A

Project Report

on

Credit Card Fraud Detection using Anomaly Classifiers

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2023-24

in

Computer Science and Engineering

By

Prerna Singh (2000290100104)

Prince Piyush (2000290100106)

Khyati Singla (2000290100082)

Under the supervision of

Assistant Professor Ms. Bharti

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2024

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Date: 12/05/2024

Signature

Name: Prerna Singh

Roll No.: 2000290100104

Signature

Name: Prince Piyush

Roll No.: 2000290100106

Signature

Name: Khyati Singla

Roll No.: 2000290100082

CERTIFICATE

This is to certify that Project Report entitled “Credit Card Fraud Detection Using Anomaly Classifiers” which is submitted by Prerna Singh, Prince Piyush and Khyati Singla in partial fulfillment of the requirement for the award of degree B. Tech. in Department of Computer Science & Engineering of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Supervisor Name: Dr. Vineet Sharma

(Designation): (HoD-Computer Science & Engineering)

Date: 12/05/2024

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe a special debt of gratitude to Ms. Bharti, Department of Computer Science & Engineering, KIET, Ghaziabad, for her constant support and guidance throughout the course of our work. Her sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Dr. Vineet Sharma, Head of the Department of Computer Science & Engineering, KIET, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members, especially faculty/industry person/any person, of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Date: 12/05/2024

Signature:
Name: Prerna Singh
Roll No.: 2000290100104

Signature:
Name: Prince Piyush
Roll No.: 2000290100106

Signature:
Name: Khyati Singla
Roll No.: 2000290100082

ABSTRACT

In the realm of financial transactions, particularly in the context of the burgeoning internet and e-commerce landscape, the specter of credit card fraud looms large. As online transactions become increasingly prevalent, so too do the opportunities for fraudulent activities to proliferate. The need for robust defense mechanisms against such threats has never been more pressing. Enter Anomaly Detection – a sophisticated arsenal of algorithms designed to discern aberrant patterns within datasets, thereby preemptively identifying fraudulent behavior.

This study delves into the realm of Anomaly Detection, spotlighting two pivotal algorithms: the Isolation Forest (IF) and the Local Outlier Factor (LOF). Its primary objective? To ascertain the efficacy of these algorithms in flagging anomalies within credit card transaction datasets, thereby bolstering fraud detection capabilities. However, the journey doesn't end there. Recognizing the inherent challenge posed by imbalanced datasets – where instances of fraudulent transactions pale in comparison to their legitimate counterparts – the study explores various resampling strategies to rectify this imbalance.

Two datasets serve as the crucible for this investigation: the European Credit Card Fraudulent transactions dataset and the German Credit Card fraud dataset. Both datasets provide fertile ground for experimentation, each laden with its own unique intricacies and challenges. However, before the algorithms can be unleashed upon these datasets, they must first be primed and optimized. This involves fine-tuning parameters, selecting appropriate feature sets, and, crucially, addressing the issue of dataset imbalance through resampling.

Enter an array of resampling techniques, each poised to tackle the challenge of dataset imbalance head-on. Random Undersampling, AllKNN, Synthetic Minority Oversampling Technique (SMOTE), and Synthetic Minority Oversampling Technique - Edited Nearest Neighbor (SMOTE-ENN) all take center stage, each vying to rebalance the datasets and level the playing field for the algorithms to come.

With the stage set and the algorithms poised for action, the study proceeds to execute its meticulously crafted experimental protocol. The results are nothing short of revelatory. The Isolation Forest classifier emerges as a veritable juggernaut, boasting an impressive accuracy of 99.81% when unleashed upon the imbalanced European credit card fraudulence dataset. Meanwhile, its counterpart, the LOF classifier, coupled with the Random Undersampling technique, delivers a commendable accuracy of 70.60% when applied to the German credit card dataset.

These findings represent a watershed moment in the ongoing battle against credit card fraud. They underscore the pivotal role that Anomaly Detection algorithms play in fortifying financial systems against the ever-evolving threat landscape. By harnessing the power of Isolation Forest, LOF, and a host of resampling techniques, financial institutions can markedly enhance their ability to detect and deter fraudulent activities. Ultimately, the study serves as a clarion call for continued innovation and vigilance in safeguarding the security of digital transactions – a mission critical to preserving consumer trust and confidence in financial systems worldwide.

TABLE OF CONTENTS

Page No.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

IF	Isolation Forest
LOF	Local Outlier Factor
SMOTE	Synthetic Minority Oversampling Technique
SMOTE-ENN	Synthetic Minority Oversampling Technique – Edited Nearest Neighbor
EDA	Exploratory Data Analysis
ANN	Artificial Neural Network
GA	Genetic Algorithm
AUC	Area Under the Precision-Recall curve
LR	Logistic Regression
NB	Naive Bayes
KNN	K-Nearest Neighbor
MLP	Multi-Layer Perceptron
LSTM	Long Short-Term Memory
DCNN	Deep Convolutional Neural Network
OCSVM	One-Class Support Vector Machine
ENN	Edited Nearest Neighbors
PCA	Principal Component Analysis

TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives

CHAPTER 1
INTRODUCTION

1.1 Introduction

Credit card fraud stands as a persistent menace in the realm of financial transactions, persistently challenging the security of digital commerce. Hackers and fraudsters, constantly innovating their techniques, exploit vulnerabilities in payment systems, capitalizing on the rapid advancement of technology. The proliferation of online banking, high-tech payment solutions, and the widespread use of debit and credit cards have significantly expanded the avenues for fraudulent activities to thrive. While these innovations offer unparalleled convenience and efficiency, they have also paved the way for a surge in financial fraud, encompassing nefarious practices like money laundering, card theft, and identity theft.

The landscape of credit card fraud is multifaceted, with distinct categories delineating the nature of illicit activities. On one front, criminals engage in the unlawful acquisition of physical credit cards, either through theft or other illicit means. On another front, the illicit procurement of sensitive card-related information, such as card numbers, CVV codes, and card types, enables fraudsters to execute unauthorized transactions and exploit unsuspecting victims for financial gain. These fraudulent activities not only jeopardize the financial well-being of individuals but also erode trust and stability within the broader financial ecosystem.

The staggering financial losses attributed to credit card fraud underscore the magnitude of the challenge at hand. From $22.8 billion in 2017 to a staggering $31 billion in 2020, the global impact of credit card fraud continues to escalate unabated. Such astronomical losses reverberate across businesses and consumers alike, with a growing number of individuals falling victim to fraudulent activities each passing year.

Identity theft emerges as a particularly pernicious subset of credit card fraud, exacting a heavy toll on unsuspecting victims. The Federal Trade Commission fields hundreds of thousands of complaints annually, highlighting the pervasive nature of this threat. Moreover, the proliferation of new credit card accounts exacerbates the problem, with identity theft incidents witnessing a sharp uptick of 48% in 2020 alone. Projections indicate a grim trajectory, with credit card theft losses poised to reach an alarming $38.5 billion by 2027.

In the quest to combat credit card fraud, technological advancements offer a glimmer of hope. Artificial intelligence, data mining, and machine learning techniques have emerged as promising tools in the detection and prevention of fraudulent activities. However, their widespread adoption is impeded by the sensitive nature of credit card data, necessitating the use of anonymized datasets in model development. Despite these challenges, ongoing research and innovation are essential to stay ahead of evolving fraud patterns and fortify the resilience of financial systems.

In summation, credit card fraud represents a formidable challenge in the digital age, demanding a concerted and multifaceted response. Through collaborative efforts, leveraging advanced technologies, and implementing robust detection mechanisms, stakeholders can mitigate the impact of credit card fraud and foster greater trust and security in financial transactions. Only through proactive measures and sustained vigilance can we safeguard the integrity of financial systems and preserve consumer confidence in the digital economy.

1.2 Project Description

The project titled "Anomaly Detection in Credit Card Transactions using Machine Learning" delves into the urgent and complex issue of credit card fraud within the burgeoning digital economy. Its abstract serves as a succinct yet illuminating overview, underscoring the pivotal role of Anomaly Detection methodologies in combating fraudulent activities. Specifically, the project focuses on harnessing the power of the Isolation Forest (IF) and Local Outlier Factor

(LOF) algorithms to identify anomalies within credit card transaction datasets. Moreover, it recognizes the challenge posed by imbalanced datasets and addresses this through the implementation of various resampling strategies, including Random Undersampling, AllKNN, SMOTE, and SMOTE-ENN, aimed at rectifying data imbalances and improving fraud detection accuracy.

At the heart of the project lies a profound motivation to fortify fraud detection mechanisms, thereby safeguarding customer trust and financial security. This imperative drives the project's emphasis on leveraging advanced techniques such as machine learning, artificial intelligence, and data analytics. A comprehensive review of existing literature underscores the efficacy of machine learning algorithms in detecting credit card fraud, providing valuable insights that inform the subsequent experimental investigation.

The project proceeds methodically, commencing with exploratory data analysis (EDA) and meticulous data preprocessing to prepare the European credit card fraud detection dataset and the German credit card fraudulent dataset for model training. Subsequently, machine learning models, including Isolation Forest and Local Outlier Factor, are trained on these preprocessed datasets, and their performance is rigorously evaluated using a suite of metrics such as accuracy, precision, recall, and F1-score. The implementation phase entails a systematic comparison of the algorithms and resampling techniques, facilitating the identification of the most effective approach for fraud detection in credit card transactions.

Upon analysis of results and ensuing discussions, the project unveils compelling insights. The Isolation Forest classifier emerges as a standout performer, exhibiting superior efficacy in detecting fraudulent transactions within the imbalanced European dataset. Conversely, the LOF algorithm, when coupled with Random Undersampling, achieves commendable accuracy in detecting fraud within the German dataset. These findings underscore the importance of proactive measures in combating credit card fraud and offer actionable recommendations for enhancing fraud detection capabilities in financial transactions.

In conclusion, the project serves as a beacon of innovation and diligence in the ongoing battle against credit card fraud. By harnessing the power of machine learning and sophisticated algorithms, it offers a promising pathway towards greater security and resilience in the digital financial landscape. Moreover, it lays the foundation for future research endeavors aimed at further refining and enhancing fraud detection mechanisms, thereby fortifying the integrity of financial transactions and preserving consumer trust in the digital era.

CHAPTER 2

LITERATURE REVIEW

Ref [1] research shows several algorithms that can be used for classifying transactions as fraud or genuine one. Credit Card Fraud Detection dataset was used in the research. Because the dataset was highly imbalanced, SMOTE technique was used for oversampling. Further, feature selection was performed and dataset was split into two parts, training data and test data. The algorithms used in the experiment were Logistic Regression, Random Forest, Naive Bayes and Multilayer Perceptron. Results show that each algorithm can be used for credit card fraud detection with high accuracy. Proposed model can be used for detection of other irregularities. Ref [2] explores the application of multiple machine learning algorithms, including Support Vector Machine (SVM), k-Nearest Neighbor (KNN), and Artificial Neural Network (ANN), for predicting credit card fraud occurrences. The study compares supervised machine learning techniques with deep learning approaches, focusing on distinguishing between fraud and non-fraud transactions. An ANN model is developed and trained, demonstrating superior accuracy compared to traditional machine learning algorithms. Data preprocessing techniques such as normalization and under-sampling are employed to address imbalanced datasets, enhancing model performance in fraud detection.

Ref [3] compares the performance of Random Forest and Adaboost algorithms for credit card fraud detection based on accuracy, precision, recall, and F1-score metrics. Results show comparable accuracy between the two algorithms, but Random Forest outperforms Adaboost in terms of precision, recall, and F1-score. The ROC curve, generated from the confusion matrix, aids in visualizing algorithm performance. Ultimately, Random Forest is deemed more effective in detecting credit card fraud due to its superior precision, recall, and F1-score metrics compared to Adaboost.

In Ref [4], hyperparameter optimization was conducted using GridSearchCV and random search. The Decision Tree classifier achieved 72.1% accuracy on the imbalanced German credit card dataset, while LDA demonstrated an impressive 98.6% accuracy on the imbalanced European credit card fraud dataset. The study also explored SMOTE, Naive Bayes, and Logistic Regression, with Decision Tree emerging as the superior performer compared to LR, LDA, and Naive Bayes. Ref [5] proposed a comprehensive ML framework, amalgamating LR, SVM, RF, NB, XGBoost, and ET algorithms with AdaBoost to enhance classification efficacy. Performance evaluation was based on metrics such as accuracy, recall, precision, MCC, and AUC. In Ref [6], a resilient deep-learning methodology was introduced, incorporating LSTM and GRU neural networks as foundational classifiers within a stacking ensemble structure. Additionally, a Multi-Layer Perceptron (MLP) functioned as the higher-level learner. To address class distribution imbalance, the SMOTE-ENN technique was employed in the dataset, ensuring balanced representation and improving model performance.

Ref [7] conducts a comprehensive performance assessment of nine distinct classifier models for credit card fraud detection. These models include Logistic Regression, K-Nearest Neighbors, Random Forest, Naive Bayes, Multi-Layer Perceptron, AdaBoost, Quadrant Discriminative Analysis, Ensemble Learning, and Pipelining. To address dataset imbalance, the ADASYN technique is employed. Remarkably, the Pipelining method emerges as the top performer across various metrics, showcasing superior performance in fraud detection.

In Ref [8], a financial fraud detection scheme is introduced utilizing a Deep Convolutional Neural Network (DCNN) and deep learning algorithms to identify fraudulent transactions. The scheme achieves an impressive accuracy rate of 99% within a 45-second timeframe. Ref [9] employs a Genetic Algorithm (GA) for feature selection, incorporating a selection of ML classifiers including Decision Tree, Random Forest, Logistic Regression, Artificial Neural Network, and Naive Bayes. Notably, the GA is integrated within the Random Forest's fitness function, resulting in the GA-RF amalgamation achieving an outstanding overall accuracy of 99.98%.

In Ref [10], a hybrid approach combining data resampling and neural network ensemble classifiers is introduced. The primary trainee in this ensemble classifier is the Long Short-Term Memory (LSTM) neural network, implemented within the AdaBoost framework. Additionally, a hybrid resampling strategy incorporating the SMOTE-ENN method is employed. Results suggest that utilizing SMOTE-ENN data resampling in conjunction with the boosted LSTM classifier is a promising strategy for effectively detecting fraudulent activities in credit card transactions.

Ref [11] research aimed to compare machine learning algorithms' ability to accurately classify fraud and non-fraud credit card transactions using the random undersampling method. Logistic Regression (LR) outperformed Naive Bayes (NB) and K-Nearest Neighbor (KNN), achieving optimal performance across all data proportions. LR demonstrated the highest accuracy of 95%, while NB achieved 91%, and KNN trailed at 75%. LR also exhibited superior Sensitivity, Specificity, Precision, and F-Measure compared to NB and KNN. The study concluded that supervised techniques like LR and NB consistently outperformed the unsupervised technique KNN in classifying credit card transactions.

In Ref [12], the Isolation Forest (IF) model was implemented for fraud detection, utilizing the Area Under the Precision-Recall curve (AUC) as a performance metric. The study demonstrated superior outcomes with an efficiency noted at 98.72%. This approach highlights the effectiveness of IF in identifying anomalies within credit card transaction data, with a focus on precision-recall trade-offs. The utilization of AUC for evaluating model performance signifies a nuanced understanding of the detection task, emphasizing the importance of capturing true positives while minimizing false positives. Ref [13] employed the one-class Support Vector Machine classifier (OCSVM) for outlier detection in credit card transactions. Hyperparameters γ and ν were optimized via grid search to maximize the AUC, showcasing the model's robustness in capturing anomalous instances. The GS-OCSVM exhibited superior fraud detection compared to the isolation forest, with higher true negative rates observed across both German and European credit card datasets. This highlights the efficacy of OCSVM in discerning fraudulent transactions from normal ones, leveraging a one-class learning approach tailored for outlier detection tasks.

In Ref [14], Python was utilized to implement the Isolation Forest (IF) and Local Outlier Factor (LOF) algorithms for credit card fraud detection. Following dataset evaluation, it was observed that the IF algorithm yielded a higher precision rate compared to LOF. This underscores the efficacy of IF in accurately identifying fraudulent transactions while minimizing false positives, contributing to enhanced fraud detection capabilities in financial transactions.

In Ref [15], the Isolation Forest (IF) algorithm's effectiveness in credit card fraud detection is discussed and contrasted with the Local Outlier Factor (LOF) algorithm. IF demonstrates superior performance over LOF across multiple metrics, including error rate, accuracy score, and recall. Notably, IF achieves a significantly higher rate of fraud detection at approximately 27%, compared to LOF's mere 2%. Moreover, the IF model exhibits exceptional precision at 99.774%, surpassing LOF's precision of 99.65%. These findings underscore IF's efficacy in accurately identifying fraudulent transactions while minimizing false positives compared to LOF. Ref [16] focuses on addressing imbalanced data through the random undersampling approach and applies three machine learning algorithms—Logistic Regression (LR), Naive Bayes (NB), and K-Nearest Neighbors (KNN)—for fraud detection. The efficacy of these algorithms is evaluated using comprehensive metrics such as accuracy score, F-measure, sensitivity, precision, specificity, and Area Under the Curve (AUC). By employing random undersampling to rebalance the dataset, the study aims to enhance the performance of these algorithms in accurately classifying fraudulent and non-fraudulent transactions, thereby contributing to the development of robust fraud detection systems in the financial domain.

CHAPTER 3

PROPOSED METHODOLOGY

3.1 Exploratory Data Analysis (EDA)

In our study, we meticulously examined two datasets sourced from Kaggle, both of which are instrumental in unraveling the complexities of credit card fraud detection. The first dataset, the European credit card fraud detection dataset, represents a vast repository of transactional data, encompassing a staggering 284,807 instances. These transactions, spanning the month of September 2013, provide a rich tapestry of consumer payment behaviors across Europe. However, what's particularly striking is the stark class imbalance within this dataset. Out of this extensive pool of transactions, a mere 492 instances - a mere 0.17% - are flagged as fraudulent. This pronounced class imbalance poses a formidable challenge for machine learning algorithms, potentially impeding their ability to discern meaningful patterns amidst such skewed data distributions. To mitigate this challenge, preprocessing steps have been undertaken, including the removal of non-essential features and the application of Principal Component Analysis (PCA) to reduce dimensionality while retaining critical information.

Turning our attention to the second dataset, the German credit card fraudulent dataset, we encounter a more modest yet equally informative collection of transactional records. Comprising 1,000 transactions in total, this dataset delineates between normal transactions and fraudulent ones, with 700 instances classified as normal and 300 flagged as fraudulent. Here, the fraud rate registers slightly higher at 0.42%, yet the underlying challenges remain akin to those observed in the European dataset. Categorical and numerical features intermingle within this dataset, necessitating rigorous preprocessing efforts to render it amenable to subsequent analysis.

Grasping the intricacies of these datasets is paramount for the development of effective fraud detection models. By delving into the distributions of features, uncovering latent patterns, and addressing class imbalances, researchers can chart a course towards the creation of machine learning algorithms endowed with the capacity to discern between legitimate and fraudulent transactions with heightened accuracy. Thus, this foundational understanding serves as a springboard for the construction of robust fraud detection frameworks capable of safeguarding financial systems and bolstering consumer trust in the digital realm.

3.2 Data Preprocessing

Data preprocessing serves as the cornerstone of any analytical endeavor, acting as the gateway to transforming raw data into a format conducive to meaningful analysis. Within the context of the dataset under examination, which features a heterogeneous mix of categorical and numerical attributes, label encoding emerges as a pragmatic technique for achieving uniformity. This process entails translating all categorical features into a standardized numerical representation, wherein normal transactions are denoted by '0' and fraudulent ones by '1'. By effecting this encoding, the data is rendered more homogenous, thereby simplifying the processing task for subsequent machine learning algorithms.

A meticulous inspection of the European dataset revealed the presence of 1,081 instances characterized by duplicate transaction records. To uphold the integrity and fairness of predictive outcomes, a judicious decision was made to systematically expunge these duplicates from the dataset. Additionally, the introduction of data scaling served to redress any instances marred by null or missing values. This fastidious endeavor not only bolsters data integrity but also underpins the credibility of subsequent analysis outcomes.

Normalization emerges as another linchpin in the preprocessing pipeline, entailing the standardization of data points to a common scale. By effecting this normalization, the dataset is imbued with a newfound consistency, a pivotal prerequisite for facilitating meaningful and accurate analysis. The normalization process bestows a semblance of uniformity upon the dataset, thereby enhancing the quality of insights gleaned and easing the comparability and interpretation of different features.

In sum, these preprocessing steps constitute a veritable crucible for refining the dataset and readying it for analysis. By deftly navigating the intricacies of data preprocessing, inconsistencies are eradicated, missing data are handled with aplomb, and the dataset is meticulously tailored to meet the exigencies of modeling. This assiduous preparation lays a robust foundation for subsequent machine learning endeavors and data analysis pursuits, ultimately culminating in the generation of more reliable insights and predictions.

3.3 Data Balancing

Both datasets, the European credit card fraud detection dataset and the German credit card fraudulent dataset, exhibit a notable class imbalance, with the vast majority of transactions categorized as normal occurrences. This incongruity in class distribution poses a significant challenge, as it can skew predictive models towards favoring the majority class and potentially overlooking minority class instances of fraudulent transactions. To rectify this imbalance and ensure fair and effective training of predictive models, it becomes imperative to implement strategies that address this class disproportionality.

One commonly employed approach to combat class imbalance involves the utilization of undersampling and oversampling techniques, or a combination thereof, to equalize class proportions within the dataset. Undersampling techniques involve reducing the number of instances in the majority class, while oversampling techniques aim to increase the representation of the minority class. By rebalancing the class distribution, these techniques aim to create a more equitable playing field for model training.

In the realm of undersampling, two distinct techniques are brought to bear: the Random Undersampler and the AllKNN technique. The Random Undersampler operates by randomly selecting instances from the majority class and discarding them, thereby reducing its scale and bringing it into closer alignment with the minority class. Conversely, the AllKNN technique adopts a more nuanced approach, identifying instances from the majority class that are in close proximity to minority class instances and subsequently removing them. This targeted approach aids in further rebalancing the dataset while preserving the integrity of the data.

In contrast, oversampling techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE), focus on augmenting the representation of the minority class. SMOTE achieves this by generating synthetic instances for the minority class through interpolation between existing minority class instances. This synthetic augmentation effectively increases the scale of the minority class, thereby mitigating the effects of class imbalance.

Additionally, a hybrid approach known as SMOTE-ENN is employed, which combines the strengths of both undersampling and oversampling techniques. Initially, SMOTE is utilized to oversample the minority class, bolstering its representation. Subsequently, undersampling is performed using the Edited Nearest Neighbors (ENN) technique, which selectively removes noisy samples from both classes. This hybridization aims to strike a balance between rectifying class imbalance and preserving the integrity of the dataset.

By leveraging these comprehensive methodologies, the distribution of the dataset is systematically restructured to alleviate the impacts of class imbalance. This ensures that predictive models are trained on a more representative and balanced dataset, leading to enhanced performance and more reliable predictions. Ultimately, the overarching goal of these methodologies is to fortify the model training process, enabling accurate learning of both minority and majority classes and laying the groundwork for more resilient and effective model outcomes.

3.4 Machine Learning Models

In this study, we delve into a thorough and exhaustive examination of two prominent anomaly detection algorithms: Local Outlier Factor (LOF) and Isolation Forest (IF). Anomaly detection is a critical facet within data analysis, dedicated to uncovering irregular or unforeseen events within datasets. Its significance spans various domains, including cybersecurity, fraud detection, industrial quality control, and health monitoring, where pinpointing anomalies swiftly and accurately is paramount for decision-making and risk mitigation.

The comparison between LOF and IF encompasses various aspects, including their theoretical foundations, computational complexities, sensitivity to parameter settings, performance under different data distributions, scalability to large datasets, and robustness against different types of anomalies. Understanding the strengths and limitations of each algorithm is essential for practitioners to make informed choices based on the specific characteristics of their datasets and the requirements of their anomaly detection tasks.

By conducting a comprehensive comparison between LOF and IF, we aim to provide valuable insights into their respective effectiveness, applicability, and performance across diverse real-world scenarios. Through empirical evaluations and rigorous analyses, this study seeks to contribute to the advancement of anomaly detection methodologies, facilitating better decision-making and risk management in various domains.

3.4.1 Isolation Forest

The Isolation Forest (IF) algorithm stands out prominently in the realm of unsupervised anomaly detection, offering a novel approach that diverges from conventional techniques. Its underlying principle revolves around the concept of isolating individual data instances, thereby providing a distinct perspective on anomaly detection.

What distinguishes IF from traditional methods is its departure from the reliance on distance and density metrics. While classic anomaly detection algorithms often rely heavily on measures such as distance from the centroid or local density estimation, IF introduces a paradigm shift by focusing on isolation. By isolating instances through the construction of random decision trees, IF effectively identifies anomalies based on their ease of separation from the majority of data points.

A notable advantage of IF lies in its resource efficiency. The algorithm is designed to operate with minimal memory allocation, making it well-suited for deployment in various computational environments, including systems with limited resources or high-throughput data processing pipelines. Furthermore, its computational overhead is remarkably low, owing to its intrinsic linear time complexity. This characteristic ensures efficient processing even with exceptionally large datasets, a feature particularly crucial in contemporary big data analytics scenarios.

Beyond its efficiency, IF offers a fresh perspective on anomaly detection, providing nuanced insights into the nature of anomalies within the dataset. By focusing on isolation rather than traditional metrics, such as distance or density, IF achieves improved accuracy in identifying outliers, especially in datasets characterized by complex structures or high dimensionality.

Moreover, the simplicity and interpretability of the Isolation Forest algorithm contribute to its appeal. Its intuitive approach, coupled with transparent decision-making processes inherent in decision tree-based methods, facilitates the understanding and interpretation of anomaly detection results. This aspect is particularly valuable in domains where actionable insights from anomaly detection are crucial for decision-making processes.

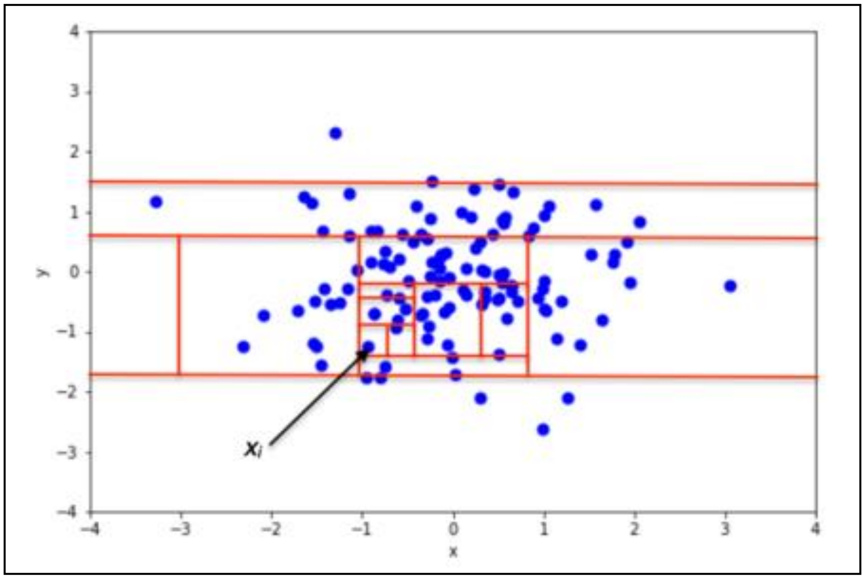


Fig 3.1 An example of isolating a non-anomalous point in a 2D Gaussian distribution

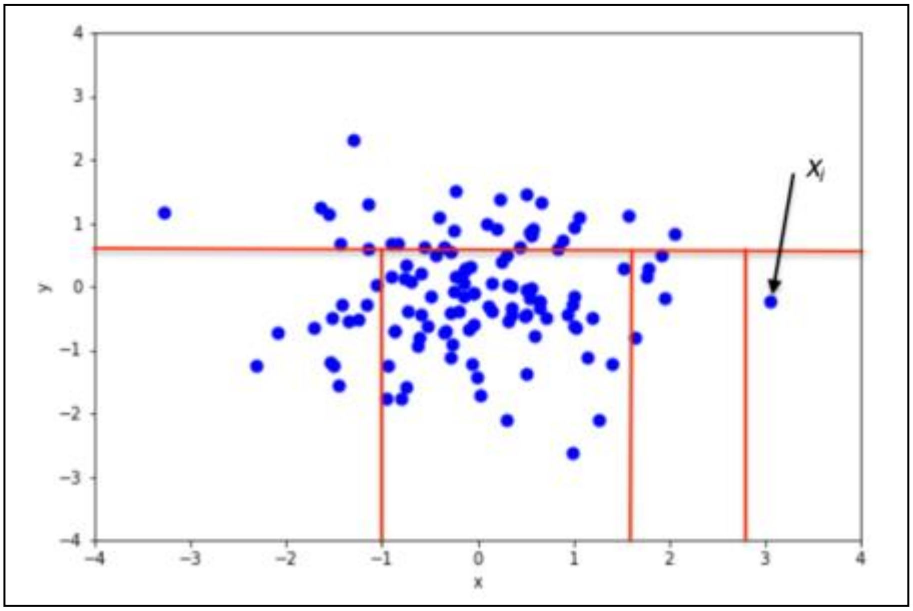


Fig 3.2 An example of isolating an anomalous point in a 2D Gaussian distribution

In summary, the Isolation Forest algorithm presents a promising solution for outlier detection tasks, offering a compelling blend of effectiveness and efficiency. Its unique approach, characterized by isolation-based anomaly identification and resource-friendly nature, renders it a valuable tool for a wide array of data analysis applications. As data volumes continue to escalate and computational resources become increasingly constrained, IF's ability to provide accurate anomaly detection with minimal resource requirements positions it as a cornerstone in the arsenal of modern anomaly detection techniques.

3.4.2 Local Outlier Factor

The Local Outlier Factor (LOF) algorithm represents a significant advancement in the realm of outlier detection, particularly distinguished by its nuanced approach to assessing anomalies within datasets. At its core, LOF operates by scrutinizing the local variances of individual data points concerning their neighboring points. This methodology deviates from conventional outlier detection techniques, which often rely on global statistical measures or distance-based metrics.

LOF's key innovation lies in its emphasis on local density estimation. Rather than treating the dataset as a homogeneous entity, LOF acknowledges the heterogeneous nature of data distributions by evaluating outliers based on the density of their immediate surroundings. By considering the density of neighboring points within a defined radius or neighborhood, LOF captures the intrinsic structure and complexity of the data, thereby enabling more precise anomaly detection.

Let k-distance(A) be the distance of the object A to the k-th nearest neighbor. Note that the set of the k nearest neighbors includes all objects at this distance, which can in the case of a "tie" be more than k objects. We denote the set of k nearest neighbors as N_k(A).

Illustration of the reachability distance. Objects *B* and *C* have the same reachability distance (k=3), while *D* is not a *k* nearest neighbor

- This distance is used to define what is called ***reachability distance***:

$$\text{reachability-distance}_k(A,B)=\max\{k\text{-distance}(B),\ d(A,B)\}$$

- The ***local reachability density*** of an object A is defined by

$$\text{lrd}_k(A):=1/(\Sigma_{B\in N_k(A)}\text{reachability-distance}_k(A,B)/|N_k(A)|)$$

- The local reachability densities are then compared with those of the neighbors using

$$\text{LOF}_k(A):=\Sigma_{B\in N_k(A)}\text{lrd}_k(B)/\text{lrd}_k(A)/|N_k(A)| = \Sigma_{B\in N_k(A)}\text{lrd}_k(B)/|N_k(A)| \cdot \text{lrd}_k(A)$$

A fundamental aspect of LOF is its ability to discern anomalies based on deviations in local density. By calculating a data point's density relative to its neighbors, LOF identifies regions within the dataset with similar densities and flags those with conspicuously lower densities as potential outliers. This localized assessment allows LOF to uncover subtle anomalies that may elude detection by traditional global methods, thus enhancing the algorithm's sensitivity to irregularities within the data.

Moreover, LOF goes beyond merely identifying outliers based on their isolated properties; it evaluates anomalies in the context of their local environment. This contextual understanding enables LOF to distinguish between genuine anomalies and data points that exhibit anomalous behavior due to their proximity to other outliers. Consequently, LOF offers a more nuanced and refined approach to anomaly detection, capable of discerning complex patterns of irregularity within diverse datasets.

Furthermore, LOF's versatility extends to its applicability across various domains and data types. Whether dealing with structured tabular data, high-dimensional feature spaces, or spatial-temporal datasets, LOF remains effective in uncovering anomalies across different contexts. Its adaptability to different data modalities underscores its utility as a versatile tool for anomaly detection in real-world scenarios.

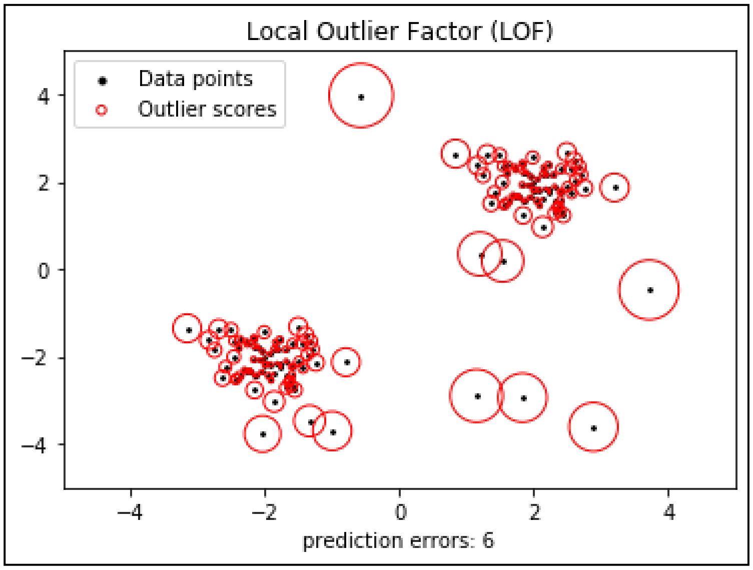


Fig 3.3 Plot of LOF on a data set

In summary, the Local Outlier Factor algorithm represents a sophisticated and powerful tool for outlier detection, leveraging local density information to uncover anomalies with exceptional precision. By prioritizing local relationships and contextual understanding, LOF offers a comprehensive and insightful perspective on the irregularities present within datasets. Its ability to capture subtle deviations and discern complex patterns of anomalous behavior underscores its significance in modern data analysis and anomaly detection tasks.

3.5 Resampling Techniques

In the pursuit of mitigating the adverse effects of class imbalance, our study incorporated four resampling techniques: Random Undersampling, AllKNN, SMOTE (Synthetic Minority Over-sampling Technique), and SMOTE-ENN (SMOTE combined with Edited Nearest Neighbors). These techniques were strategically chosen to address the disparity in class distribution within the datasets and to bolster the model's capacity to discern patterns from both the majority and minority classes.

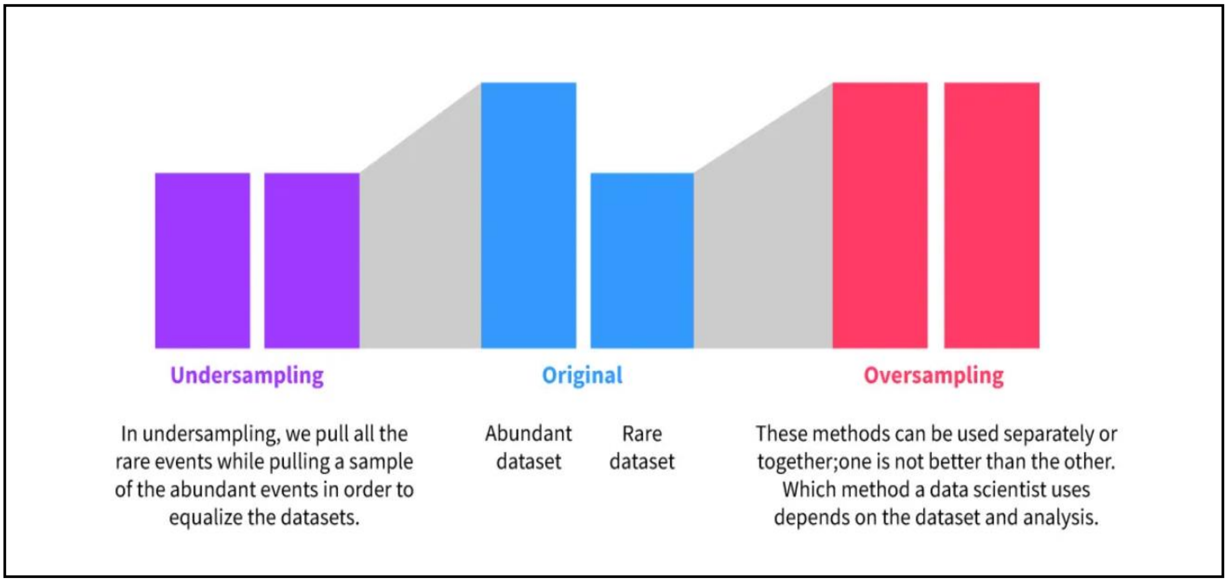

Undersampling
Original
Oversampling
In undersampling, we pull all the rare events while pulling a sample of the abundant events in order to equalize the datasets.
Abundant dataset
Rare dataset
These methods can be used separately or together;one is not better than the other. Which method a data scientist uses depends on the dataset and analysis.

Fig 3.4 Pictorial representation of undersampling and oversampling

3.5.1 Random UnderSampling

Random undersampling is a data preprocessing technique employed to rectify class imbalance within datasets, particularly prevalent in binary classification tasks. In scenarios where one class (typically the minority class) is notably underrepresented compared to the other, random undersampling aims to rectify this imbalance by randomly removing instances from the majority class until a more equitable distribution between the classes is attained. By reducing the prevalence of the majority class, random undersampling seeks to mitigate the bias towards the dominant class, thereby promoting a fairer representation of both classes in the dataset. This technique is instrumental in enhancing the performance of machine learning models, ensuring that classifiers trained on balanced datasets can effectively capture patterns and make accurate predictions, even in the presence of imbalanced classes.

3.5.2 AllKNN Undersampling

AllKNN undersampling is a sophisticated resampling technique designed to address class imbalance within datasets. Unlike random undersampling, which removes instances from the majority class indiscriminately, AllKNN undersampling employs the k-nearest neighbors (KNN) algorithm to selectively identify and eliminate instances from the majority class.

This method operates by examining each instance in the majority class and identifying its k-nearest neighbors within the same class. By assessing the local density and distribution of instances, AllKNN identifies outliers or instances situated in densely populated regions. These instances are considered potentially redundant or less informative for model training, and thus are candidates for removal.

The selection of the appropriate value for k (the number of nearest neighbors) is crucial in determining the effectiveness of AllKNN undersampling. A higher value of k may lead to a more conservative approach, retaining more instances from the majority class, while a lower value of k may result in a more aggressive undersampling strategy.

One of the advantages of AllKNN undersampling is its ability to selectively remove instances based on their local context, thereby reducing the risk of information loss and preserving the intrinsic structure of the dataset. By focusing on instances that deviate from their local neighborhoods, AllKNN undersampling aims to rebalance the class distribution while minimizing the impact on the overall dataset characteristics.

Furthermore, AllKNN undersampling can be particularly effective in scenarios where the class imbalance is severe or where the majority class contains clusters of outliers or noisy instances. By targeting these outliers and reducing their influence on the training process, AllKNN helps to improve the generalization ability of machine learning models, leading to more robust and reliable predictions on unseen data.

Overall, AllKNN undersampling represents a nuanced and data-driven approach to addressing class imbalance, leveraging the power of the KNN algorithm to selectively remove instances from the majority class while preserving the integrity and representativeness of the dataset.

Through careful consideration of local density and distribution, AllKNN offers a valuable tool for enhancing the performance of machine learning models in imbalanced classification tasks.

3.5.3 Synthetic Minority Over-sampling TEchnique (SMOTE)

SMOTE, or Synthetic Minority Over-sampling Technique, is a popular resampling method designed to address class imbalance in datasets, particularly prevalent in binary classification tasks where one class is significantly underrepresented compared to the other. Unlike undersampling techniques that remove instances from the majority class, SMOTE focuses on the minority class by generating synthetic instances to augment its representation.

By synthesizing new instances, SMOTE effectively expands the minority class, thereby achieving a more balanced distribution between classes. This is accomplished by interpolating between existing minority class instances, creating synthetic data points along the line segments connecting pairs of nearest neighbors. The placement of synthetic instances is guided by the underlying distribution of the minority class, ensuring that the generated samples are representative of the class characteristics.

One of the key advantages of SMOTE is its ability to address class imbalance without discarding potentially valuable information from the majority class. By generating synthetic instances rather than removing existing ones, SMOTE preserves the integrity and richness of the dataset, thereby enhancing the robustness and generalization ability of machine learning models trained on the resampled data.

SMOTE's effectiveness in rebalancing class distributions and improving model performance has made it a widely adopted technique in the field of imbalanced classification. However, it is important to note that SMOTE may not always be suitable for every dataset, particularly in cases where the minority class is already well-represented or when the distribution of data points is highly complex. Additionally, careful consideration should be given to the selection of parameters, such as the number of synthetic samples to generate and the method used to determine nearest neighbors, to ensure optimal performance and avoid potential pitfalls such as overfitting.

3.5.4 Synthetic Minority Oversampling Technique - Edited Nearest Neighbor (SMOTE-ENN)

SMOTE-ENN, or Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors, is a resampling technique specifically designed to address class imbalance in datasets. It integrates the strengths of SMOTE, which generates synthetic instances to augment the minority class, with the refinement provided by Edited Nearest Neighbors (ENN), which selectively removes noisy or misclassified instances.

In SMOTE-ENN, the SMOTE algorithm is first applied to the minority class to generate synthetic instances, thereby increasing its representation in the dataset. This helps to rebalance the class distribution and alleviate the issues associated with class imbalance. Subsequently, the edited nearest neighbors algorithm is employed to identify and remove instances that are considered outliers or noisy based on their proximity to neighboring data points.

By combining oversampling and undersampling techniques, SMOTE-ENN aims to enhance the quality of the resampled dataset while preserving its structural integrity. This iterative process iteratively refines the dataset by adding synthetic instances to the minority class and then removing potentially problematic instances, leading to a more balanced and representative dataset for model training.

One of the key advantages of SMOTE-ENN is its ability to effectively address class imbalance while also mitigating the potential drawbacks associated with oversampling techniques, such as the generation of synthetic instances that may introduce noise or reduce model performance. By iteratively refining the dataset, SMOTE-ENN helps to improve the robustness and generalization ability of machine learning models trained on imbalanced data, leading to more reliable and accurate predictions. However, like any resampling technique, careful parameter tuning and validation are essential to ensure optimal performance and avoid potential pitfalls such as overfitting.

3.6 Implementation

In the meticulous process of developing and refining machine learning models, the division of data into distinct subsets - namely, the training, validation, and testing sets - plays a pivotal role. This segmentation ensures a systematic approach to model development, hyperparameter tuning, and performance evaluation while guarding against overfitting and ensuring generalization to unseen data.

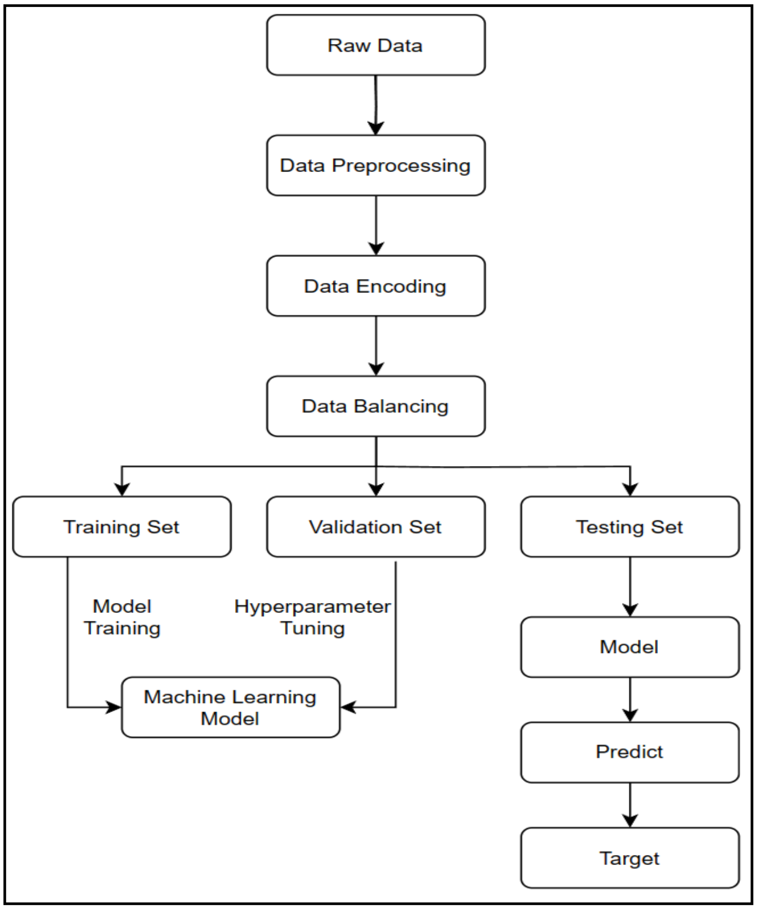


Fig 3.5 Block Diagram of the Proposed Model

Initially, the classifier undergoes training on the designated training dataset, where it learns the underlying patterns and structures present in the data. Following this, the model's hyperparameters are fine-tuned using the validation dataset, optimizing its performance without direct exposure to the testing data. This separation is critical to prevent any bias or overfitting that may arise from tuning the model based on the testing set.

The testing dataset serves as the ultimate litmus test for the trained model's performance. Here, the model's efficacy is rigorously evaluated using metrics such as accuracy, precision, recall, and F1-score, providing insights into its ability to make accurate predictions on unseen data. Additionally, the confusion matrix, encompassing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), offers a granular understanding of the model's predictive performance across different classes or categories.

In the realm of unsupervised learning, where labels are absent, model training relies solely on the available variables within the dataset. The objective shifts towards uncovering inherent patterns, anomalies, or deviations within the data. Evaluation metrics such as the detection of outliers, precision, recall, and F1-score remain instrumental in assessing the efficacy of unsupervised learning models, providing insights into their ability to discern meaningful patterns amidst complex data distributions.

Our research project encompasses two distinct datasets - the European and German credit card fraud transaction datasets. Initially, we evaluate the performance of models without altering the class distribution, providing a baseline for comparison. Subsequently, we employ resampling techniques, including random undersampling, AllKNN, SMOTE, and SMOTE-ENN, on both datasets. Each resampling technique is meticulously applied, and the resulting models are evaluated using comprehensive metrics to gauge the impact on model accuracy and effectiveness.

By systematically analyzing the effects of various resampling techniques on model performance, our research aims to elucidate the efficacy of these techniques in addressing imbalanced class distributions and enhancing the robustness of machine learning models in fraud detection tasks. Through rigorous experimentation and analysis, we endeavor to contribute valuable insights to the field of anomaly detection and imbalanced data classification, facilitating more reliable and effective fraud detection mechanisms in real-world scenarios.

CHAPTER 4

RESULTS AND DISCUSSION

In the pursuit of identifying the most effective algorithm for detecting fraudulent transactions, a comprehensive evaluation framework was established, encompassing various criteria for comparing algorithms. Among the foremost metrics utilized for this evaluation were accuracy, recall, and precision, which are widely regarded as fundamental indicators of a machine learning model's success. These metrics were computed using a Confusion Matrix, a structured approach to assessing model performance that provides insights into true positives, true negatives, false positives, and false negatives.

To ascertain the efficacy of different algorithms, testing was conducted using both the original dataset and a resampled dataset. This dual approach allowed for a thorough examination of algorithm performance under different conditions. Notably, the findings underscored the pivotal role of resampling techniques in influencing model performance, with resampling demonstrating a significant impact on algorithm effectiveness across various scenarios.

In the context of the imbalanced European and German credit card datasets, two prominent anomaly detection models, Isolation Forest (IF) and Local Outlier Factor (LOF), were deployed. The primary objective was to enhance the performance of these models, particularly by reducing false negatives and improving overall accuracy. To address the challenge posed by imbalanced data, a spectrum of resampling techniques was explored, aimed at rebalancing class distributions and augmenting the representation of minority classes.

Through rigorous experimentation and analysis, the study aimed to not only identify the most effective anomaly detection algorithm but also to elucidate the role of resampling techniques in enhancing model performance. By leveraging insights gained from these evaluations, the research sought to contribute valuable insights to the domain of fraud detection and anomaly detection, ultimately facilitating the development of more robust and reliable detection mechanisms in real-world applications.

Table 4.1 Model performance on European Dataset without resampling

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	No. of Outliers
IF	99.81	63	53	54	107
LOF	99.81	50	50	50	104

Table 4.2 Model performance on German Dataset without resampling

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	No. of Outliers
IF	70.50	35	50	41	59
LOF	70.50	35	50	41	59

To mitigate class imbalance, four resampling techniques – Random Undersampling, AllKNN, SMOTE, and SMOTE-ENN – were employed. These methods sought to rectify the imbalance in class distribution, thereby improving the models' ability to detect patterns from both majority and minority classes. The performance of Isolation Forest (IF) and Local Outlier Factor (LOF) models was evaluated using these resampling strategies on the European and German datasets. Results are presented in Tables 4.3, 4.4, 4.5, and 4.6, showcasing the impact of each resampling technique on the models' performance metrics. Through this analysis, the study aims to identify the most effective resampling strategy for enhancing the accuracy and reliability of anomaly detection in credit card fraud detection tasks.

Table 4.3 IF Model performance on European Dataset with various resampling techniques

Resampling Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	No. of Outliers
Random Undersampling	98.98	49	50	50	101
AllKNN	98.80	53	78	55	682
SMOTE	49.76	25	50	33	57130
SMOTE-ENN	48.77	25	49	33	58201

Table 4.4 LOF Model performance on European Dataset with various resampling techniques

Resampling Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	No. of Outliers
Random Undersampling	98.99	49	50	50	100
AllKNN	98.96	50	50	50	591
SMOTE	49.98	67	50	34	56880
SMOTE-ENN	49.53	42	50	34	57341

Table 4.5 IF Model performance on German Dataset with various resampling techniques

Resampling Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	No. of Outliers
Random Undersampling	70.56	35	50	41	53
AllKNN	51.61	48	48	48	45
SMOTE	52.85	47	50	36	132
SMOTE-ENN	43.80	22	50	30	59

Table 4.6 LOF Model performance on German Dataset with various resampling techniques

Resampling Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	No. of Outliers
Random Undersampling	70.60	35	50	41	52
AllKNN	54.83	55	55	54	42
SMOTE	51.42	26	48	34	136
SMOTE-ENN	43.80	47	50	33	59

This research underscores the crucial fusion of cutting-edge algorithms and meticulous data preparation methodologies as a linchpin in the ongoing battle against credit card fraud and the safeguarding of financial transactions. By amalgamating sophisticated algorithms with rigorous data preparation processes, financial institutions can fortify their defenses against increasingly sophisticated fraudulent activities, thereby fostering trust and confidence in the financial ecosystem.

In Table 4.7, a comprehensive comparison of all models utilized on both the European and German datasets, incorporating various resampling techniques, is presented to identify the optimal model. The results highlight performance metrics such as accuracy, precision, recall, and F1-score, aiding in the selection of the most effective approach for credit card fraud detection. Furthermore, Table 4.8 presents a comparative analysis between our proposed methodology and existing research findings, offering insights into the novel contributions and advancements made in the field of fraud detection through our study.

Table 4.7 Best Model on respective dataset

Dataset	Best Model
European	Isolation Forest
German	Local Outlier Factor – Random Undersampling

Table 4.8 Comparative analysis with previous reports on European Dataset

Reference	Model	Accuracy (%)
Ref [11]	Local Outlier Factor	97
Ref [12]	Isolation Forest	98.72
Ref [14]	Isolation Forest	99.72
Proposed	Isolation Forest	99.81

Moreover, the study illuminates the profound positive impact of resampling techniques on the efficacy of fraud detection systems. Notably, resampling interventions resulted in the complete eradication of false negatives (FN), indicative of a significant enhancement in the models' acumen in accurately pinpointing fraudulent transactions. This remarkable achievement not only underscores the prowess of resampling but also highlights its pivotal role in bolstering the precision and reliability of fraud detection mechanisms.

Additionally, the discernible reduction in false positives following resampling interventions underscores the nuanced capability of these techniques to differentiate between genuine and fraudulent transactions. This nuanced refinement minimizes misclassifications on both ends of the spectrum, ensuring that legitimate transactions are not erroneously flagged as fraudulent while effectively identifying and isolating instances of fraudulent activity.

In essence, the synergistic integration of advanced algorithms and meticulous data preparation methodologies, coupled with the strategic deployment of resampling techniques, constitutes a formidable defense against credit card fraud. This holistic approach not only enhances the efficacy of fraud detection systems but also reinforces the resilience of financial institutions in safeguarding the integrity of financial transactions and preserving consumer trust.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

This study delves into the realm of credit card fraud detection by utilizing datasets derived from both German and European credit card transactions. The primary objective is to evaluate the performance of various machine learning classification techniques in accurately identifying fraudulent transactions. The process involves several stages, including data importation, preprocessing, encoding, and model configuration, all geared towards facilitating machine learning model training.

The dataset undergoes meticulous preprocessing to ensure its suitability for model training. This involves tasks such as data cleaning, handling missing values, and encoding categorical variables into numerical representations. Once the dataset is prepared, it is partitioned into training, validation, and testing sets, following best practices in machine learning workflow.

The training phase involves feeding the prepared dataset into different classification models, including but not limited to Isolation Forest (IF) and Local Outlier Factor (LOF). These models are trained on the training data and then fine-tuned using the validation set to optimize their performance.

After training, the models are deployed to make predictions on unseen data from the testing set. Their performance is evaluated using various metrics such as accuracy, precision, recall, and F1-score. These metrics provide insights into how well the models are able to distinguish between fraudulent and legitimate transactions.

In addition to evaluating individual models, the study explores the impact of resampling techniques on model performance. Resampling methods such as random undersampling are applied to address class imbalance issues commonly encountered in fraud detection datasets.
The results of the study indicate that the Isolation Forest model achieves a high accuracy of 99.81% when applied to the European credit card fraudulent transactions dataset. On the other hand, the Local Outlier Factor algorithm performs best on the German credit card fraudulent transaction dataset, achieving an accuracy of 70.60% when combined with random undersampling.

Overall, the study provides valuable insights into the effectiveness of different machine learning models and resampling techniques in detecting credit card fraud, paving the way for more robust fraud detection systems in the financial industry.

5.2 Future Scope

In the pursuit of advancing credit card fraud detection, a multifaceted approach integrating state-of-the-art anomaly detection methods beyond conventional algorithms like Isolation Forest and Local Outlier Factor is imperative. Exploring cutting-edge techniques such as neural networks, autoencoders, and deep learning architectures offers the potential to unlock intricate patterns and relationships within credit card transaction data, thus augmenting the accuracy and sophistication of fraudulent activity identification. These advanced methods leverage the inherent complexity of transactional data, enabling models to discern subtle anomalies and fraudulent patterns that may elude traditional algorithms.

Moreover, the incorporation of ensemble learning methodologies presents a compelling avenue for enhancing fraud detection efficacy. By combining the strengths of diverse models, ensemble techniques create robust and resilient detection frameworks capable of capturing a broader spectrum of fraudulent behaviors and adapting to evolving fraud tactics.

Key Points to Consider:

- **Real-Time Data Streaming and Continuous Monitoring:** Developing mechanisms for real-time data ingestion and continuous monitoring allows for proactive fraud detection and swift response to emerging threats. By leveraging streaming analytics, financial institutions can detect suspicious activity in near-real-time, enabling timely intervention to mitigate potential losses.

- **Integration of External Threat Intelligence:** Supplementing internal data with external threat intelligence sources enhances the fraud detection capabilities of models. By leveraging external data feeds, such as fraud consortiums and threat intelligence platforms, financial institutions can stay ahead of emerging fraud trends and sophisticated attack vectors.

- **Dynamic Feature Engineering:** Adopting dynamic feature engineering techniques enables models to adaptively modify and optimize feature sets based on evolving transaction dynamics. By continuously refining feature selection and engineering processes, models can maintain relevance and efficacy in detecting evolving fraud patterns.

- **Ethical Implications and Regulatory Compliance:** Addressing ethical considerations and ensuring compliance with regulatory mandates is paramount in the deployment of machine learning models for fraud detection. Safeguarding data privacy and security, and adhering to regulations such as GDPR and PCI-DSS, is essential to maintaining trust and integrity in financial systems.

- **Cross-Domain Collaboration and Knowledge Sharing:** Collaboration across industry sectors and knowledge sharing initiatives can enhance the collective resilience against fraud. By exchanging insights and best practices, financial institutions can leverage collective intelligence to develop more robust fraud detection mechanisms.

- **Continuous Model Evaluation and Improvement:** Establishing processes for continuous model evaluation and improvement ensures that fraud detection systems remain effective and adaptive. By leveraging feedback loops and performance metrics, models can be iteratively refined to address emerging threats and evolving fraud tactics.

By integrating these advanced methodologies and considerations into credit card fraud detection systems, financial institutions can bolster their defenses against increasingly sophisticated fraud schemes, safeguarding the integrity of financial transactions and preserving consumer trust in the digital economy.

REFERENCES

[1] Varmedja, Dejan, Mirjana Karanovic, Srdjan Sladojevic, Marko Ar- senovic, and Andras Anderla. “Credit card fraud detection-machine learning methods.” In 2019 18th International Symposium INFOTEH- JAHORINA (INFOTEH), pp. 1-5. IEEE, 2019.

[2] Asha, R. B., and Suresh Kumar KR. “Credit card fraud detection using artificial neural network.” Global Transitions Proceedings 2, no. 1 (2021): 35-41.

[3] Sailusha, Ruttala, V. Gnaneswar, R. Ramesh, and G. Ramakoteswara Rao. “Credit card fraud detection using machine learning.” In 2020 4th international conference on intelligent computing and control systems (ICICCS), pp. 1264-1270. IEEE, 2020.

[4] Chugh, Bharti, and Nitin Malik. “Machine Learning Classifiers for De- tecting Credit Card Fraudulent Transactions.” In Information and Com- munication Technology for Competitive Strategies (ICTCS 2021) ICT: Applications and Social Interfaces, pp. 223-231. Singapore: Springer Nature Singapore, 2022.

[5] Ileberi, Emmanuel, Yanxia Sun, and Zenghui Wang. “Performance evaluation of machine learning methods for credit card fraud detection using SMOTE and AdaBoost.” IEEE Access 9 (2021): 165286-165294.

[6] Mienye, Ibomoiye Domor, and Yanxia Sun, “A Deep Learning Ensemble With Data Resampling for Credit Card Fraud Detection.” IEEE Access 11 (2023): 30628-30638.

[7] Bagga, Siddhant, Anish Goyal, Namita Gupta, and Arvind Goyal. “Credit card fraud detection using pipeling and ensemble learning.” Procedia Computer Science 173 (2020): 104-112.

[8] Chen, Joy Iong-Zong, and Kong-Long Lai. “Deep convolution neural network model for credit-card fraud detection and alert.” Journal of Artificial Intelligence and Capsule Networks 3, no. 2 (2021): 101-112.

[9] Shirodkar, Nikita, Pratikesh Mandrekar, Rohit Shet Mandrekar, Rahul Sakhalkar, KM Chaman Kumar, and Shailendra Aswale. “Credit card fraud detection techniques–A survey.” In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic- ETITE), pp. 1-7. IEEE, 2020.

[10] Esenogho, Ebenezer, Ibomoiye Domor Mienye, Theo G.Swart, Kehinde Aruleba, and George Obaido. “A neural network ensemble with feature engineering for improved credit card fraud detection.” IEEE Access 10 (2022): 16400-16407.

[11] John, Hyder, and Sameena Naaz, “Credit card fraud detection using local outlier factor and isolation forest.” Int. J. Comput. Sci. Eng 7, no. 4 (2019): 1060-1064.

[12] Gupta, Swati, Sanjay Patel, Surender Kumar, and Goldi Chauhan, “Anomaly detection in credit card transactions using machine learning.” (2020).

[13] Kittidachanan, Kittikun, Watha Minsan, Donlapark Pornnopparath, and Phimphaka Taninpong, “Anomaly detection based on GS-OCSVM classification” In 2020 12th International Conference on Knowledge and Smart Technology (KST), pp. 64-69. IEEE, 2020.

[14] Vijayakumar, V., Nallam Sri Divya, P. Sarojini, and K.Sonika. “Isolation forest and local outlier factor for credit card fraud detection system.” International Journal of Engineering and Advanced Technology (IJEAT) 9 (2020): 261-265.

[15] Rajeev, Haritha, and Uma Devi, “Detection of credit card fraud using isolation forest algorithm.” In Pervasive Computing and Social Network- ing: Proceedings of ICPCSN 2021, pp. 23-34. Springer Singapore, 2022.

[16] Itoo, Fayaz, Meenakshi, and Satwinder Singh, “Comparison and analysis of logistic regression, Na ̈ive Bayes and KNN machine learning algorithms for credit card fraud detection.” International Journal of Information Technology 13 (2021): 1503-1511.

[17] Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua, “Isolation Forest” Data Mining, 2008. ICDM’08. Eighth IEEE International Conference on

[18] Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua, “Isolation-based anomaly detection.” ACM Transactions on Knowledge Discovery from Data (TKDD) 6.1 (2012): 3

[19] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In ACM sigmod record.

APPENDIX 1

Anomaly Detection Classifiers for Detecting Credit Card Fraudulent Transactions

Prerna Singh
Department of Computer Science and Engineering
KIET Group of Institutions, Delhi-NCR
Ghaziabad-Meerut Road, Ghaziabad-201206, India
prernasingh51002@gmail.com

Khyati Singla
Department of Computer Science and Engineering
KIET Group of Institutions, Delhi-NCR
Ghaziabad-Meerut Road, Ghaziabad-201206, India
khyatisingla1805@gmail.com

Prince Piyush
Department of Computer Science and Engineering
KIET Group of Institutions, Delhi-NCR
Ghaziabad-Meerut Road, Ghaziabad-201206, India
prince.piyush.2019@gmail.com

Bharti Chugh
Department of Computer Science and Engineering
KIET Group of Institutions, Delhi-NCR
Ghaziabad-Meerut Road, Ghaziabad-201206, India
bharti.cse@kiet.edu

***Abstract*—The internet and e-commerce have grown quickly, which has increased credit card use but also, regrettably, increased credit card fraud. To address this, Anomaly Detection has emerged as a crucial method for identifying unusual events and data in datasets. It uses advanced algorithms to detect deviations from normal patterns, helping authorities proactively combat fraudulent activities. While digital advancements offer convenience, they also expose vulnerabilities. Anomaly Detection offers a modern defense, safeguarding financial systems by early spotting of anomalies. In this study, we employed two algorithms - the Isolation Forest (IF) and the Local Outlier Factor (LOF) for identification of anomalies. To improve the performance of these models, we also employed a variety of resampling strategies. Specifically, we used techniques like Random Undersampling, AllKNN, Synthetic Minority Oversampling Technique (SMOTE), and Synthetic Minority Oversampling Technique - Edited Nearest Neighbor (SMOTE-ENN) to balance the European Credit Card Fraudulent transactions dataset and the German Credit Card fraud dataset. Out of the different configurations, the Isolation Forest classifier demonstrated the highest accuracy, reaching 99.81%, when applied to the initially imbalanced European credit card fraudulence dataset. On the other hand, the German credit card dataset achieved a remarkable accuracy of 70.60% through the implementation of the LOF classifier, coupled with the Random Undersampling technique to address its imbalanced nature.**

***Index Terms*—Anomaly Detection, IF, LOF, Credit Card Fraud, Random Undersampling, AllKNN, SMOTE, SMOTE-ENN**

I. Introduction

Hackers and fraudsters have targeted credit cards for years, and this is unlikely to change anytime soon. Unfortunately, as technology advanced, so did the most prevalent scams, so con artists are constantly coming up with new ruses to con unsuspecting victims. In recent years, there has been a massive increment in the use of online banking, high-tech payments, and buying using debit/credit cards as a result of the proliferation of businesses, online services, and internet users. While this offers convenience and efficiency, it has also given rise to financial frauds like money laundering, card theft, and identity theft. Despite benefits such as cashless transactions and time saving, these frauds pose serious risks modern tech advancements have fueled more sophisticated fraud techniques, causing substantial losses to businesses and users.

Fraud detection systems employing data mining and machine learning have been developed, but challenges remain. There exist two distinct categories of credit card fraudulent activities. The first involves the unlawful acquisition of the physical credit card itself, while the second pertains to the illicit procurement of sensitive card-related information, encompassing the card number, CVV code, card type, and related data. Through the unlawful acquisition of credit card information, an unauthorized person can potentially gain access to substantial financial resources or execute significant transactions prior to the cardholder’s awareness of such activities. Fraudulent activities include unauthorized transactions, credit card theft, and more, affecting customer trust and business stability credit card fraud, both through application and behavior, is a major concern, leading to billions in losses.

In 2017, global credit card fraud losses reached $22.8 billion, and rise to $31 billion in 2020. Detection of these frauds is crucial due to their impact on financial transactions and institutions. Sixty-five percent of cardholders, up from 58 percent the year before, have experienced fraud of some kind. As the second-largest category for identity theft fraud, the Federal Trade Commission claims that it received roughly 390,000 allegations of credit card identity theft fraud in 2021. By new credit card accounts in 2020, identity theft climbed by 48%. With reference to Nilson report, credit card theft is increasing and is expected to touch an astounding figure of

$38.5 billion by 2027.

Credit card fraud detection has been aided by the use of artificial intelligence, data mining and machine learning techniques, among others. Unfortunately, these initiatives have not produced any noteworthy results. One major obstacle to using ML methodologies for credit card fraud detection is the personal and confidential nature of that data. Therefore, datasets with anonymized properties are used in the development of ML models for credit card fraud detection. Therefore, identifying fraudulent transaction is very tough because new patterns, deviations and features emerge with each fraudulent transaction.

A. *Motivation*

The increasing prevalence of credit card fraud has prompted a compelling need for robust fraud detection mechanisms. As financial transactions shift towards digital platforms, the vulnerability to fraudulent activities escalates. The motivation behind this endeavor lies in preventing substantial financial losses, ensuring customer trust, and upholding the integrity of electronic payment systems. By working in this domain seeks to address this critical issue, safeguarding both consumers and financial institutions. By harnessing state-of-the-art technical automation such as ML, artificial intelligence and data analytics, the aim is to create proactive and accurate systems that identify and mitigate fraudulent transactions, ultimately fostering a more secure and resilient financial landscape.

B. *Literature Review*

The substantial financial losses incurred as a result of fraudulent activities have spurred researchers to seek a remedy capable of pre-emptively identifying and thwarting such illicit actions. Numerous approaches have been posited and subjected to testing in pursuit of this objective.

Ref [1] used the SMOTE technique to balance the dataset. They then ran the Multilayer Perceptron (MP), Random Forest (RF), Naive Bayes (NB), and Logistic Regression (LR) algorithms. Their research showed that the Random Forest algorithm produced the best outcomes. Ref [2] conducted an investigation involving the utilization of SVM, KNN and Artificial Neural Network (ANN) techniques to forecast instances of fraud. A comparative analysis was conducted between ML algorithms and ANN. The results indicated that the implementation of an ANN yielded an accuracy level approaching 100%. Ref [3] employed the RF and AdaBoost algorithms, utilizing the confusion matrix to generate ROC curves. The outcomes derived from both algorithms yielded congruent conclusions. When evaluating precision, recall, and the F1-score, RF algorithm exhibited the highest values.

Ref [4] executed GridSearchCV and random search for hyperparameter optimization. The Decision Tree classifier achieved 72.1% accuracy on the imbalanced German credit card dataset. LDA reached an impressive 98.6% accuracy on the imbalanced European credit card fraud dataset. Additionally looked into were SMOTE, Naive Bayes, and Logistic Regression. As a result, the Decision Tree model performs better than the LR, LDA, and Naive Bayes algorithms. Ref [5] crafted an ML-based framework fusing LR, SVM, RF, NB, XGBoost, and ET algorithms with AdaBoost. The amalgamation aimed to heighten classification efficacy, evaluated via metrics like accuracy, recall, precision, MCC, and AUC. Ref [6] introduces a resilient deep-learning methodology incorporating LSTM and GRU neural networks as foundational classifiers within a stacking ensemble structure and a MLP functions as the higher-level learner. To address class distribution imbalance, the SMOTE-ENN technique is utilized in the dataset.

Ref [7] conducts a performance assessment of the effectiveness of 9 distinct classifier models on credit card fraud detection data. The models are Logistic Regression, K-Nearest Neighbors, Random Forest, Naive Bayes, Multi Layer Perceptron, AdaBoost, quadrant discriminative analysis, ensemble learning and pipelining. To rectify dataset imbalance, the ADASYN technique is used. In conclusion, Pipelining method demonstrated superior performance across diverse metrics. Ref [8] introduced a financial fraudulence detection scheme utilizing a Deep Convolution Neural Network (DCNN) and algorithms of deep learning to detect fraudulent transactions. Within a 45-second timeframe, an accuracy rate of 99% in detection was achieved. Ref [9] employed the GA for the purpose of feature selection. The devised detection mechanism incorporated a selection of ML classifiers - DT, RF, LR, ANN, and NB. Notably, the genetic algorithm was incorporated within the RF's fitness function. This amalgamation, termed GA-RF (utilizing version 5), yielded a remarkable overall accuracy of 99.98%. A combination of data resampling strategy and a neural network ensemble classifier are introduced by Ref [10]. LSTM neural network is implemented as the primary trainee to create the ensemble classifier within the AdaBoost framework. Simultaneously, the hybrid resampling strategy incorporates the SMOTE-ENN method. According to their results, using the SMOTE-ENN data resampling method in tandem with the boosted LSTM classifier is a promising strategy for spotting fraudulent acts in financial transactions using credit cards.

Ref [11] compared the LOF and IF algorithms using Python to identify fraudulent transactions, yielding a 97% accuracy for LOF and 76% for IF. Ref [12] implemented the IF model and utilized the Area Under the Precision-Recall curve (AUC), which demonstrated superior outcomes compared to the Area Under the ROC curve. The efficiency of their approach in a fraud detection model was noted as 98.72%. Ref [13] employed the one-class support vector machine classifier (OCSVM) for outlier detection, optimizing hyperparameters γ and ν through grid search based on maximizing the AUC.

The GS-OCSVM exhibited superior fraud detection to the isolation forest, with higher true negative rates for both German and European credit card datasets. Ref [14] utilized Python to apply the IF and LOF algorithms. After evaluating the datasets, they observed that the IF algorithm delivered a higher precision rate compared to the LOF algorithm. Ref [15] discusses the IF algorithm's role in credit card fraud detection and contrasts it with the LOF algorithm. IF outperforms LOF in terms of error, accuracy-score and recall for two algorithms. The rate of fraud detection is about 27% with Isolation Forest, In comparison to LOF's mere 2%. IF model boasts 99.774% precision, surpassing LOF's 99.65%. Ref [16] employed the random under-sampling approach on imbalanced data and applied three machine learning algorithms - LR, NB, KNN. The algorithms efficacy was assessed using metrics encompassing accuracy-score, F-measure, sensitivity, precision, specificity and AUC.

I. PRELIMINARIES

A. EDA (Exploratory Data Analysis)

In this study, we conducted an evaluation on two distinct datasets. The European credit card fraud detection dataset and the German credit card fraudulent dataset, both sourced from Kaggle. These datasets are presented in CSV format (.csv).

The dataset pertaining to European credit card transactions encompasses an extensive count of 284,807 instances, reflecting the payments conducted by consumers across Europe in the month of September 2013. Remarkably, out of this extensive pool, a mere 492 transactions are identified as fraudulent, constituting an exceedingly minute proportion of approximately 0.17%. This stark discrepancy in statistics emphasizes how unbalanced the European credit card fraud dataset is. There are two categories for these transactions: legitimate and fraudulent. The original functionality and other contextual information are removed from the training data to meet security issues. Using only integral attributes as its emphasis, the Principal Component Analysis (PCA) transformation's outcomes are presented. The PCA-derived principal components are denoted by the variables V1, V2, V3,....,V28. Notably, the PCA conversion process has not been applied to the 'Time' and 'Number' properties.

The original German credit card dataset comprises a total of 21 distinct features, with 13 of them being categorical in nature and the remaining 7 being numerical attributes. This dataset encompasses a total of 1000 credit card transactions. Among these transactions, 700 are classified as normal transactions, while the remaining 300 are categorized as fraudulent transactions, resulting in a proportion of 0.42% representing the occurrence of fraudulent activities.

B. Data Preprocessing

Data preprocessing stands as a pivotal initial stage in the journey of refining raw data into a more structured and suitable form. Given the dataset's combination of categorical and numerical attributes, the technique of label encoding emerges as a practical choice to translate all features into a uniform numerical representation. Within this encoding scheme, normal transactions are denoted by '0', whereas fraudulent transactions are signified as '1'. Within the context of the European dataset, it was uncovered that a total of 1081 instances exhibited duplicated transaction records. In order to uphold precision and fairness in predictive outcomes, a strategic decision was taken to eliminate these instances of duplication from the dataset. In conjunction with this, the process of data scaling was introduced, involving the meticulous removal of any instances featuring null or missing values. This diligent effort contributes to data integrity and the credibility of analysis outcomes. Furthermore, a critical step involved the process of normalization, whereby data points are standardized to a common scale. This procedure fosters consistency across the dataset and facilitates a more meaningful and accurate analysis, ultimately enhancing the quality of insights derived from the data.

C. Data Balancing

Both of the datasets originally display a compelling class imbalance, with preponderance of transactions are categorized as normal occurrences. Given this substantial disparity, addressing the imbalance becomes an imperative undertaking to ensure the efficient and unbiased training of predictive models. Commonly employed methodologies to rectify the skewed class distribution encompass strategies such as equalizing the class proportions through techniques like decreasing the scale of the majority class (undersampling), magnifying the scale of the minority class (oversampling), or implementing a combination of both strategies. To address the imbalance, two distinct techniques for undersampling are employed - the Random Undersampler and the AllKNN technique. In contrast, for oversampling, the SMOTE methodology is harnessed to perform data balancing. Furthermore, a hybrid approach termed the SMOTE-ENN is incorporated. This hybrid approach merges the benefits of both undersampling and oversampling, aiming to restructure the dataset's distribution.

The overarching goal of these methodologies is to redress the inherent class imbalance, thereby fortifying the model training process by enabling accurate learning of both the minority and majority classes. This comprehensive approach lays the foundation for more resilient and effective model outcomes.

I. Machine Learning Models

In the scope of this study, we embark on a comprehensive comparison between two distinct algorithms - the LOF and the IF. These algorithms hold a pivotal role in the domain of anomaly detection, a field aimed at identifying unusual or unexpected occurrences within datasets.

A. *Isolation Forest*

The Isolation Forest algorithm, in particular, stands out as an unsupervised methodology meticulously designed for the explicit purpose of outlier detection. Its core principle revolves around the concept of ”isolation”, a procedure where an individual data instance is segregated from the rest of the dataset. This distinctive approach sets it apart from conventional methods, introducing a fresh perspective to anomaly detection. What distinguishes the Isolation Forest algorithm is its departure from the customary reliance on distance and density metrics. Instead, it innovatively incorporates the concept of isolation, thereby elevating the efficacy and efficiency of anomaly identification. This strategic shift leads to improved accuracy and a more nuanced understanding of anomalies within the data. An additional advantage of the Isolation Forest algorithm is its resource efficiency. It demands minimal memory allocation, making it suitable for diverse computational environments. Furthermore, its computational overhead is remarkably low, thanks to its intrinsic linear time complexity.

B. *Local Outlier Factor*

To identify outliers, the LOF algorithm compares a data point’s local variances to those of its neighbors. It gauges outliers based on local density, influenced by proximity to nearest neighbors. Calculating the data point’s density against neighboring points enables identification of regions with akin densities, singling out those with notably lower densities. LOF measures the extent of density deviation of data point from its neighbors; if significantly lower, it’s considered an outlier. By emphasizing local relationships and densities, LOF unveils nuanced anomalies often unnoticed by conventional methods. This approach enriches anomaly detection’s granularity, offering a more thorough comprehension of irregularities within the dataset.

C. *Implementation*

There are three distinct subsets of the dataset - training set, validation set, and testing set. The classifier is initially taught on the training dataset, then hyperparameter tuning is performed on the validation dataset, and finally its performance is evaluated on the testing data. Notably, the model is solely exposed to the training and validation datasets. A crucial point of emphasis is the strict prohibition of utilizing the test dataset for classifier training.

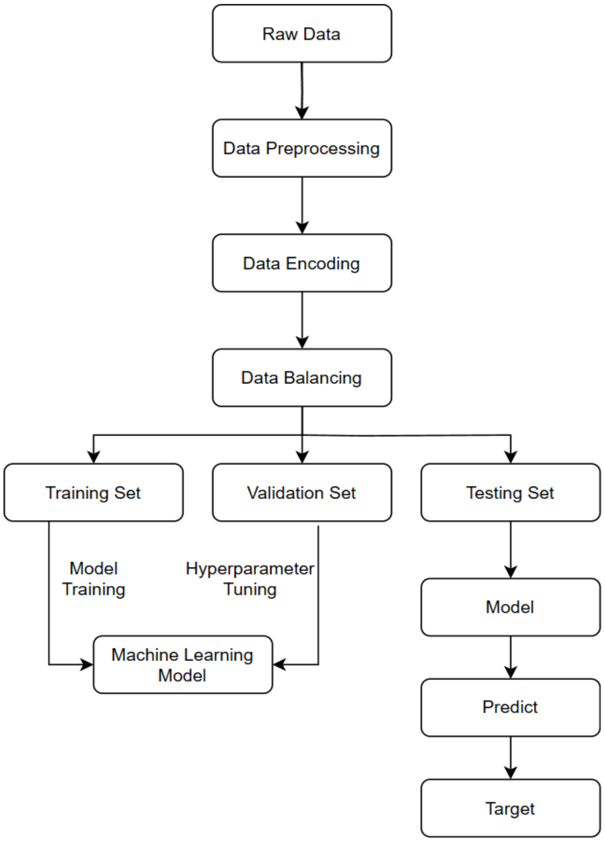


Fig. 1: Block Diagram of the Proposed Model

As depicted in Fig. 1 above, the block diagram delineates the configuration of the planned model. The model’s operational sequence comprises several key steps - data collection, data processing, identification of the suitable model based on data characteristics, subsequent model training and testing, and culminating in a comprehensive evaluation. This stage is pivotal in driving the refinement of machine learning model accuracy.

In unsupervised learning, the training process involves using only the variables present in the dataset, without any predefined labels. The goal is to build models that can identify patterns or deviations within the data. Accuracy rate, precision, recall, and F1-score are just some of the criteria used to assess these models’ efficacy, the detection of outliers, and a confusion matrix, which includes metrics like true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These models are then put to the test by making predictions during the testing phase.

In our research project, we have two sets of data. For both of these datasets, we begin by evaluating the models without making any adjustments to the class distribution. Then, we proceed to apply resampling techniques on both the European and German credit card fraud transaction datasets. These techniques include random undersampling, AllKNN, SMOTE, and SMOTE-ENN. After each resampling technique is applied, we record and analyze the evaluation metrics to understand how they have changed. This procedure allows us to evaluate the effect of the various resampling techniques on the accuracy of the models applied to both data sets.

I. Results and Discussions

To evaluate the most efficient algorithm for detecting fraudulent transactions, various criteria for comparing algorithms have been implemented. Accuracy, recall, and precision are the uttermost popular metrics used for the evaluation of the success of machine learning algorithms. All of these measurements can be computed using a Confusion matrix, which is a structured way of evaluating model performance. The assessment of how well a model performs was carried out based on these metrics. The models were subjected to testing using both the original dataset and a dataset that had been resampled. The findings clearly demonstrated that in most of the cases the procedure of resampling plays a significant aspect in influencing the performance of the models.

On the imbalanced European and German credit card dataset, two prominent anomaly detection models, IF and LOF were employed. The aim was to improve their performance, reduce false negatives, and enhance the total accuracy of the models. Additionally, various resampling techniques were explored to tackle the challenge posed by imbalanced data.

Table I: Model performance on European Dataset without resampling.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	No. of Outliers
IF	99.81	63	53	54	107
LOF	99.81	50	50	50	104

Table II: Model performance on German Dataset without resampling.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	No. of Outliers
IF	70.50	35	50	41	59
LOF	70.50	35	50	41	59

To diminish the impact of class imbalance, four resampling techniques were applied - Random Undersampling, AllKNN, SMOTE, and SMOTE-ENN. These techniques aimed to balance the class distribution and enhance the models ability to capture patterns from both majority and minority classes. The results of the IF and LOF models using a variety of resampling strategies on European and German dataset are shown in Tables 3, 4, 5, and 6.

Table III: IF Model performance on European Dataset with various resampling techniques.

Resampling Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	No. of Outliers
Random under-sampling	98.98	49	50	50	101
AllKNN	98.80	53	78	55	682
SMOTE	49.76	25	50	33	57130
SMOTE-ENN	48.77	25	49	33	58201

Table IV: LOF Model performance on European Dataset with various resampling techniques.

Resampling Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	No. of Outliers
Random under-sampling	98.99	49	50	50	100
AllKNN	98.96	50	50	50	591
SMOTE	49.98	67	50	34	56880
SMOTE-ENN	49.53	42	50	34	57341

Table V: IF Model performance on German Dataset with various resampling techniques.

Resampling Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	No. of Outliers
Random under-sampling	70.56	35	50	41	53
AllKNN	51.61	48	48	48	45
SMOTE	52.85	47	50	36	132
SMOTE-ENN	43.80	22	50	30	59

Table VI: LOF Model performance on German Dataset with various resampling techniques.

Resampling Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	No. of Outliers
Random under-sampling	70.60	35	50	41	52
AllKNN	54.83	55	55	54	42
SMOTE	51.42	26	48	34	136
SMOTE-ENN	43.80	47	50	33	59

A comparison of all models employed on European and German dataset, along with resampling techniques, is conducted to discover the optimal model, and the results are summarized in Table 7. Whereas Table 8 illustrates the comparative study of our proposed approach with the already published research findings.

Table VII: Best Model on respective dataset

Dataset	Best Model
European	IF
German	LOF - Random under-sampling

Table VIII: Comparative analysis with previous reports on European Dataset.

Reference	Model	Accuracy (%)
Ref [11]	LOF	97
Ref [12]	IF	98.72
Ref [14]	IF	99.72
Proposed	IF	99.81

To effectively prevent credit card fraud and protect the security of financial transactions, this research highlights the need of employing both modern algorithms and data preparation methodologies.

Moreover, the positive impact of resampling was evident in the complete elimination of false negatives (FN), as well as a reduction in the count of false positives. This indicates that fraudulent transactions were not only effectively identified but also distinguished from normal transactions, thus minimizing both types of misclassifications.

I. Conclusion

Through the use of a dataset built from German credit card transactions and European credit card transactions, this study compares the performance of various machine learning classification techniques. Identifying specific transactions as fraudulent is the main goal. The central aim is to discern the fraudulent nature of specific transactions. Both the credit card dataset underwent a series of stages, including importation, preprocessing, encoding, and configuration, in preparation for model training facilitated by the machine learning workflow mechanism. Subsequently, the dataset was subjected to training, deployment, and evaluation for each classification model, encompassing the utilization of multiple resampling techniques.

In particular, when conducting a comparative analysis between the IF model and the LOF model for the European credit card fraudulent transactions dataset, the IF model outperforms the LOF model in terms of recognizable performance, with an accuracy of 99.81%.
On the German credit card fraudulent transaction dataset, the LOF algorithm performs at its highest accuracy of 70.60% over the IF model when executed in integration with the random undersampling methodology.

References

[1] Varmedja, Dejan, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, and Andras Anderla. ”Credit card fraud detection-machine learning methods.” In 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), pp. 1-5. IEEE, 2019.

[2] Asha, R. B., and Suresh Kumar KR. ”Credit card fraud detection using artificial neural network.” Global Transitions Proceedings 2, no. 1 (2021): 35-41.

[3] Sailusha, Ruttala, V. Gnaneswar, R. Ramesh, and G. Ramakoteswara Rao. ”Credit card fraud detection using machine learning.” In 2020 4th international conference on intelligent computing and control systems (ICICCS), pp. 1264-1270. IEEE, 2020.

[4] Chugh, Bharti, and Nitin Malik. ”Machine Learning Classifiers for Detecting Credit Card Fraudulent Transactions.” In Information and Communication Technology for Competitive Strategies (ICTCS 2021) ICT: Applications and Social Interfaces, pp. 223-231. Singapore: Springer Nature Singapore, 2022.

[5] Ileberi, Emmanuel, Yanxia Sun, and Zenghui Wang. ”Performance evaluation of machine learning methods for credit card fraud detection using SMOTE and AdaBoost.” IEEE Access 9 (2021): 165286-165294.

[6] Mienye, Ibomoiye Domor, and Yanxia Sun. ”A Deep Learning Ensemble With Data Resampling for Credit Card Fraud Detection.” IEEE Access 11 (2023): 30628-30638.

[7] Bagga, Siddhant, Anish Goyal, Namita Gupta, and Arvind Goyal. ”Credit card fraud detection using pipeling and ensemble learning.” Procedia Computer Science 173 (2020): 104-112.

[8] Chen, Joy Iong-Zong, and Kong-Long Lai. ”Deep convolution neural network model for credit-card fraud detection and alert.” Journal of Artificial Intelligence and Capsule Networks 3, no. 2 (2021): 101-112.

[9] Shirodkar, Nikita, Pratikesh Mandrekar, Rohit Shet Mandrekar, Rahul Sakhalkar, KM Chaman Kumar, and Shailendra Aswale. ”Credit card fraud detection techniques–A survey.” In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1-7. IEEE, 2020.

[10] Esenogho, Ebenezer, Ibomoiye Domor Mienye, Theo G. Swart, Kehinde Aruleba, and George Obaido. ”A neural network ensemble with feature engineering for improved credit card fraud detection.” IEEE Access 10 (2022): 16400-16407.

[11] John, Hyder, and Sameena Naaz. ”Credit card fraud detection using local outlier factor and isolation forest.” Int. J. Comput. Sci. Eng 7, no. 4 (2019): 1060-1064.

[12] Gupta, Swati, Sanjay Patel, Surender Kumar, and Goldi Chauhan. ”Anomaly detection in credit card transactions using machine learning.” (2020).

[13] Kittidachanan, Kittikun, Watha Minsan, Donlapark Pornnopparath, and Phimphaka Taninpong. Anomaly detection based on GS-OCSVM classification. In 2020 12th International Conference on Knowledge and Smart Technology (KST), pp. 64-69. IEEE, 2020.

[14] Vijayakumar, V., Nallam Sri Divya, P. Sarojini, and K. Sonika. ”Isolation forest and local outlier factor for credit card fraud detection system.” International Journal of Engineering and Advanced Technology (IJEAT) 9 (2020): 261-265.

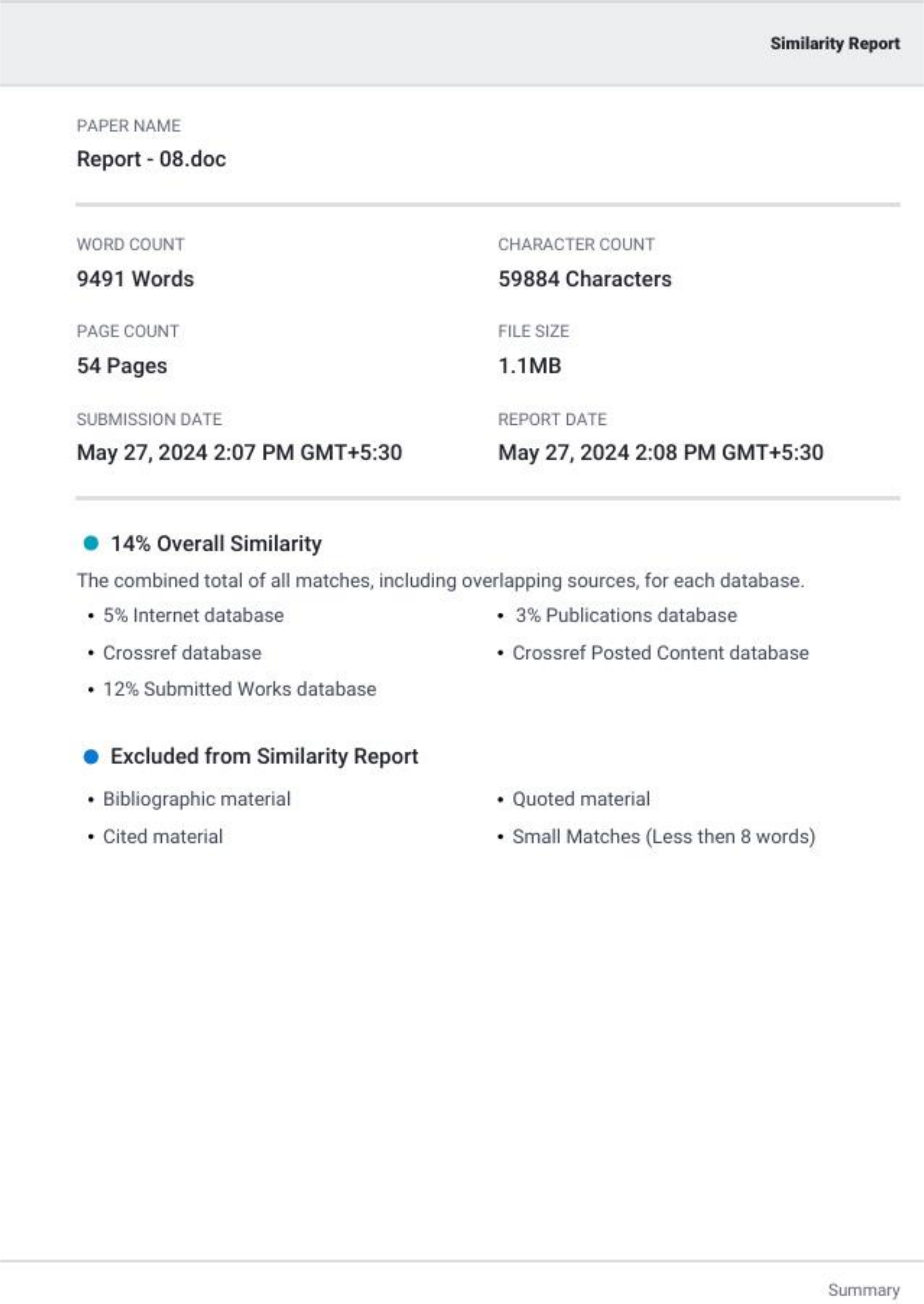
[15] Rajeev, Haritha, and Uma Devi. ”Detection of credit card fraud using isolation forest algorithm.” In Pervasive Computing and Social Networking: Proceedings of ICPCSN 2021, pp. 23-34. Springer Singapore, 2022.

[16] Itoo, Fayaz, Meenakshi, and Satwinder Singh. ”Comparison and analysis of logistic regression, Na¨ıve Bayes and KNN machine learning algorithms for credit card fraud detection.” International Journal of Information Technology 13 (2021): 1503-1511.

[17] Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. “Isolation forest.” Data Mining, 2008. ICDM’08. Eighth IEEE International Conference on

[18] Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. “Isolation-based anomaly detection.” ACM Transactions on Knowledge Discovery from Data (TKDD) 6.1 (2012): 3

[19] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In ACM sigmod record.



Similarity Report

PAPER NAME

Report - 08.doc

WORD COUNT **9491 Words**	CHARACTER COUNT **59884 Characters**
PAGE COUNT **54 Pages**	FILE SIZE **1.1MB**
SUBMISSION DATE **May 27, 2024 2:07 PM GMT+5:30**	REPORT DATE **May 27, 2024 2:08 PM GMT+5:30**

● 14% Overall Similarity

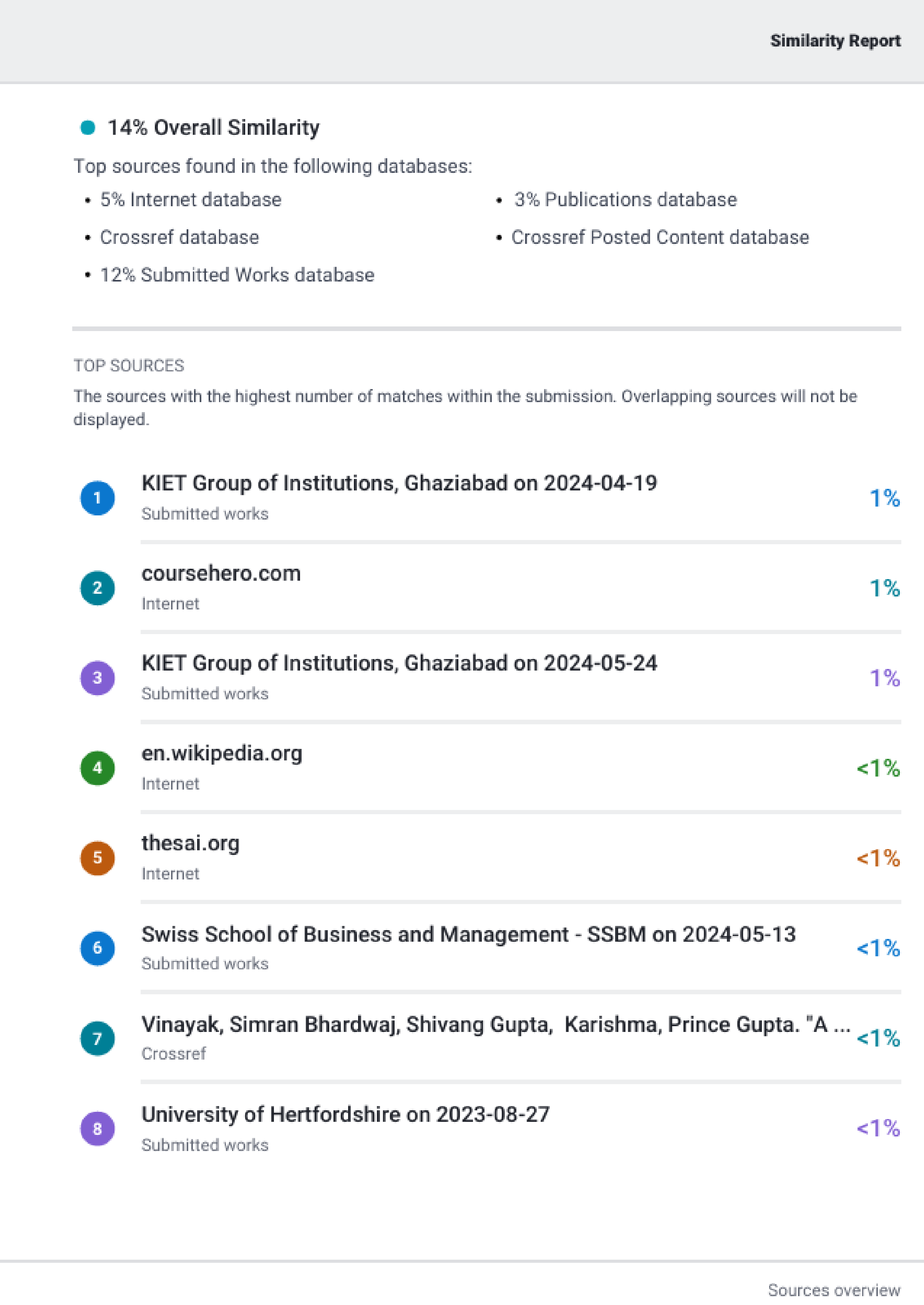
The combined total of all matches, including overlapping sources, for each database.

- 5% Internet database
- 3% Publications database
- Crossref database
- Crossref Posted Content database
- 12% Submitted Works database

● Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 8 words)

Summary



Similarity Report

● 14% Overall Similarity

Top sources found in the following databases:

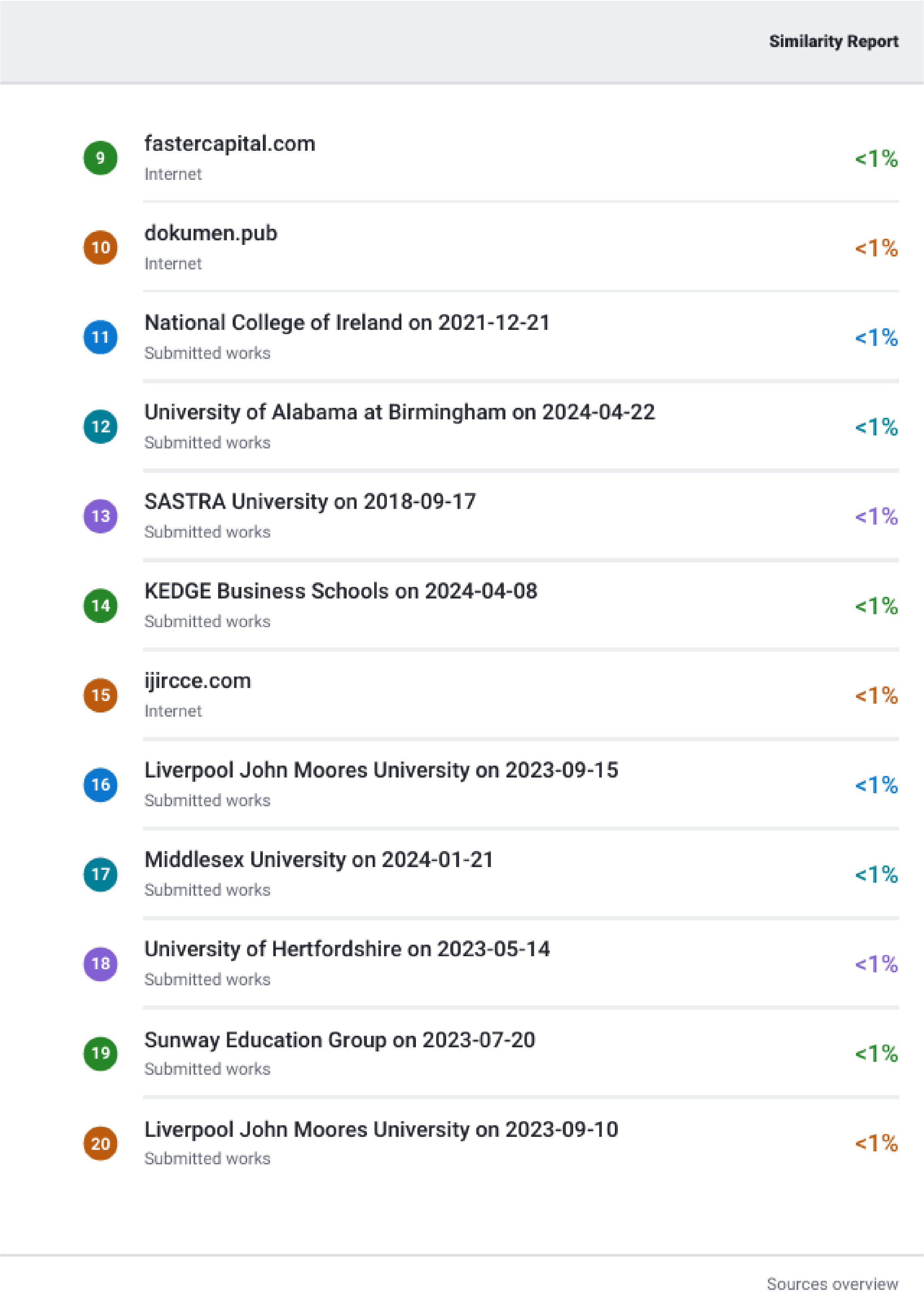
- 5% Internet database
- 3% Publications database
- Crossref database
- Crossref Posted Content database
- 12% Submitted Works database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	KIET Group of Institutions, Ghaziabad on 2024-04-19 Submitted works	1%
2	coursehero.com Internet	1%
3	KIET Group of Institutions, Ghaziabad on 2024-05-24 Submitted works	1%
4	en.wikipedia.org Internet	<1%
5	thesai.org Internet	<1%
6	Swiss School of Business and Management - SSBM on 2024-05-13 Submitted works	<1%
7	Vinayak, Simran Bhardwaj, Shivang Gupta, Karishma, Prince Gupta. "A ... Crossref	<1%
8	University of Hertfordshire on 2023-08-27 Submitted works	<1%

Sources overview



Similarity Report

9	fastercapital.com Internet	<1%
10	dokumen.pub Internet	<1%
11	National College of Ireland on 2021-12-21 Submitted works	<1%
12	University of Alabama at Birmingham on 2024-04-22 Submitted works	<1%
13	SASTRA University on 2018-09-17 Submitted works	<1%
14	KEDGE Business Schools on 2024-04-08 Submitted works	<1%
15	ijircce.com Internet	<1%
16	Liverpool John Moores University on 2023-09-15 Submitted works	<1%
17	Middlesex University on 2024-01-21 Submitted works	<1%
18	University of Hertfordshire on 2023-05-14 Submitted works	<1%
19	Sunway Education Group on 2023-07-20 Submitted works	<1%
20	Liverpool John Moores University on 2023-09-10 Submitted works	<1%

Sources overview