**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 Figures**

[**Figure 1** *Classification Through Clustering Framework* 21](#_Toc184834421)

[**Figure 2** *COVID-19 Deaths, by Week, in The United States* 29](#_Toc184834422)

[**Figure 3** *Labor Market Policy Responses to Recessions the United States and Europe* 32](#_Toc184834423)

[**Figure 4** *PPP Loan Application Process Flowchart* 35](#_Toc184834424)

[**Figure 5** *Logistic Regression Response Function* 37](#_Toc184834425)

[**Figure 6** *Graphic Representation of a Random Forest Decision Tree* 38](#_Toc184834426)

[**Figure 7** *SVM Hyperplane* 39](#_Toc184834427)

[**Figure 8** *Illustration of K-Means Clustering Method* 40](#_Toc184834428)

[**Figure 9** *Graphic Depiction of DBSCAN Clustering* 41](#_Toc184834429)

[**Figure 10** *Hierarchical Clustering Dendrogram* 42](#_Toc184834430)

[**Figure 11** *Study Processflow Diagram* 54](#_Toc184834431)

# 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 and 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 and 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 newarly 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 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

The K-Means, Hierarchical, T-SNE, and DBSCAN models used in conjunction with SVM, logistic regression, neural networks, Naïve Bayesian, tree models, and ensembles perform identically in detecting fraudulent activity: 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

This chapter details the research methodology used to investigate fraudulent loan applications within the Paycheck Protection Program (PPP) dataset by outlining the approach to data preprocessing, feature selection, and model implementation necessary to detect fraudulent patterns effectively. The methodology is aligned with the study’s two primary research objectives: identifying key fraud indicators and evaluating effective model combinations for detecting anomalies. The chapter is organized to include a restatement of the problem and purpose statements, an overview of the research design, a description of the population and sampling frame, the operational definitions of variables, study procedures, data analysis methods, and ethical assurances. Each section builds on the methodological structure to ensure that the research design aligns directly with the study’s research questions and hypotheses.

**Problem Statement**

The PPP, implemented as part of emergency relief measures during the COVID-19 pandemic, provided essential financial support to small businesses. However, due to the program’s rapid rollout, the PPP has been vulnerable to fraudulent applications that exploit its high-volume, high-value structure. Identifying fraudulent activity within this dataset presents unique challenges, as traditional fraud detection methods are not always effective for large-scale, government-administered programs. This study addresses the need for a robust fraud detection methodology capable of analyzing the PPP dataset’s scale, complexity, and diversity to identify fraud patterns that may otherwise go undetected.

**Purpose of the Study**

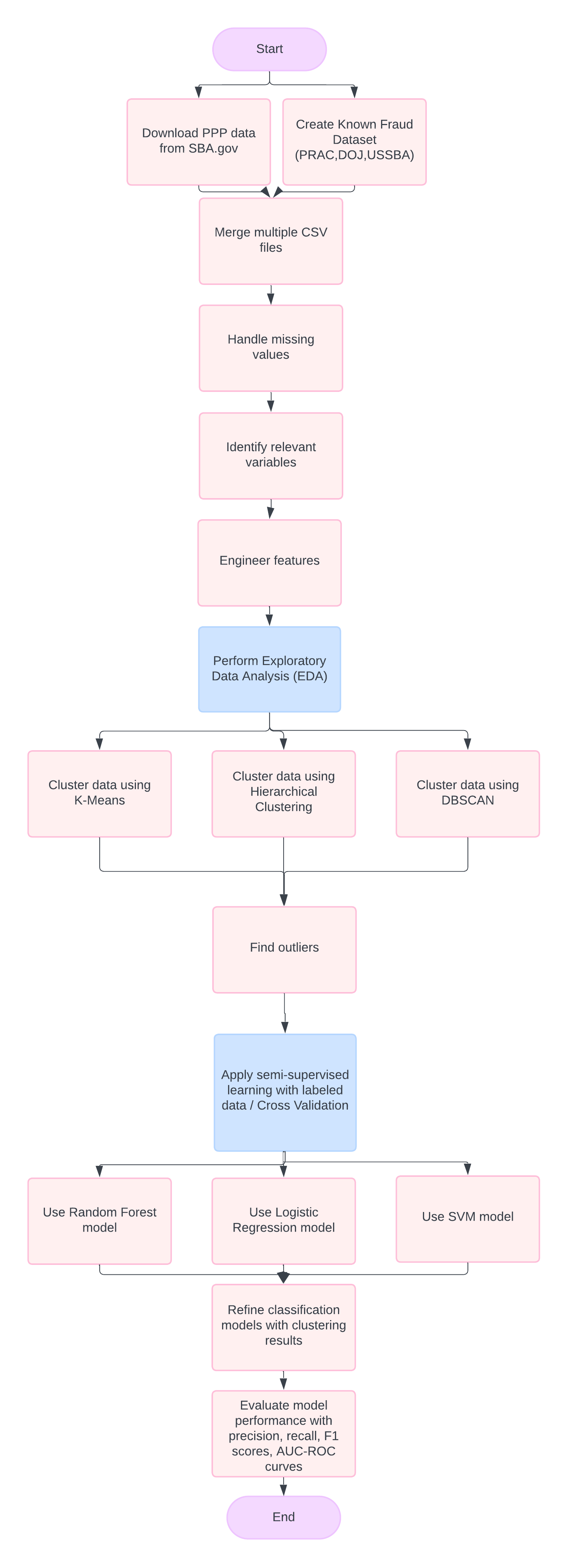
The purpose of this study is to identify patterns indicative of fraud within the PPP dataset by applying machine learning techniques. Specifically, the study investigates both unsupervised and semi-supervised learning models to identify key fraud indicators and to test the efficacy of combining clustering methods with supervised models for fraud detection. By analyzing the effectiveness of various model pairings and exploring feature selection’s impact, this study aims to uncover optimal machine learning configurations for detecting fraud within large, complex datasets such as the PPP.

## Research Methodology and Design Process Diagram

The flowchart below (Figure 11) outlines each stage in this process, beginning with data collection from public sources and government reports on confirmed fraud cases. Each step will be described in detail in this chapter.

Key steps include data preprocessing to handle missing values and anonymize sensitive information, followed by dimensionality reduction using PCA to focus on the most relevant features. Unsupervised clustering methods are then used to identify patterns and potential fraud clusters within the data. These clusters are further refined to detect outliers, which often represent anomalies linked to fraud. Finally, the clustered data is combined with labeled instances to train semi-supervised classification models, including Random Forest, Logistic Regression, and SVM, which are evaluated based on precision, recall, F1 scores, and AUC-ROC.

**Figure 11***Study Processflow Diagram*



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

This study employs a Classification through Clustering framework that combines unsupervised clustering and semi-supervised classification techniques. The design follows a systematic, hypothesis-driven structure, ensuring that each phase directly addresses the research questions and hypotheses:

* **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**: The K-Means, Hierarchical, T-SNE, and DBSCAN models used in conjunction with SVM, logistic regression, neural networks, Naïve Bayesian, tree models, and ensembles perform identically in detecting fraudulent activity: 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 methodology is suitable for addressing the research questions and hypotheses due to its structured, iterative design that leverages both feature analysis and model comparisons. 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

While the Classification through Clustering approach is well-suited for this study, alternative methodologies were considered but found to be less appropriate based on the specific challenges and goals of the research. Below are three alternative approaches, explaining why they were ultimately not selected.

**1. Supervised Learning Only**

In a traditional supervised learning approach, a model is trained entirely on labeled data to classify new examples. While this method is effective when there is a large, balanced labeled dataset available, it is less suitable in the context of PPP loan fraud detection due to the scarcity of labeled fraud cases.

* **Challenges with Supervised Learning**:
  + **Limited labeled data**: In the case of the PPP, only a small subset of loans is definitively labeled as fraudulent, as determined by prosecuted cases and investigative reports from sources like the DOJ and PandemicOversight.gov. Supervised learning requires extensive labeled data, and relying on this limited dataset would result in a model that overfits to the known fraud cases and fails to generalize well to new, unlabeled data (Ali et al., 2021).
  + **Imbalance in class distribution**: Fraud cases represent a small fraction of the overall dataset, making it difficult for supervised models to detect fraud without overemphasizing the non-fraudulent loans. This class imbalance leads to biased models that often predict loans as non-fraudulent to achieve higher accuracy scores, at the expense of missing potential fraud (López et al., 2012).
  + **Overfitting**: With a small number of labeled fraud cases, a supervised model may overly fit the patterns in the training data, performing well on these cases but poorly on new data, resulting in unreliable fraud detection.

Due to the scarcity of labeled fraud cases and the risk of overfitting, a purely supervised learning approach was deemed less appropriate for this study. Semi-supervised learning, which incorporates both labeled and unlabeled data, allows the model to learn from patterns in the larger, unlabeled dataset while benefiting from the fraud labels available.

**2. Unsupervised Learning Only**

An unsupervised learning approach, in which no labeled data is used, was also considered. In this method, clustering algorithms such as K-Means or DBSCAN would group loans based on their similarities without predefined fraud labels. Unsupervised learning excels at anomaly detection, making it attractive for identifying unusual loan patterns that could be indicative of fraud.

* **Challenges with Unsupervised Learning**:
  + **Lack of refinement**: While unsupervised clustering can detect potential anomalies or outliers, it lacks the refinement provided by real-world fraud examples. Without the incorporation of labeled fraud cases, the model may misclassify benign anomalies as fraud or fail to catch sophisticated fraud schemes that don’t exhibit obvious anomalies.
  + **Higher false-positive rates**: Clustering algorithms may identify loans that deviate from normal patterns as suspicious, but not all deviations are fraudulent. This can lead to a high number of false positives, overwhelming investigators and undermining the model’s credibility (Sarker, 2021). Fraud detection requires more than anomaly detection—it also requires context and understanding of known fraudulent behaviors, which unsupervised learning alone cannot provide.
  + **Difficult to evaluate performance**: Without labeled data, it is challenging to quantitatively evaluate the model’s performance. Metrics like precision, recall, and accuracy depend on the presence of labeled data to verify predictions.

Although unsupervised learning is useful for detecting patterns, it is insufficient on its own for fraud detection. Incorporating labeled fraud cases via semi-supervised learning enhances the model’s ability to detect both anomalies and fraudulent behavior with greater precision.

**3. Manual Fraud Detection**

An entirely manual approach to detecting fraud was also considered, in which human investigators would review the PPP loan data and flag suspicious cases. While manual review remains an important component of fraud detection in many fields, it is not feasible given the size of the PPP dataset, which contains millions of records.

* **Challenges with Manual Review**:
  + **Scalability**: The PPP loan dataset is massive, containing over 11 million loans, making manual review inefficient and time-consuming. Human investigators could only review a small fraction of the dataset, leaving most loans unchecked.
  + **Subjectivity**: Manual detection can introduce subjectivity and inconsistency. Different reviewers may flag loans based on different criteria, and cognitive biases may lead to overlooking certain types of fraud (Giupponi et al., 2022).
  + **Cost and resource-intensive**: Investigating each loan manually would require extensive time and financial resources. Given the scale of the PPP, this approach would not be feasible for real-time or large-scale fraud detection efforts.

For these reasons, a manual review is impractical for a dataset of this size. Machine learning offers a scalable and automated solution for detecting fraud while preserving human oversight for reviewing flagged cases.

### Appropriateness of the Methodology

The Classification through Clustering framework, coupled with semi-supervised learning, is highly appropriate for detecting fraud within the PPP loan dataset. Several key factors make this methodology well-suited to the study’s goals:

**1. Handling Limited Labeled Data**

A core challenge in detecting fraud in the PPP dataset is the limited availability of labeled fraud cases. Only a small fraction of loans have been confirmed as fraudulent through investigations by the DOJ, SBA OIG, and PandemicOversight.gov. A purely supervised learning approach would require a much larger labeled dataset to function effectively.

* **Semi-supervised learning**, as used in this study, overcomes this challenge by leveraging both labeled and unlabeled data. The labeled fraud cases serve as a foundation for refining the clusters generated during unsupervised clustering, allowing the model to detect patterns of fraudulent behavior without requiring a large labeled dataset (Ali et al., 2021).
* This method enables the discovery of fraud patterns that may not be immediately visible in the limited labeled dataset, improving the model’s ability to generalize and detect previously unseen cases of fraud.

**2. Fraudulent Behavior as Anomalies**

Fraud detection often involves identifying behaviors that deviate from normal patterns, which are known as anomalies. The unsupervised clustering component of this methodology is highly effective at detecting such deviations.

* Clustering algorithms like K-Means and DBSCAN allow for the identification of loan applications that behave unusually compared to the rest of the dataset (López et al., 2012). For example, loans with unusually high loan-to-employee ratios or businesses applying for loans in unexpected geographic regions can be flagged for further investigation.
* By incorporating real-world labeled fraud data in the semi-supervised phase, the model can further refine these clusters and distinguish between legitimate anomalies and actual fraud, reducing the number of false positives.

**3. Flexibility and Scalability**

Given the size and complexity of the PPP loan dataset, the chosen methodology must be flexible and scalable to handle large volumes of data. The Classification through Clustering framework is highly scalable, as clustering techniques can efficiently process large datasets, and the integration of labeled fraud cases enables the model to focus on the most relevant patterns.

* PCA helps reduce the dimensionality of the dataset, ensuring that the clustering algorithms operate efficiently, even with a high number of variables (Sarker, 2021).
* The use of ensemble learning methods like Random Forest ensures that the classification models remain robust across various data configurations, further enhancing the model’s scalability and applicability to real-world datasets.

**4. Mitigating Class Imbalance**

Fraud cases typically represent a small proportion of the total dataset, leading to issues of class imbalance. This imbalance can cause traditional models to underperform in detecting fraud, as they are overwhelmed by the majority class (non-fraudulent loans).

* This methodology addresses class imbalance by applying techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to create synthetic examples of fraud, ensuring that the model receives sufficient training data for both classes (Sarker, 2021).
* Additionally, stratified cross-validation will be used to evaluate the model’s performance, ensuring that each data fold contains a representative number of fraud cases.

In conclusion, the Classification through Clustering approach is not only appropriate but highly effective for detecting fraud in the PPP dataset. By integrating unsupervised clustering with labeled fraud cases and applying semi-supervised learning, the model is able to scale, generalize, and refine its detection capabilities across a large, imbalanced, and complex dataset. This methodology offers the flexibility needed to uncover hidden fraud patterns while maintaining high accuracy and reducing false positives, making it ideal for the study’s objectives.

### Comparative Analysis and Risk Mitigation

While supervised and unsupervised methods are commonly used in fraud detection, they present limitations in the context of PPP fraud:

* **Supervised Models**: Depend heavily on labeled data, which is scarce in PPP datasets. This limits their ability to adapt to novel fraud patterns.
* **Unsupervised Models**: Lack refinement and often produce high false-positive rates without labeled data for validation.

By combining clustering with semi-supervised learning, this study leverages the strengths of both approaches. The semi-supervised phase ensures that the labeled fraud cases refine the clusters, enhancing accuracy while minimizing false positives. This hybrid approach is especially effective in contexts like PPP fraud detection, where labeled data is limited and anomalies are diverse.

However, the methodology acknowledges several risks:

1. **Clustering Sensitivity**: Algorithms like K-means are sensitive to initialization and the number of clusters, potentially leading to unstable results. To mitigate this, multiple initializations and validation metrics (e.g., silhouette score and Calinski-Harabasz index) are employed to ensure stability and robustness.
2. **PCA Limitations**: PCA may obscure interpretable features, which can affect downstream classification accuracy. This risk is mitigated by monitoring the explained variance ratio and carefully analyzing feature contributions during PCA implementation.
3. **Imbalanced Data Challenges**: Despite the use of oversampling techniques like SMOTE, class imbalance remains a potential limitation. To address this, stratified cross-validation ensures that training and testing data maintain proportional fraud representation, improving generalizability.

By addressing these risks proactively and balancing the trade-offs of supervised and unsupervised methods, the Classification through Clustering framework demonstrates strong potential for robust fraud detection in the PPP dataset.

## Population and Sample

The population for this study consists of all businesses and entities that applied for and received loans exceeding **$150,000** through the **PPP,** made publicly available by the **SBA**. This subset of PPP loans provides critical insights for detecting potential fraudulent activities, as larger loans are often associated with higher scrutiny and risk.

The decision to focus on loans above $150,000 is based on the greater financial impact and the availability of detailed loan-level data. By analyzing the full population of these loans, the study aims to uncover patterns of fraudulent behavior that may be more prevalent among high-value loans.

### ****Key Variables and Characteristics****

Based on the **PPP Data Dictionary**, the dataset includes several important fields that are critical for the study’s fraud detection objectives:

* **LoanAmount**: This field represents the total loan provided to each borrower. This study specifically focuses on loans over **$150,000**, as these larger loans are more likely to attract both scrutiny and potential fraud.
* **BorrowerName**: Contains the name of the business or entity that applied for the loan. This variable will allow for cross-referencing with publicly available databases on prosecuted fraud cases.
* **BorrowerCity** and **BorrowerState**: Geographic indicators that will enable the identification of fraud patterns across different regions of the U.S. Regional analysis is important, as certain areas may have higher concentrations of fraud due to differing levels of oversight or economic conditions.
* **NAICSCode**: The **North American Industry Classification System (NAICS) Code** identifies the industry sector of the borrower. This variable is crucial for analyzing industry-specific fraud trends, as certain industries may be more prone to misuse of funds or may have experienced higher levels of financial distress during the pandemic, leading to greater opportunities for fraud.
* **JobsReported**: This field indicates the number of jobs supported by the loan, a key variable for detecting inconsistencies. Fraudulent borrowers may inflate their job numbers to qualify for larger loans or to appear compliant with PPP requirements, making this variable central to fraud detection.
* **LoanStatus**: Tracks the current status of the loan, such as whether the loan has been forgiven, repaid, or is still active. Loans that have not been forgiven, particularly for reasons related to improper use of funds, may signal potential fraud.
* **LenderName**: Identifies the financial institution that issued the loan. This variable can be used to detect patterns of fraud across specific lenders, especially those with high concentrations of fraudulent loans.

These variables provide a comprehensive view of each loan, allowing the study to analyze the full range of borrower behavior, loan status, and geographic distribution. By leveraging these data fields, the study will be able to uncover complex patterns of fraud that may not be apparent when examining individual cases.

### ****Rationale for Using the Entire Population****

The study will not use a sample but will instead analyze the **entire population** of loans exceeding $150,000. This approach offers several advantages:

1. **High-Risk Focus**:
   * Loans over $150,000 represent a significant financial risk to the PPP. These loans are often subject to more scrutiny and may have a higher likelihood of fraud. By focusing on the entire population of large loans, the study targets high-value cases where the financial impact of fraud would be more substantial (PRAC, 2020).
2. **Comprehensive Data Coverage**:
   * Analyzing the entire dataset ensures that no potentially fraudulent loans are overlooked. This comprehensive approach allows for the detection of fraud across a wide spectrum of industries, regions, and loan sizes, improving the generalizability and accuracy of the findings.
   * By including all loans over $150,000, the study avoids the biases and limitations that could arise from sampling, providing a full representation of the population.
3. **Diversity of Loan Characteristics**:
   * The population includes businesses of various sizes, industries, and geographic locations. This diversity ensures that the model can generalize fraud detection across different sectors and regions, providing robust and scalable results that can be applied to future fraud detection efforts.
4. **Efficient Processing with Machine Learning**:
   * Advances in machine learning algorithms, particularly **clustering techniques** and **semi-supervised learning**, allow the model to process and analyze the full population of PPP loans efficiently. By using techniques such as **PPP** to reduce the dimensionality of the dataset, the model will be able to handle the large number of variables while maintaining computational efficiency (Sarker, 2021).
   * The entire population dataset will be used to train the model, ensuring that fraud patterns are detected across all loans, without the need for human intervention in the data selection process.

### ****Fraud Detection Focus****

Focusing on loans over $150,000 is especially relevant for detecting fraud within the PPP because larger loans:

* Present Higher Financial Risk: Fraud involving large loans has a more significant financial impact, making these cases a priority for detection efforts.
* Enable Complex Fraud Schemes: Larger loans may be subject to more sophisticated fraud, including inflated payroll numbers, falsified documentation, or misrepresentation of business size, all of which are easier to identify in this high-value subset.

By narrowing the focus to loans over $150,000, this study addresses the most impactful cases of potential fraud while also leveraging the Fraud Triangle to understand the behavioral underpinnings of fraudulent activity. This focus enables a deeper analysis of how financial pressures and opportunities interact within the unique context of high-value PPP loans.

## Instrumentation

This study leverages a combination of **software, hardware**, and specialized **libraries** to carry out data preprocessing, semi-supervised learning, and model evaluation on the **PPP** loan dataset. The choice of instrumentation ensures scalability, efficiency, and accuracy, given the dataset’s size and the need for advanced machine learning techniques to detect fraudulent activity.

### ****Software and Tools****

1. **Programming Language: Python**
   * Python is the primary language for this study due to its versatility, ease of use, and extensive ecosystem of libraries for data science and machine learning. Its rich set of libraries enables efficient data manipulation, statistical analysis, and machine learning, making it ideal for this research.
   * **Advantages of Python**:
     + Large community support.
     + Extensive library of open-source tools.
     + Seamless integration with cloud-based environments like **Google Colab**.
     + Flexibility to integrate both deep learning and traditional machine learning techniques.
2. **Machine Learning Libraries**:
   * **Scikit-learn**: A powerful, widely-used library for machine learning in Python, Scikit-learn provides an array of tools for model evaluation, data preprocessing, and clustering algorithms like **K-Means and DBSCAN**. In this study, Scikit-learn will be crucial for implementing the unsupervised learning techniques used to identify patterns in the PPP dataset, as well as for evaluating the accuracy and efficiency of the semi-supervised learning model.
     + **Key Functions**:
       - **Clustering Algorithms**: Used for unsupervised learning to group similar loans and detect anomalies that may signal fraud.
       - **Cross-Validation**: Scikit-learn’s k-fold cross-validation will ensure the model is evaluated on various data splits, improving generalization and reducing overfitting.
       - **Dimensionality Reduction**: Scikit-learn’s implementation of **PCA** will be used to reduce the dataset’s complexity, making it easier to identify key features while retaining critical patterns.
   * **Keras**: A high-level neural network library running on top of **TensorFlow**, Keras will be used for the semi-supervised learning phase of the study. Keras allows for rapid prototyping of machine learning models and provides a user-friendly interface to build complex models for deep learning.
     + **Advantages of Keras**:
       - Simplified API for building and training deep learning models.
       - Integration with **Google Colab**, allowing for the use of GPU acceleration to speed up the training of large models.
       - Well-suited for handling complex fraud detection tasks due to its ability to build models that learn intricate patterns in data.
   * **XGBoost**: XGBoost will be utilized in the supervised learning phase of the study to boost model performance. Known for its speed and accuracy, XGBoost excels at tasks involving structured datasets like the PPP loan data, and its built-in mechanisms to prevent overfitting make it a good choice for the classification of fraudulent loans.
     + **Key Strengths**:
       - **Gradient Boosting**: XGBoost uses a gradient-boosting framework that iteratively improves the model’s predictive capabilities.
       - **Handling Class Imbalance**: Given the small number of fraudulent loans compared to non-fraudulent ones, XGBoost is effective at managing class imbalances by assigning higher weights to underrepresented classes.
3. **Data Management and Storage**:
   * The dataset will be stored and managed using **CSV files** hosted on **Google Drive**. CSV storage offers simplicity and ease of use, allowing seamless integration with Python’s **Pandas** library for data manipulation and preprocessing.
     + **Advantages of CSV Storage**:
       - Universally supported format for tabular data.
       - Easily accessible for batch processing and manipulation.
       - Provides flexibility in scaling the dataset as required for additional analyses.

### ****Hardware Resources****

1. **Google Colab Pro**:
   * **Google Colab Pro** will be the primary environment for running the analysis. Colab provides an easy-to-use platform for Python programming, combined with the ability to scale computational resources as needed. With access to **upgraded GPU resources,** Colab Pro ensures that the model training processes, particularly those involving deep learning models built with Keras, can be executed efficiently.
   * **Advantages of Google Colab Pro**:
     + **GPU Acceleration**: Enables faster model training by offloading computationally intensive tasks to powerful GPU hardware, which is essential for handling large datasets like the PPP loan data.
     + **Cloud-Based Infrastructure**: Allows for scalable computation without the need for local hardware, enabling continuous work without constraints on local processing power.
2. **GPU Accelerators**:
   * **GPU accelerators** will be used to enhance the performance of deep learning models, particularly during the training of the semi-supervised learning models. GPU resources will reduce the time needed to train complex models, improving overall workflow efficiency.

### ****Data Cleaning and Preprocessing****

To ensure the quality and usability of the dataset, various libraries will be employed to clean and preprocess the data:

1. **Pandas**:
   * **Pandas** will serve as the primary tool for managing the dataset. With its powerful DataFrame structure, Pandas allows for efficient handling of missing values, data filtering, and dataset transformations. Given the size of the PPP dataset, Pandas' ability to process large amounts of data in a flexible and readable manner is critical.
     + **Key Functions**:
       - **Handling Missing Values**: Pandas provides built-in functions to fill missing data, either through imputation or interpolation, ensuring a clean dataset for analysis.
       - **Filtering and Transforming Data**: Enables the identification and removal of outliers, transforming data as needed for model input.
2. **NumPy**:
   * **NumPy** will complement Pandas by providing efficient numerical operations. NumPy is especially useful for matrix manipulations and will assist in mathematical computations required during preprocessing and normalization stages.

### ****Modeling and Evaluation****

1. **Modeling Tools**:
   * **Scikit-learn** will be used for **dimensionality reduction, model evaluation**, and cross-validation techniques. These functions are essential to ensure that the model generalizes well to unseen data and is not overfitting.
     + **Principal Component Analysis (PCA)**: This technique will reduce the number of input features, simplifying the data while retaining the most important information.
     + **Cross-Validation**: K-fold cross-validation will ensure robust model evaluation, splitting the data into training and test sets to assess model performance across different samples.
2. **XGBoost**:
   * In the semi-supervised learning phase**, XGBoost** will be used for model evaluation and classification tasks. Its strong predictive performance and ability to handle imbalanced datasets make it ideal for the task of detecting fraudulent loans.

### ****Data Visualization****

1. **Matplotlib**:
   * **Matplotlib** will be used to create visual representations of the data and results. The library will be essential for plotting confusion matrices, visualizing model performance metrics, and presenting key findings of the fraud detection analysis.

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

1. **Primary Dataset**:
   * The dataset for this study is obtained from the **SBA**. It includes details of all loans exceeding $150,000 that were distributed under the PPP. The dataset is publicly available and will be downloaded in **CSV format** from the SBA's website, providing key information such as loan amount, business type, geographic location, and loan forgiveness status.
2. **Supplementary Fraud Data**:
   * To augment the dataset, information on fraud cases will be manually scraped from sources such as:
     + **PandemicOversight.gov**: This platform consolidates reports on fraud, waste, and abuse related to pandemic relief programs, including the PPP.
     + **SBA Office of Inspector General (OIG)**: Reports and audits from the OIG will help identify loans that were flagged or under investigation.
     + **Department of Justice (DOJ)**: Records of prosecuted fraud cases will be collected to build the labeled fraud dataset, which will be critical for the semi-supervised learning phase

## Operational Definitions of Variables

In this study, the dataset from the PPP provided by the SBA will be used to detect fraudulent loans. The variables for the study are based on the loan data and additional labels sourced from prosecuted fraud cases. Each variable is defined based on its role in the semi-supervised learning model and its relevance to fraud detection.

### ****Independent Variables (Predictors)****

1. **LoanAmount**:
   * **Data Type**: Continuous
   * **Level of Measurement**: Ratio
   * **Description**: This variable represents the total loan amount received by the borrower under the PPP. The larger the loan, the higher the financial risk, and thus, it is a key predictor for potential fraud.
   * **Instrument**: Derived directly from the SBA dataset, with additional validation from publicly available databases such as **PandemicOversight.gov** and **DOJ** for fraudulent loans.
   * **Feature Engineering**: A **Loan-to-Employee Ratio** variable will be derived from this variable and the number of employees to help detect anomalies where loans may be disproportionately large relative to the business size.
2. **JobsReported**:
   * **Data Type**: Continuous
   * **Level of Measurement**: Ratio
   * **Description**: Indicates the number of employees the loan was intended to support. Significant discrepancies between this value and the size of the loan may indicate fraudulent behavior.
   * **Instrument**: Collected directly from the SBA dataset.
   * **Feature Engineering**: A feature will be engineered by normalizing the **JobsReported** value in conjunction with the business type and geographic region, which helps identify outliers.
3. **BorrowerState**:
   * **Data Type**: Categorical (Nominal)
   * **Level of Measurement**: Nominal
   * **Description**: The U.S. state where the borrower’s business is located. Geographic location can influence the likelihood of fraud due to varying levels of economic conditions and oversight.
   * **Instrument**: Directly sourced from the SBA loan dataset.
   * **Feature Engineering**: Location-based features will be created to assess whether certain states show higher instances of fraud. These features will incorporate regional fraud data obtained from sources such as **PandemicOversight.gov**.
4. **NAICSCode**:
   * **Data Type**: Categorical (Nominal)
   * **Level of Measurement**: Nominal
   * **Description**: The industry sector of the borrower, as classified by the North American Industry Classification System (NAICS). Certain industries may have higher susceptibility to fraud based on past trends.
   * **Instrument**: Collected from the SBA loan dataset.
   * **Feature Engineering**: Industry-specific fraud patterns will be identified through clustering and supervised learning models. **Dummy variables** will be created for each industry group.
5. **LoanStatus**:
   * **Data Type**: Categorical (Nominal)
   * **Level of Measurement**: Nominal
   * **Description**: Describes the current status of the loan, such as forgiven, repaid, or active. Loans that were not forgiven due to improper usage may indicate fraudulent activity.
   * **Instrument**: Directly sourced from the SBA dataset.
   * **Feature Engineering**: Fraud detection algorithms will leverage the **LoanStatus** field as an indicator of misuse or anomalies in repayment and forgiveness trends.

### ****Dependent Variable (Criterion)****

1. **FraudLabel**:
   * **Data Type**: Binary (Fraud/No Fraud)
   * **Level of Measurement**: Nominal
   * **Description**: This is a binary variable indicating whether a loan is flagged as fraudulent. Fraud cases are identified using external data sources, including **DOJ**, **PandemicOversight.gov**, and **SBA OIG**.
   * **Instrument**: Labeled fraud cases from publicly available sources. Fraudulent loans are identified through manual data collection from **PandemicOversight.gov**, **SBA OIG**, and **DOJ** records of prosecuted fraud cases.
   * **Feature Engineering**: No further transformations; the variable is used directly in supervised learning.

### ****Artificial Variables/Feature Engineering****

To enhance the model’s ability to detect fraud, several new variables will be created through **feature engineering** techniques. These include:

1. **Loan-to-Employee Ratio**:
   * **Data Type**: Continuous
   * **Level of Measurement**: Ratio
   * **Description**: This derived feature represents the loan amount relative to the number of employees reported. A high ratio might indicate fraud, particularly if a small business received a disproportionately large loan. This variable will be key in identifying businesses that may have misrepresented their payroll to obtain higher loans.
2. **LoanApprovalDate**:
   * **Data Type**: Date/Time
   * **Level of Measurement**: Interval
   * **Description**: The date on which the loan was approved. Time-based features will be engineered, such as loan approval trends over time and detection of irregularities in loan spikes during specific periods, particularly when program guidelines were updated.
3. **Location-Based Fraud Risk Score**:
   * **Data Type**: Continuous
   * **Level of Measurement**: Interval
   * **Description**: This feature assigns a fraud risk score based on the business’s location. Using geographic clustering and prior fraud data, regions with higher instances of fraud will be given a higher risk score.

### ****Distribution Determination and Bias****

The variables in this study are expected to follow different distributions based on their types (continuous or categorical). A **distributional analysis** will be conducted to understand whether each variable is normally distributed or follows other patterns, such as skewness in loan amounts or regional clusters of fraud.

* **Handling Imbalanced Classes**: Given that fraud cases are expected to represent a small portion of the dataset, the **FraudLabel** variable will exhibit significant class imbalance. To address this, techniques such as **SMOTE** will be applied to generate synthetic fraud cases, balancing the dataset and improving model performance.
* **Selection Bias**: The dataset is limited to loans over $150,000, meaning smaller loans are excluded. While this focuses the study on high-value loans with higher fraud risk, it may limit generalizability to smaller loans.
* **Data Availability Bias**: Fraud labels are based on publicly available information, which may not cover all fraudulent cases due to reporting or prosecutorial delays.

## Study Procedures

### Data Collection and Preprocessing

The first stage of this study involves gathering and preparing data from two primary sources: the publicly available PPP loan dataset and a labeled fraud dataset created from public records of prosecuted fraud cases. Each dataset serves a distinct purpose, allowing for the identification of potential fraud through clustering techniques and enabling semi-supervised learning by combining labeled and unlabeled data.

**PPP Loan Data**  
The primary dataset for this study is the PPP loan dataset, which is publicly accessible on SBA.gov and provides a comprehensive view of the approved PPP loans. This dataset includes features critical to fraud detection, such as loan amounts, business types, reported employee counts, loan forgiveness status, geographic information, and loan approval dates (USSBA, 2023). These features will be analyzed through unsupervised clustering to identify patterns within the general loan data, which will subsequently support the semi-supervised learning phase by highlighting unusual data clusters that may indicate fraudulent activity.

**Labeled Fraud Dataset**  
A crucial element of this study is the creation of a labeled fraud dataset, derived from multiple authoritative sources, to provide real-world examples of fraudulent loans. This labeled dataset will be used to train semi-supervised models that can then classify potential fraud within the PPP loan data.  
The labeled fraud dataset is populated using the following sources:

**Department of Justice (DOJ)**: DOJ press releases and prosecution reports provide verified examples of prosecuted loans identified as fraudulent. Each case includes detailed descriptions of fraud schemes, allowing for comprehensive analysis (Department of Justice, 2020).

**Small Business Administration Office of Inspector General (SBA OIG)**: SBA OIG investigative audits and reports expand the labeled dataset by identifying loans flagged as suspicious or ineligible based on investigative criteria (USSBA OIG, 2023).

**PandemicOversight.gov**: The Pandemic Response Accountability Committee (PRAC) offers a consolidated platform through PandemicOversight.gov, detailing cases of fraud, waste, and abuse related to COVID-19 relief programs. PRAC’s data includes reports from various oversight agencies and provides insights into complex fraud schemes, misuse of funds, and organized fraud efforts within the PPP program (PRAC, 2020).

**Data Preprocessing**

The preprocessing phase involves systematic transformations across both datasets to ensure compatibility with machine learning algorithms and data integrity, particularly for clustering and classification.

1. **Data Integration and Labeling**: The labeled fraud dataset is merged with the general SBA PPP loan data to create a combined dataset with two classifications:
   * **Known Fraud Cases**: Records in the manually labeled fraud dataset are marked with a fraud indicator (fraud = 1).
   * **General PPP Loan Data**: The remaining records are treated as unlabeled (fraud = 0), as they have not been verified as fraudulent. These records will be analyzed in the unsupervised clustering phase to identify potential anomalies that may indicate fraud.
2. **Handling PII through Hashing**: To ensure confidentiality, fields containing personally identifiable information (PII), such as BorrowerName, BorrowerAddress, FranchiseName, ServicingLenderName, and OriginatingLender, are hashed. This process protects sensitive information while maintaining unique identifiers essential for analysis.
3. **Encoding Categorical Variables**: Categorical variables are encoded based on their structure:
   * **Binary Variables**: Indicators such as RuralUrbanIndicator and HubzoneIndicator are one-hot encoded, allowing for clear binary classification.
   * **Multi-Class Variables**: Multi-class categorical variables (e.g., BorrowerState, BusinessType, NAICSCode) are label-encoded to allow compatibility with distance-based clustering and classification methods.
4. **Scaling Numeric Features**: Numeric features, including financial variables, employee counts, and loan allocation details, are standardized using the StandardScaler. Standardization ensures consistent scaling across features, facilitating accurate clustering and improving model performance.
5. **Imputation of Missing Values**: Missing values are not imputed by default, as doing so may inadvertently introduce bias or obscure patterns indicative of fraudulent behavior. Instead, categorical variables with missing values, such as NAICSCode, are assigned a separate “Missing” category to retain the integrity of the original dataset. For numeric fields, missingness is handled on a case-by-case basis: variables with excessive missingness may be excluded, while others may be flagged using binary indicators. This approach ensures that potential signals relevant to fraud are preserved, supporting the study’s objective of detecting anomalous or incomplete records through semi-supervised learning techniques.
6. **Feature Engineering**: Additional features are created to capture relevant fraud indicators. For example, the ForgivenessAmountRatio is calculated by dividing ForgivenessAmount by CurrentApprovalAmount, providing insights into forgiveness patterns that may reveal irregularities suggestive of fraud.

**Feature Selection**

This study evaluates model performance using both a complete feature set and a selected subset of key features, addressing RQ1 and testing H1 to determine the impact of feature selection on clustering and classification accuracy.

1. **Full Feature Set vs. Key Feature Subset**: Two feature configurations are applied to assess the impact of dimensionality on model effectiveness:
   * **Full Feature Set**: The entire preprocessed dataset, used as a baseline to understand the contribution of all available features.
   * **Key Feature Subset**: A subset of features, selected through feature importance analysis, expected to contribute most significantly to fraud detection.
2. **Feature Importance Analysis**:
   * **Correlation Analysis**: Identifies highly correlated features, reducing redundancy and improving interpretability.
   * **Mutual Information and Feature Importance Metrics**: Algorithms such as random forest feature importance and mutual information scores are applied to highlight features with the strongest contributions to identifying fraudulent patterns.
3. **Evaluation of Feature Impact**: The effectiveness of the full and key feature subsets will be evaluated across clustering and classification stages to determine which configuration optimally aids in fraud detection. This testing phase addresses RQ1 and validates H1 by examining the role of specific features in distinguishing fraudulent cases.

### Clustering Methodology

The clustering methodology employs a range of unsupervised algorithms to detect patterns and potential fraud within the general PPP loan dataset, which consists of loans not explicitly labeled as fraudulent. By identifying clusters and outliers, this phase aims to detect anomalies that may indicate fraud. Results from clustering are later combined with labeled fraud data to enhance the accuracy of fraud detection in the semi-supervised learning phase. This section details the four clustering techniques applied in this study: K-Means, Hierarchical Clustering, DBSCAN, and T-SNE for visualization.

PCA is applied prior to clustering to reduce dimensionality and mitigate the curse of dimensionality. By retaining components that explain 95% of the variance in the dataset, PCA simplifies data representation, enhancing clustering performance while preserving critical information.

**K-Means Clustering**

K-Means clustering is applied as a foundational method due to its efficiency and effectiveness in grouping data into distinct clusters. The algorithm iteratively assigns data points to a fixed number of clusters (K) based on their proximity to calculated cluster centroids.

**Procedure**: K-Means is applied to both the full feature set and the key feature subset to evaluate the impact of feature selection. The Elbow Method is used to determine the optimal number of clusters by identifying the point at which adding more clusters yields minimal reduction in within-cluster variance.

**Evaluation**: Cluster quality is assessed using metrics like the Silhouette Score and the Davies-Bouldin Index, which measure cohesion within clusters and separation between clusters. These metrics ensure that clusters are well-defined, facilitating accurate anomaly detection.

**Hierarchical Clustering**

Hierarchical clustering is used to explore nested structures within the data, which may reveal hierarchical or layered relationships among potentially fraudulent cases.

**Procedure**: This study uses agglomerative clustering, a form of hierarchical clustering, with complete linkage to gradually group data points into clusters. Hierarchical clustering is applied to both feature configurations to examine whether the selected features improve the clarity of detected fraud patterns.

**Evaluation**: Cluster quality is visualized using a Dendrogram, which represents the merging of clusters in hierarchical order, and further validated with the SilhouetteCoefficient to assess the separation between clusters. Hierarchical clustering’s ability to highlight nested relationships is valuable in identifying layered or complex fraud patterns.

**DBSCAN for Anomaly Detection**

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is particularly effective in detecting anomalies within the PPP dataset. By identifying regions of dense data points and labeling sparse regions as noise, DBSCAN helps isolate outliers that may correspond to fraudulent loans.

**Procedure**: DBSCAN is applied to the general PPP dataset, where it labels dense clusters while marking noise points (labeled as -1) as potential anomalies. Parameters eps (distance threshold) and min\_samples (minimum number of points for a dense region) are tuned to balance sensitivity to outliers while minimizing false positives.

**Application for Outlier Detection**: DBSCAN's noise points are flagged as high-risk loans and marked for further analysis in the semi-supervised classification stage. This approach aligns with the goal of identifying previously unmarked loans that exhibit anomalous characteristics consistent with fraud.

**T-SNE for Dimensionality Reduction and Visualization**

T-SNE (t-distributed Stochastic Neighbor Embedding) is applied to visualize clustering results and support interpretability of high-dimensional data. Although primarily a visualization tool, T-SNE assists in examining the separation and cohesion of clusters identified by other methods, adding interpretive depth to clustering results.

**Procedure**: T-SNE reduces the dataset to two-dimensional space, capturing the local and global relationships between data points in clusters. By applying T-SNE to both the full and selected feature sets, this study visualizes the clusters formed by K-Means, Hierarchical, and DBSCAN methods.

**Visualization of Cluster Separation**: Visualizations produced by T-SNE help assess the density and separation of clusters, particularly for complex fraud patterns. These visual insights support cluster validation and offer a preliminary understanding of potential outliers, reinforcing the clustering results and aiding in the interpretation of potential fraud cases.

Each clustering method offers a unique approach to organizing and interpreting the PPP loan data, aiming to detect potential fraud by identifying dense clusters and isolating outliers. Cluster quality and anomaly detection results from these methods will be assessed and combined with labeled fraud cases in the semi-supervised classification stage to improve fraud detection accuracy. This combined approach will test whether the clusters identified here correspond to known fraudulent patterns, thereby addressing RQ1 and testing H1.

To ensure robustness and reproducibility in clustering, this study applies cross-validation techniques adapted for unsupervised learning. These include:

* **Split-Half Reliability**: The dataset is divided into two subsets, and clustering is performed on each. Consistency in cluster assignments across the two subsets is measured using the Adjusted Rand Index (ARI), which evaluates the degree of agreement between clustering results.
* **Bootstrap Sampling**: Repeatedly samples subsets of the data and applies clustering to assess the stability of cluster assignments. This method helps identify how sensitive the clustering model is to variations in the data.
* **Cluster Validity Indices**:
  + **Silhouette Score**: Quantifies the compactness of data points within clusters compared to their separation from other clusters. Higher scores indicate more distinct and well-defined clusters.
  + **Calinski-Harabasz Index**: Measures the ratio of intra-cluster cohesion to inter-cluster separation. This index is particularly useful in comparing clustering results across algorithms or hyperparameter settings.

These metrics will be used iteratively to refine the clustering models, ensuring optimal performance before integrating them into the semi-supervised classification phase.

### Hybrid Approach and Combined Models

The semi-supervised classification phase builds on the results of the clustering analysis by combining labeled fraud data with pseudo-labels generated from unsupervised clustering. This approach leverages both the labeled fraud dataset and the clusters from the PPP loan data to create a more comprehensive model capable of detecting potential fraud cases. The goal is to answer RQ2 by testing various combinations of clustering and classification models, comparing their effectiveness in identifying fraud within the PPP dataset.

**Supervised Models**

Supervised machine learning models are trained on the labeled dataset, which consists of cases from DOJ, SBA OIG, and PRAC reports. These models establish a baseline performance for fraud detection, providing a comparison for later combined model approaches.

**Model Selection**: The study includes a variety of supervised algorithms to test their individual and combined effectiveness:

* 1. **Support Vector Machine (SVM)**: Known for high accuracy in binary classification tasks, SVM is used to establish a high-precision baseline for fraud detection.
  2. **Logistic Regression**: Logistic regression offers interpretability and efficiency, making it suitable for initial fraud classification tasks.
  3. **Neural Networks**: Multi-layer perceptrons are included to capture complex patterns and interactions in the data, particularly valuable for detecting subtle fraud indicators.
  4. **Naïve Bayes**: This probabilistic model is included for its ability to handle categorical data and deliver fast predictions, offering a comparison to more complex models.
  5. **Decision Trees and Ensemble Models**: Decision trees, along with ensemble methods like random forests and gradient boosting, are chosen for their ability to capture non-linear relationships and mitigate overfitting, crucial in high-dimensional fraud detection tasks.

**Baseline Training**: Each supervised model is trained on the labeled fraud dataset, using fraud indicators (fraud = 1) and general loans (fraud = 0) as target labels. Model performance is measured through metrics such as accuracy, AUC-ROC, precision, recall, and F1-Score, which collectively assess each model’s initial classification power.

### Combined Clustering and Classification

The unsupervised clusters from the PPP loan data are integrated with supervised classification models to improve fraud detection performance. By combining clustering and classification outputs, this approach enhances the detection of unlabeled fraud cases, allowing for an evaluation of various unsupervised + supervised model pairings.

**Pseudo-Labeling and Cluster Integration**: The integration of clustering results into classification models transforms unsupervised outputs into actionable inputs for semi-supervised learning. This process combines clustering-derived pseudo-labels with limited labeled data to enhance classification accuracy. The implementation is as follows:

1. **K-Means Clustering**:
   * The K-Means algorithm generates clusters grouping loans with similar attributes (e.g., loan amount, business type).
   * Each loan is assigned a cluster ID, treated as a categorical feature (cluster\_id) in the classification model. This feature captures underlying groupings, enhancing the model’s ability to distinguish between normal and anomalous patterns.
2. **DBSCAN Clustering**:
   * DBSCAN identifies outliers, representing points that deviate significantly from normal data. These outliers are flagged with a binary pseudo-label (pseudo\_fraud = 1), while non-outliers are labeled as pseudo\_fraud = 0.
   * The pseudo-label is integrated into the classification dataset as an additional feature, enabling the model to prioritize high-risk cases flagged by DBSCAN.
3. **Hierarchical Clustering**:
   * Hierarchical clustering provides a multi-level grouping of data points, revealing nested relationships.
   * Loans are assigned to hierarchical cluster levels (e.g., top-level cluster = 1, sub-cluster = 2), encoded as ordinal features representing similarity between data points.
4. **Integration with Classification Models**:
   * The dataset is augmented with clustering-derived features (cluster\_id, pseudo\_fraud, hierarchical cluster level) and labeled fraud cases.
   * Semi-supervised classification models use this hybrid dataset to iteratively refine predictions, leveraging labeled data to validate and enhance clustering outputs.
5. **Feedback and Validation**:
   * The classification models are validated using traditional metrics (e.g., precision, recall, F1-score).
   * Clustering outputs are iteratively refined by adjusting parameters (e.g., the number of clusters in K-Means or epsilon in DBSCAN) to align pseudo-labels with known fraud patterns.

By transforming clustering outputs into features and pseudo-labels, this integration bridges the gap between unsupervised and semi-supervised learning, enabling a robust and interpretable approach to fraud detection in the PPP dataset.

**Model Pairing and Combination**: Each clustering algorithm’s results are paired with supervised models, creating combinations like:

**K-Means + Logistic Regression**: Using K-Means clusters to identify primary groups and logistic regression to classify fraud within each cluster.

**DBSCAN + SVM**: Employing DBSCAN to isolate anomalies and SVM to refine the classification of high-risk cases.

**Hierarchical + Neural Network**: Leveraging hierarchical clusters to represent complex fraud structures, with a neural network trained to detect subtle fraud signals within each cluster.

**T-SNE + Ensemble Models**: Using T-SNE-reduced features for visual cluster patterns and ensemble models to enhance robustness in fraud classification.

**Evaluation of Model Combinations**: The effectiveness of each unsupervised + supervised model pairing is evaluated using metrics such as:

**AUC-ROC**: Measures the model’s ability to distinguish between fraud and non-fraud cases across all thresholds.

**Precision and Recall**: Precision assesses the accuracy of fraud predictions, while recall measures the model’s sensitivity to true fraud cases.

**F1-Score**: Provides a balance between precision and recall, useful for optimizing fraud detection performance where both false positives and false negatives are significant concerns.

**Hyperparameter Tuning**: To ensure optimal performance, each model combination undergoes hyperparameter tuning using grid search or cross-validation. This process fine-tunes model parameters (e.g., SVM’s kernel, decision tree depth) to maximize fraud detection accuracy and reliability.

The combined approach enables the study to test H2 by determining if specific unsupervised + supervised pairings yield superior fraud detection performance. Each combination’s effectiveness is analyzed against the baseline performance of individual models, allowing for a comprehensive evaluation of model interactions and contributions to fraud detection.

### Statistical Analysis and Hypothesis Testing

This phase evaluates the performance of model combinations and feature configurations to test the study’s hypotheses and determine the optimal approach for detecting fraud within the PPP dataset. By using a series of statistical tests and evaluation metrics, this stage aims to validate whether specific combinations of unsupervised and supervised models perform differently and whether a targeted subset of features improves model accuracy in fraud detection.

**Hypothesis Testing for Feature Sets**

To address Hypothesis 1 (H1), this study evaluates the effectiveness of the full feature set compared to the selected subset of key features. The goal is to determine if clustering and classification models perform optimally with all features or if a subset yields statistically significant improvements in fraud detection accuracy.

**Method**: Models trained on both the full and subset feature configurations are compared using accuracy, AUC-ROC, precision, recall, and F1-Score metrics. These performance metrics offer insights into how each feature set influences model performance and detection power.

**Model Performance Comparison Across Combinations**

To address Hypothesis 2 (H2), this phase evaluates the performance of different unsupervised and supervised model pairings. The aim is to determine whether specific model combinations provide statistically significant advantages in fraud detection, reflecting the effectiveness of integrated clustering and classification.

**Method**: Each unsupervised + supervised model pairing is evaluated on key metrics, including AUC-ROC, precision, recall, and F1-Score, with results compared against baseline models trained on labeled data only. By analyzing multiple model pairings, this study tests whether some combinations perform better than others in detecting fraudulent cases.

**Statistical Tests**:

**ANOVA**: ANOVA tests are used to examine the variance in model performance metrics across different model pairings. This test helps identify any significant differences in mean performance across combinations, indicating which pairing is most effective.

**McNemar’s Test**: For paired binary classifications (e.g., detecting fraud or non-fraud), McNemar’s Test assesses differences in predictive outcomes between models, identifying statistically significant improvements in fraud detection from specific model pairings.

**Paired t-tests or Wilcoxon Signed-Rank Test**: For non-normally distributed metrics, the Wilcoxon Signed-Rank Test serves as a non-parametric alternative to t-tests, enabling comparisons between model pairs in detecting fraud cases.

**Interpretation**: Statistically significant differences in performance metrics between model pairings would support H2ₐ, indicating that certain unsupervised + supervised combinations perform better in detecting fraud. If no significant difference is found across model pairings, this would support H2₀, suggesting that all model pairings perform comparably in detecting fraud.

**Evaluation Metrics for Validation**

Each hypothesis testing stage uses specific metrics to validate model performance:

**AUC-ROC**: Evaluates the model’s ability to distinguish between fraud and non-fraud across all probability thresholds, providing a robust measure of classification performance.

**Precision and Recall**: Precision focuses on minimizing false positives, while recall emphasizes sensitivity to true positives, both critical in assessing fraud detection efficacy.

**F1-Score**: Balances precision and recall, offering a single metric to evaluate the trade-off between detecting fraud accurately and reducing false positives.

**Silhouette Score (for Clustering)**: Measures cohesion within clusters and separation between clusters, offering insight into the clarity of fraud-related patterns identified through clustering.

These metrics ensure that model performance is measured comprehensively, capturing both the accuracy of fraud detection and the consistency of clustering.

## Assumptions

This study is built on several assumptions to support the research methodology, data interpretation, and conclusions:

1. **Data Completeness and Accuracy**: It is assumed that the PPP loan dataset provided by SBA.gov accurately represents the entire population of PPP loan applications. Additionally, it is assumed that records in the labeled fraud dataset, obtained from sources such as DOJ, SBA OIG, and PRAC, are verified instances of fraud.
2. **Independence of Fraudulent Patterns**: The study assumes that fraud patterns within the PPP dataset are not random and exhibit detectable statistical regularities, allowing machine learning models to learn and generalize these patterns for accurate fraud detection.
3. **Model Generalizability**: The machine learning models trained and validated in this study are assumed to generalize well to new and unseen PPP loan data. This assumption implies that fraud patterns identified in the study’s training data are representative of broader trends in PPP fraud.
4. **Feature Integrity**: It is assumed that feature engineering, encoding, and transformations do not introduce bias or distort the original data’s characteristics, allowing models to focus on genuine indicators of fraud.

## Limitations

Despite the rigorous methodology, several limitations may affect this study's generalizability and interpretability. The following limitations are noted, along with strategies implemented to mitigate their impact:

1. **Data Bias in Labeled Fraud Cases**:

**Limitation**: The labeled fraud dataset relies on publicly reported and prosecuted cases, which may not represent the full spectrum of fraudulent activity within the PPP. Fraud cases that remain undetected or unprosecuted could introduce bias, limiting the model's ability to generalize to unseen types of fraud.

**Mitigation**: To address this limitation, the study includes unsupervised clustering methods on the unlabeled PPP data to identify patterns and anomalies that might indicate other types of fraud beyond those already labeled. This clustering step helps broaden the model’s exposure to potential fraud patterns, even in the absence of explicit labels.

1. **Feature Constraints Due to Data Privacy**:

**Limitation**: Personally identifiable information (PII) is hashed to maintain confidentiality, potentially limiting the interpretability of some features. Additionally, the study relies solely on publicly available attributes, restricting access to deeper financial data or internal audit findings that might improve model precision.

**Mitigation**: Although hashing reduces interpretability, it retains the uniqueness of each record, preserving valuable pattern associations. To enhance interpretability further, the study applies feature engineering on non-PII variables to create derived indicators, such as ForgivenessAmountRatio, which capture meaningful aspects of financial behavior that may correlate with fraud.

1. **Cluster Interpretability in High Dimensions**:

**Limitation**: Clusters formed by unsupervised models like K-Means and DBSCAN may be difficult to interpret, particularly in high-dimensional spaces, which can obscure the specific factors contributing to each cluster.

**Mitigation**: Dimensionality reduction techniques, specifically T-SNE, are applied for visualization to aid in interpreting cluster patterns. Additionally, model interpretability is enhanced through the application of semi-supervised models, where labeled fraud data provides clearer classification boundaries, allowing for a more interpretable analysis of fraud patterns within clusters.

1. **Hyperparameter Sensitivity in Clustering and Classification Models**:

**Limitation**: The performance of clustering and classification algorithms can be highly sensitive to hyperparameter choices. Incorrect or suboptimal parameter settings could lead to either overfitting or underfitting, affecting model reliability and generalizability.

**Mitigation**: To minimize this sensitivity, the study employs cross-validation and grid search techniques to systematically optimize key hyperparameters, such as eps in DBSCAN and the number of clusters in K-Means. This optimization procedure is applied iteratively to ensure stable model performance and reduce the risk of overfitting.

1. **Imbalance Between Fraud and Non-Fraud Cases**:

**Limitation**: Fraud cases constitute a small proportion of the PPP loan dataset, potentially leading to class imbalance in the semi-supervised model, where the model may struggle to identify rare fraud cases among the more prevalent non-fraud cases.

**Mitigation**: To address class imbalance, the study applies resampling techniques, including Synthetic Minority Over-sampling Technique (SMOTE), to balance the labeled training dataset. Additionally, performance metrics like precision, recall, and F1-Score are selected to focus on the model’s ability to detect fraud cases effectively, ensuring that evaluation prioritizes sensitivity to the minority (fraud) class.

## Delimitations

This study is defined by specific boundaries set intentionally to maintain focus and feasibility, given the research objectives and available resources. These delimitations are outlined below, with rationales for each decision:

1. **Dataset Focus on PPP Loans**:

**Delimitation**: This study exclusively utilizes data from the PPP and does not integrate datasets from other COVID-19 relief programs, such as the Economic Injury Disaster Loan (EIDL) program.

**Rationale**: The PPP dataset is selected for its large volume and the widespread public concern surrounding PPP-related fraud. Focusing on a single program allows for a detailed analysis tailored to the unique features and fraud patterns within PPP, enhancing the specificity of findings and avoiding potential confounding effects from differences in program requirements or disbursement mechanisms.

1. **Limitation to Publicly Available Features**:

**Delimitation**: Only publicly available features within the PPP dataset are used, and additional data fields are not included due to privacy concerns and access limitations.

**Rationale**: Publicly available data ensures transparency and reproducibility in research findings while respecting data privacy regulations. By restricting the study to accessible fields, this approach maintains ethical standards and allows future researchers to replicate the study using the same data sources, supporting validation and comparison efforts.

1. **Temporal Boundaries of the PPP Data**:

**Delimitation**: The study is limited to PPP loans issued during the COVID-19 pandemic and does not consider loans or fraud cases beyond this specific period.

**Rationale**: The temporal focus on the COVID-19 relief period allows the study to analyze fraud patterns in the unique context of a large-scale, rapidly implemented government aid program. This boundary enhances the relevance of the findings by situating them within a specific period of heightened financial vulnerability, where fraud mechanisms may differ from standard loan environments.

1. **Focus on Feature Selection Relevant to Fraud Detection**:

**Delimitation**: Feature engineering is limited to variables directly relevant to fraud detection, and extraneous features that do not meaningfully contribute to identifying fraud indicators are excluded from analysis.

**Rationale**: Limiting features to those pertinent to fraud detection reduces noise within the dataset, increasing model efficiency and interpretability. By focusing on features with clear relevance, this delimitation ensures that computational resources are devoted to the most impactful data, improving model clarity and enhancing fraud detection accuracy.

**Ethical Considerations (Secondary data)**

Ethical considerations are a priority in this study due to the sensitivity of the PPP dataset and the involvement of PII. To protect participant privacy and comply with ethical standards, the following measures are implemented:

1. **Data Privacy and Confidentiality**: All PII, such as BorrowerName, BorrowerAddress, and ServicingLenderName, is hashed to ensure privacy without compromising the uniqueness of each record. This transformation follows data protection standards, safeguarding confidentiality while maintaining data integrity.
2. **Use of Publicly Available Data**: The study exclusively uses data that is publicly accessible. Fraud labels are derived from verified public sources, such as DOJ press releases, SBA OIG reports, and PRAC data. This approach ensures transparency and eliminates reliance on private or confidential sources.
3. **Ethical Oversight and IRB Compliance**: The study has obtained approval from the Institutional Review Board (IRB) to confirm compliance with ethical standards in research. This includes assurances regarding data use, privacy protections, and informed consent protocols, even for secondary data use.
4. **Transparency in Methodology**: The research methods, including data processing, feature engineering, and model design, are documented transparently to ensure reproducibility and integrity in results. This transparency allows other researchers to replicate or extend the study while maintaining ethical consistency.

These ethical assurances ensure that the study adheres to privacy standards and ethical principles, protecting data integrity and participant confidentiality.

## Summary

Chapter 3 outlined the comprehensive methodology employed to investigate fraudulent loan applications within the Paycheck Protection Program (PPP) dataset, focusing on two key objectives: identifying primary fraud indicators and evaluating effective machine learning model combinations for fraud detection. The chapter began with a detailed description of the data sources, encompassing both the publicly available PPP loan dataset and a manually labeled fraud dataset derived from verified cases from DOJ, SBA OIG, and PRAC records. This dual-source approach provides a foundation for both unsupervised and semi-supervised learning phases.

Data preprocessing steps were carefully executed to ensure privacy, consistency, and quality, incorporating techniques such as PII hashing, categorical encoding, and imputation of missing values. Feature engineering and feature selection further refined the dataset by deriving fraud-relevant indicators and testing the impact of selected feature subsets on model performance. These steps directly support the research’s hypothesis-driven approach to evaluating the effectiveness of specific features in fraud detection.

The chapter also detailed the clustering methodologies, including K-Means, Hierarchical Clustering, DBSCAN, and T-SNE, applied to detect potential fraud patterns within the general PPP loan data. These unsupervised models provided a basis for anomaly detection, where clusters and outliers serve as preliminary indicators of possible fraud, aligning with the study’s aim to detect fraud patterns without extensive labeled data.

Semi-supervised classification techniques were then introduced, integrating labeled fraud cases with clustering results to improve model performance. By testing combinations of unsupervised clustering with supervised models such as SVM, logistic regression, neural networks, and ensemble methods, the study evaluates the effectiveness of different model pairings for enhanced fraud detection.

The final sections discussed the study’s assumptions, limitations, and delimitations, acknowledging the challenges in balancing data privacy, model interpretability, and class imbalance while addressing each with targeted mitigation strategies. Statistical validation methods and evaluation metrics were selected to ensure that the results of model performance testing are both statistically sound and practically relevant, with metrics tailored to accurately measure clustering quality and classification accuracy in fraud detection.

Together, these methodological steps provide a structured approach to answering the study’s research questions and testing its hypotheses, ensuring that each phase of the research design aligns with the overarching goals of detecting and analyzing fraud patterns within the PPP dataset. Chapter 4 will present the results of these analyses, including model performance evaluations and interpretations of clustering and classification outcomes in the context of PPP fraud detection.

# 

# Chapter 4: Findings

Begin writing here…

Checklist:

Begin with an introduction and restatement of the problem and purpose sentences verbatim and the organization of the chapter.

Organize the entire chapter around the research questions/hypotheses.

Identify all the steps that were described in Chapter 3.

Link to your GitHub code and data should be included here

License codes and links should be included here.

## Data Preprocessing and Modeling Process Diagram

Include the Diagram created as a diagramming tool that describes the process of your study and depicts all the details of all stages of your completed research. Small diagrams that refer to components of your product, algorithm, or design can be added below.

## Data Preprocessing and Modeling

Begin writing here…

Checklist:

Describe the different stages of your study, referring to all details recorded.

Describe the exact steps of your study plan:

### Data Cleaning

Include the size, observations, variables, types of variables, missing values, affected variables of the original dataset, the methods of imputation you plan to include, size, and limitations after the cleaning (if known otherwise, you should consist of those in Chapter 4), etc.

Detail the cleansing processes, such as outlier detection, handling missing data through imputation and deduplication, and ensuring data quality.

### Data Preprocessing

Explain the transformations of variables, including normalization, standardization, encoding categorical data, and introducing artificial variables. Document the methods used with supporting literature.

Discuss strategies like limiting sample size, bootstrapping, and resampling techniques. Validate these processes with appropriate statistical methods.

Mention the software programs, platforms, and their versions used for data preparation. Include programming scripts or modules in the Appendix.

### Data Integration

Describe the process and tools for merging data from different digital sources or datasets, addressing issues like schema integration, entity resolution, and data format normalization.

Include the programming module for data integration in the Appendix.

### Data Feature Engineering

Explain how and why you are creating interaction features between variables.

Discuss the creation of polynomial features and their relevance to the model.

Detail any aggregated features from multiple data sources or grouped observations.

Describe the methods used to scale features and the rationale behind the choice.

Justify the use of log transformation for reducing skewness in continuous variables.

Outline using PCA or other techniques to reduce the feature space.

Discuss methods for selecting the most relevant features (e.g., filter, wrapper, and embedded methods).

If relevant, detail how date and time information is transformed into features useful for prediction (e.g., extracting day of the week, lag features).

For datasets with textual data, describe preprocessing steps like tokenization, stemming, and lemmatization.

Present an analysis of feature importance and how it influenced the feature engineering process.

### Data Exploration

You should include the data exploration, identifying the similarities, contrasts, and anomalies you observed. Documentation should include graphs, charts, and tables created programmatically. (Do not include ALL different graphs for all variables; instead, focus on the findings that provide new insights, guide your research, and alter your steps or identified constraints).

Examples can be qqplots proving normality and histograms. Parallel boxplots and time series plots identify differences in distributions and boxplots for initial populations with vast amounts of data. 5-number summaries in tables along with standard deviations, recognizing differences in internal variation of comparing populations, loess graphs that compare time series, interactive graphs identifying differences of tendency measures, heatmaps, scatterplot matrices, etc…

Document only the most insightful visualizations and tables that guide the research direction.

## Data Mining

Follow the CRISP-DM model, iterating as necessary, and report only the final results.

Compare them with alternatives and support your choice with evidence from literature or analysis findings.

Detail the input, training choices, and ratio for training. Justify these choices with statistical validation.

Include outputs like decision trees, charts, or diagrams, explaining their relevance.

Document any challenges faced and insights for future research.

Include relevant programming modules in the Appendix.

### Model Validation and Hyperparameter Tuning

Validate the model to ensure its accuracy and generalizability. Employ methods like cross-validation, ROC curves, precision-recall curves, and confusion matrices. The above are just examples of measures for validation; depending on your research focus, another type of measure may be appropriate. (For LLMs, Blue Score, Rouge Score, Hallucinations Score etc.)

Discuss hyperparameter tuning strategies, such as grid or random search, and their impact on model performance.

Compare the model with alternative models or algorithms, focusing on error metrics and performance measures.

If applicable, use triangulation or other methods to validate the model's results.

Include the programming modules for model validation and hyperparameter tuning in the Appendix.

### Data Modeling

In this paragraph, you should discuss the modeling method or methods you followed in every detail.

Discuss in detail why the requirements of using the specific model are satisfied by your sample. (you may need to support this with evidence, graphs, or diagnostics)

Describe the input selection process, training, testing, the data you used for testing, and the included or excluded variables.

Explain the different approaches you followed to your resulting model.

Support with evidence about the superiority of the resulting model and discuss the methods you followed to reach that. This may include measures of error, methods of dropping variables, RSME, entropy, AIC/BIC or stepwise regression, ANOVA tables followed by Tukey’s test, etc., and/or other diagnostics.

Present relevant graphs, charts, and tables generated by the algorithm or created for analysis.

Include programming modules related to data modeling in the Appendix.

## Results

Begin writing here…

Checklist:

Briefly discuss the overall study. Organize the presentation of the results by the research questions/hypotheses.

Objectively report the results of the analysis without discussion, interpretation, or speculation.

Provide an overview of the demographic information collected. It can be presented in a table. Ensure no potentially identifying information is reported.

### Data Modeling Evaluation

Explanation and metrics in tabular and graphical format.

If applicable, present the results of any predictive models or machine learning algorithms.

Discuss model performance metrics (accuracy, precision, recall, F1 score, etc.).

Compare different models if you experimented with multiple approaches.

### Data Analysis

Present post-EDA and comprehensive analysis for drawing inferences at the end of the study.

Present findings in written and visual forms, including discussions on graphs, charts, and tables.

Include the programming module used for data analysis in the Appendix.

### Model Comparisons and Diagnostics

Complete written, tabular, and graphical explanations.

#### Model Fit and Diagnostics

Evaluate the model fit using appropriate statistical tests and diagnostic plots like residual plots, qqplots, and influence plots.

Discuss the methods used for diagnosing issues with the model, such as overfitting or underfitting, and the steps taken to address these issues.

Include evidence supporting the model's fit, such as R-squared, adjusted R-squared, F-statistics, and p-values.

For Classification Models: Elaborate on classification metrics such as accuracy, precision, recall, F1 score, and the AUC-ROC curve. Explain the significance of each metric and the scenarios in which they are instrumental. (consider entropy as well when applicable)

For Regression Models: Discuss regression metrics like MSE, MAE, and R-squared. Provide insight into what these metrics convey about model performance.

Custom Metrics: If any custom metrics are used, describe them and justify their relevance to the project's objectives.

Optimization Metrics and Findings

#### Optional Sections based on the project

Validation and Sensitivity Analysis:

* Discuss how the validity of your findings was assessed.
* Perform sensitivity analysis if relevant.

Subgroup Analysis:

* If your study involves different subgroups, present findings for each subgroup.
* Discuss any variations or patterns observed across subgroups.

Time-Series Analysis (if applicable):

* If your data involves a temporal dimension, present time-series analyses.
* Identify trends or patterns over time.

Spatial Analysis (if applicable):

* If your data has a spatial component, present spatial analyses.
* Use maps or spatial visualizations to convey information.

In the case of Inferential Hypothesis Testing:

* + - State the hypotheses tested in your study.
    - Present the results of statistical tests (t-tests, ANOVA, chi-square, etc.).
    - Include p-values and confidence intervals.

#### Error Analysis and Model Refinement:

Conduct a thorough error analysis to understand the model's limitations and areas for improvement.

Refine the model based on the error analysis, adjusting features, model parameters, or algorithm.

Document the model refinement process and the subsequent improvements in model performance.

#### Model Interpretation and Explainability:

Provide a clear interpretation of the model results, explaining the significance and impact of critical features.

If applicable, use techniques for model explainability, such as feature importance scores, SHAP values, or LIME, to make the model's decisions more transparent.

### Research Question # (Hypothesis when necessary)

Repeat for each question

Text…

Report all the results (without discussion) salient to the research question/hypothesis. Identify common themes or patterns.

Use tables and/or figures to report the results as appropriate.

For quantitative studies, report any additional descriptive information as appropriate. Identify the assumptions of the statistical test and explain how the extent to which the data met these assumptions was tested. Report any violations and describe how they were managed as appropriate. Make decisions based on the results of the statistical analysis. Include relevant test statistics, *p* values, and effect sizes in accordance with APA requirements.

## Evaluation of the Findings

Begin writing here…

Checklist:

Interpret the results considering the existing research and theoretical or conceptual framework (as discussed in Chapters 1 and 2). Briefly indicate the extent to which the results were consistent with existing research and theory.

Organize this discussion by research question/s

Do not draw conclusions beyond what can be interpreted directly from the results.

Devote approximately one to two pages to this section.

## Limitations

Discuss any limitations of your study, including data limitations, methodological constraints, or external factors that may impact the findings.

## Summary

Begin writing here…

Checklist:

Summarize the key points presented in the chapter.

# 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…