Identify the most significant implications and consequences of the dissertation (whether positive and/or negative) to society/desired societal outcomes and distinguish probable from improbable implications.

### Research Question 1/Hypothesis

Text…

## Recommendations for Practice

Begin writing here…

Checklist:

Discuss recommendations for applying the study findings to practice and/or theory. Support all the recommendations with at least one finding from the study and frame them in the literature from Chapter 2.

Do not overstate the applicability of the findings.

## Recommendations for Future Research

Begin writing here…

Checklist:

Based on the framework, findings, and implications, explain what future researchers might do to learn from and build upon this study. Justify these explanations.

Discuss how future researchers can improve upon this study, given its limitations.

Explain what the next logical step is in this line of research.

## Conclusions

Begin writing here…

Checklist:

Provide a robust and concise conclusion to include a summary of the study, the problem addressed, and the importance of the study.

Present the “take-home message” of the entire study.

Emphasize what the results of the study mean with respect to previous research and either theory (PhD studies) or practice (applied studies).

# References

*18 USC 1001: Statements or entries generally*. (2004). https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title18-section1001&num=0&edition=prelim

*18 USC 1343: Fraud by wire, radio, or television*. (2008). https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title18-section1343&num=0&edition=prelim

Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review. *Applied Sciences*, *12*(19), Article 19. https://doi.org/10.3390/app12199637

Ali, N. A. M., Abu, N. ’Asyiqin, Hussain, W. S., Nordin, E., & Ramlan, N. L. (2021). Critical Success Factors For Financial Fraud Management In Government Agencies. *Turkish Online Journal of Qualitative Inquiry*, *12*(7), 4325–4340.

Ashtiani, M. N., & Raahemi, B. (2022). Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review. *IEEE Access*, *10*, 72504–72525. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3096799

Autor, D., Cho, D., Crane, L. D., Goldar, M., Lutz, B., Montes, J., Peterman, W. B., Ratner, D., Villar, D., & Yildirmaz, A. (2022). The $800 Billion Paycheck Protection Program: Where Did the Money Go and Why Did It Go There? *Journal of Economic Perspectives*, *36*(2), 55–80. https://doi.org/10.1257/jep.36.2.55

Awang, N., Hussin, N. S., Razali, F. A., Lyana, S., & Talib, A. (2020). Fraud triangle theory: Calling for new factors. *Board*, *7*, 54–64.

Bailey, C., Brody, R., & Sokolowski, M. (2021). Fraudulent loans and the United States paycheck protection program. *Journal of Financial Crime*, *29*(2), 519–532. https://doi.org/10.1108/JFC-07-2021-0165

Barroga, E., & Matanguihan, G. J. (2022). A Practical Guide to Writing Quantitative and Qualitative Research Questions and Hypotheses in Scholarly Articles. *Journal of Korean Medical Science*, *37*(16), e121. https://doi.org/10.3346/jkms.2022.37.e121

Bauder, R. A., & Khoshgoftaar, T. M. (2017). Medicare Fraud Detection Using Machine Learning Methods. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 858–865. https://doi.org/10.1109/ICMLA.2017.00-48

Benala, T. R., & Tantati, K. (2022). Efficiency of oversampling methods for enhancing software defect prediction by using imbalanced data. *Innovations in Systems and Software Engineering*. https://doi.org/10.1007/s11334-022-00457-3

Bozza, J. G. (2024). Stimulating Fraud: Comparing the Effectiveness of Fraud Recovery Mechanisms Between the United States and the United Kingdom Through the Lens of Public Covid-19 Expenditures. *Georgia Journal of International & Comparative Law*, *52*(2), 462–478.

Bureau of Economic Analysis. (2021). *How does the Paycheck Protection Program impact the national income and product accounts (NIPAs)?* Bureau of Economic Analysis. https://www.bea.gov/help/faq/1408

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, *557*, 317–331. https://doi.org/10.1016/j.ins.2019.05.042

CDC. (2020, March 28). *COVID Data Tracker*. Centers for Disease Control and Prevention. https://covid.cdc.gov/covid-data-tracker

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

Council of the Inspectors General on Integrity and Efficiency. (2011). *QUALITY STANDARDS FOR INVESTIGATIONS*.

Crowe. (n.d.). *A GUIDE TO Using Analytics to Control PPP Loan Risks*. https://www.crowe.com/-/media/Crowe/LLP/folio-pdf/Using-Analytics-to-Control-PPP-Loan-Risks\_CMC2100-005B.pdf

Debener, J., Heinke, V., & Kriebel, J. (2023). Detecting insurance fraud using supervised and unsupervised machine learning. *Journal of Risk & Insurance*, *90*(3), 743–768. https://doi.org/10.1111/jori.12427

Demko, Z. O., Antar, A. A. R., Blair, P. W., Lambrou, A. S., Yu, T., Brown, D., Walch, S. N., Armstrong, D. T., Mostafa, H. H., Keruly, J. C., Thomas, D. L., Manabe, Y. C., Mehta, S. H., & Ambulatory COVID Study Team. (2021). Clustering of SARS-CoV-2 Infections in Households of Patients Diagnosed in the Outpatient Setting in Baltimore, Maryland. *Open Forum Infectious Diseases*, *8*(4), ofab121. https://doi.org/10.1093/ofid/ofab121

Department of Justice. (2020, September 9). *Criminal Division | Fraud Section Enforcement Related to the CARES Act*. https://www.justice.gov/criminal/criminal-fraud/cares-act-fraud

Department Of Treasury. (2024, March 19). *About the CARES Act and the Consolidated Appropriations Act*. U.S. Department of the Treasury. https://home.treasury.gov/policy-issues/coronavirus/about-the-cares-act

Dridi, S. (2022a). *Supervised Learning—A Systematic Literature Review*. OSF Preprints. https://doi.org/10.31219/osf.io/tysr4

Dridi, S. (2022b). *Unsupervised Learning—A Systematic Literature Review*. OSF Preprints. https://doi.org/10.31219/osf.io/kpqr6

Emilio Ferrara. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, *6*(1), 3–3. https://doi.org/10.3390/sci6010003

Emmons, W. R., & Dahl, D. (2022). *Was the Paycheck Protection Program Effective?* Federal Reserve Bank of St Louis. https://www.stlouisfed.org/publications/regional-economist/2022/jul/was-paycheck-protection-program-effective

Giupponi, G., Landais, C., & Lapeyre, A. (2022). Should We Insure Workers or Jobs During Recessions? *Journal of Economic Perspectives*, *36*(2), 29–54. https://doi.org/10.1257/jep.36.2.29

Gui, Q., Zhou, H., Guo, N., & Niu, B. (2024). A survey of class-imbalanced semi-supervised learning. *Machine Learning*, *113*(8), 5057–5086. https://doi.org/10.1007/s10994-023-06344-7

Halkidi, M. (2009). Hierarchial Clustering. In L. LIU & M. T. ÖZSU (Eds.), *Encyclopedia of Database Systems* (pp. 1291–1294). Springer US. https://doi.org/10.1007/978-0-387-39940-9\_604

Harrington, B., & Leslie, C. A. (2023). Toward a Multilevel Sociology of Fraud. *Northwestern University Law Review*, *118*(1), 139–166.

Humphries, J. E., Neilson, C. A., & Ulyssea, G. (2020). Information frictions and access to the Paycheck Protection Program. *Journal of Public Economics*, *190*, 104244. https://doi.org/10.1016/j.jpubeco.2020.104244

Itri, B., Mohamed, Y., Mohammed, Q., & Omar, B. (2019). Performance comparative study of machine learning algorithms for automobile insurance fraud detection. *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, 1–4. https://doi.org/10.1109/ICDS47004.2019.8942277

Kadens, E. (2023). The Persistent Limits of Fraud Prevention in Historical Perspective. *Northwestern University Law Review*, *118*(1), 167–192.

King, D. L., McLanahan, M. L., & Case, C. J. (2023). Frauds Do Not Pause for a Pandemic. *Journal of Business & Educational Leadership*, *13*(1), 67–81.

Koreff, J., Baudot, L., & Sutton, S. G. (2023). Exploring the Impact of Technology Dominance on Audit Professionalism through Data Analytic-Driven Healthcare Audits. *Journal of Information Systems*, *37*(3), 59–80. https://doi.org/10.2308/ISYS-2022-023

Larson, Benjamin James. (2020). *False Positive Reduction in Credit Card Fraud Prediction: An Evaluation of Machine Learning Methodology on Imbalanced Data*. https://www.proquest.com/openview/197642c9490f064f94d434ff0ed8f9e7/1?pq-origsite=gscholar&cbl=18750&diss=y

Li, L., & Strahan, P. (2020). *Who Supplies PPP Loans (And Does it Matter)? Banks, Relationships and the COVID Crisis* (w28286; p. w28286). National Bureau of Economic Research. https://doi.org/10.3386/w28286

López, M. I., Luna, J. M., Romero, C., & Ventura, S. (2012). Classification via clustering for predicting final marks based on student participation in forums. *Proceedings of the 5th International Conference on Educational Data Mining*.

Ma, K. W. F., & McKinnon, T. (2020). *COVID-19 and Cyber Fraud: Emerging Threats During the Pandemic* (SSRN Scholarly Paper 3718845). https://doi.org/10.2139/ssrn.3718845

Maalouf, M. (2011). Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies*, *3*, 281–299. https://doi.org/10.1504/IJDATS.2011.041335

Miller, K. S., & Bertozzi, A. L. (2024). Model Change Active Learning in Graph-Based Semi-supervised Learning. *Communications on Applied Mathematics and Computation*, *6*(2), 1270–1298. https://doi.org/10.1007/s42967-023-00328-z

Minastireanu, E., & Mesnita, G. (2019). *An Analysis of the Most Used Machine Learning Algorithms for Online Fraud Detection*. *Informatica Economica, 23*, 5–16. https://doi.org/10.12948/issn14531305/23.1.2019.01

Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. (2021). Machine Learning for Anomaly Detection: A Systematic Review. *IEEE Access*, *9*, 78658–78700. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3083060

PRAC. (2020, April 15). *Reporting on Oversight | Pandemic Oversight*. https://www.pandemicoversight.gov/oversight

Rixom, B. A., Rixom, J. M., Pippin, S., & Wong, J. (2021). Contrasting Public Perceptions of Government versus Certified Public Accounting Firm Oversight of Relief Packages. *Accounting & the Public Interest*, *21*(1), 39–63. https://doi.org/10.2308/API-2020-019

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLoS ONE*, *14*(1), e0210236. https://doi.org/10.1371/journal.pone.0210236

Sabasteanski, N., Brooks, J., & Chandler, T. (2021). Saving lives and livelihoods: The Paycheck Protection Program and its efficacy. *EconomiA*, *22*(3), 278–290. https://doi.org/10.1016/j.econ.2021.11.004

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Stoner, J. L., Felix, R., & Stadler Blank, A. (2023). Best practices for implementing experimental research methods. *International Journal of Consumer Studies*, *47*(4), 1579–1595. https://doi.org/10.1111/ijcs.12878

The National Law Review. (2023). *The DOJ is Continuing to Target PPP Recipients for Fraud*. The National Law Review. https://www.natlawreview.com/article/doj-continuing-to-target-ppp-recipients-fraud

USGAO. (2020). *Telecommunications: FCC Should Take Action to Better Manage Persistent Fraud Risks in the Schools and Libraries Program | U.S. GAO*. https://www.gao.gov/products/gao-20-606

USSBA. (2021). *Process Flow Diagram*. US SBA PPP Loan API. https://ussbappp.github.io/usecase/usecase-process-flow.html

USSBA. (2023). *PPP FOIA - U.S. Small Business Administration (SBA) | Open Data*. https://data.sba.gov/dataset/ppp-foia

USSBA OIG. (2023). *COVID-19 Pandemic EIDL and PPP Loan Fraud Landscape* (23–09). USSBA. https://www.sba.gov/document/report-23-09-covid-19-pandemic-eidl-ppp-loan-fraud-landscape

West, D. M. (2021). *Using AI and machine learning to reduce government fraud*. https://policycommons.net/artifacts/4142718/using-ai-and-machine-learning-to-reduce-government-fraud/4952011/

Xu, J. J., Chen, D., Chau, M., Li, L., & Zheng, H. (2022). Peer-to-peer loan fraud detection: Constructing features from transaction data. *MIS Quarterly*, *46*(3).

Zhang, X., Shen, X., & Ouyang, T. (2022). Extension of DBSCAN in Online Clustering: An Approach Based on Three-Layer Granular Models. *Applied Sciences*, *12*(19), Article 19. https://doi.org/10.3390/app12199402

Zhao, G., Li, G., Qin, Y., Zhang, J., Chai, Z., Wei, X., Lin, L., & Yu, Y. (2024). Exploration and Exploitation of Unlabeled Data for Open-Set Semi-supervised Learning. *International Journal of Computer Vision*. https://doi.org/10.1007/s11263-024-02155-y

Zhou, N., Zhang, Z., Nair, V. N., Singhal, H., & Chen, J. (2022). Bias, Fairness and Accountability with Artificial Intelligence and Machine Learning Algorithms. *International Statistical Review*, *90*(3), 468–480. https://doi.org/10.1111/insr.12492

# 

# Appendix A XXX

Insert Appendix A content here…

Note that you must include your Programming Modules and/or the final standalone (if applicable here). An alternative that I would propose is to include the link to your personal GitHub webpage and all modules and the data you worked with there.

You should include a static link to ensure the **GitHub page stays live.**

# Appendix B XXX

Insert/type Appendix n content here…

# Appendix C XXX

Insert/type Appendix n content here…