Unveiling the Transformative Potential of Dark Data: An Optimized Extraction Method for Accurate Accident Prediction in Big Data

Masroor Shah*, Fazal Malik*, Abd Ur Rub, Muhammad Suliman*, Irfan ullah*, Sana Ullah*, Romaan Khan***, Salman Alam******

*Department of Computer Science, Iqra National University Peshawar, Khyber Pakhtunkhwa (KPK), Pakistan
**School of Electronics and Information, Northwestern Polytechnical University, Xi'an Shaanxi, China
*** City University of Science and Information Technology Peshawar, KPK, Pakistan
****COMSATS University Islamabad (CUI), Pakistan

Abstract- Amidst the era of data-driven decision-making, the persistent challenge of dark data—a reservoir of untapped information within routine transactions—necessitates the development of robust approach for its extraction. The exponential growth of big and dark data, particularly the escalating complexity of the latter, poses formidable challenges. This research introduces an optimized extraction method for accessing dark data using AdaBoost and Random Forest classifiers, showcasing transformative potential for precise predictions and uncovering latent insights. This method encompasses phases of analysis, implementation, and evaluation, aiming to address challenges in extracting insights from dark data. The evaluation of AdaBoost and Random Forest classifiers reveals AdaBoost's commendable overall accuracy of 78.4%, while the Random Forest classifier excels further with an impressive overall accuracy of 89.48%. This research pioneers the application of dark data in accident prediction, resulting in a substantial accuracy boost from 78.4% to 89.48%. Emphasizing the transformative potential of using dark data for precise predictions, the study underscores tangible benefits for decision-makers and urban planners. As data volumes escalate, the increasingly evident pivotal role of dark data in advancing decision-making and enhancing predictive model accuracy establishes a precedent for unlocking latent insights within previously untapped datasets.

Index Terms- Big data, Data quality, Dark data, Complexity of dark data, Accident prediction

I. Introduction

The introduction emphasizes data's foundational role across diverse domains and explores challenges in big data, including issues of quality, storage, security, and analytics. It introduces dark data, emphasizing its untapped potential within organizational datasets. The narrative underscores the exponential growth and escalating complexity of both big and dark data. The conclusion highlights the imperative for strategic interventions to unlock the full potential of big data for future applications and decision-making processes.

A. Data

Data represents the raw material of information, sourced from diverse domains such as business, economic, social networking sites, the Internet of Things, scientific areas, and sensor devices. All data inherently carries meaningful information, and the amalgamation of information contributes to knowledge [1].

B. Big Data

In contemporary activities, substantial volumes of intricate data are generated across various sectors, encompassing business, economic, social networking sites, the Internet of Things, and scientific domains. The aggregation and collection of extensive and complex data in routine transactions characterize what is commonly referred to as big data [1]. This dataset includes traditional system data, general and web data, and is notably influenced by machine-generated, sensor-derived, and social media data, with a predominant focus on platforms like Facebook, Twitter, email, blogs, and other social media channels [2].

Furthermore, advancements in data capture devices result in the generation of vast datasets, comprising unstructured data, semi-structured data, voice recordings, videos, audios, and graphics, both within and outside organizational boundaries. Historically, organizations faced challenges in managing big data due to limited technologies, storage capacities, and the high costs associated with tools [3]. Crucial issues such as data cleaning, analysis, processing, security, and information extraction from large datasets posed considerable difficulties for traditional technology [4]. However, contemporary big data projects worldwide have ushered in unprecedented opportunities. They offer the latest tools, frameworks, models, security measures, processing capabilities, storage capacities, and real-time analysis, all of which are more scalable, flexible, and performance-oriented compared to traditional technologies [5].

C. Types of Big Data

Different classifications of big data are elucidated [6] in Figure 1.

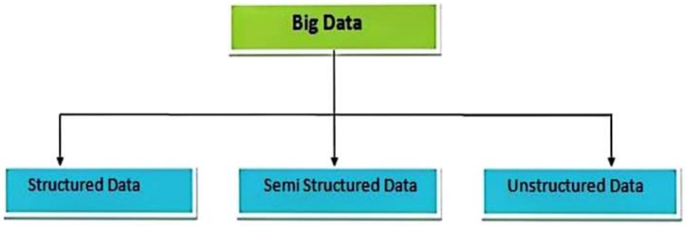


Figure 1. *Types of Big Data*

1. Structured Data

Structured data adheres to the formal structure of data models linked with relational databases or other data tables. Essentially, structured data refers to information with a specific predefined format.

2. Semi-Structured Data

Semi-structured data is a variant of structured data that deviates from the formal structure of data models associated with relational databases. While it lacks strict conformity, it still includes tags or other semantic elements to organize hierarchical data within records and fields.

3. Unstructured Data

Unstructured data lacks compatibility with traditional databases and lacks a discernible internal structure. This stands in contrast to structured data stored in databases. A significant portion—up to 80%—of business data is categorized as unstructured, and this proportion is on the rise annually.

D. Characteristics of Big Data Technology

In the exploration of big data technology, key characteristics play a pivotal role, as outlined [8] in Figure 2.

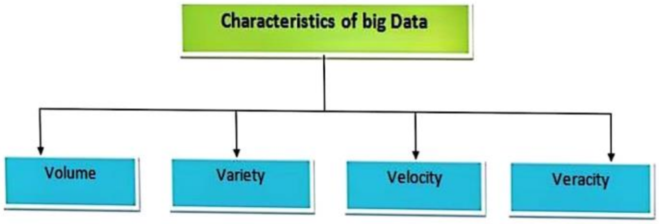


Figure 2. *Characteristics of Big Data*

1. Volume

The term "volume" in the context of big data technology refers to the sheer quantity of data, encompassing both the amount and the storage requirements. It involves the capacity to store data securely during routine activities, quantified in units such as Kilo byte (KB), Megabyte (MB), Giga byte (GB), and Tera byte (TB).

2. Variety

Variety, a critical facet of big data, pertains to the diverse array of data types that the system can accommodate. This encompasses the nature of the stored data, including structured, semi-structured, unstructured, and machine-generated data.

3. Velocity

Velocity, as a defining property of big data, denotes the speed at which data is processed and accessed. It encapsulates the rapidity of data processing, including streaming, real-time operations, and remote control functionalities.

4. Veracity

Veracity highlights the paramount importance of security in big data systems, particularly given the substantial volume of data involved. The transition to cloud storage magnifies the significance of safeguarding data from external threats and unknown entities.

5. Value

Following Veracity, Velocity, Variety, and Volume, the preeminent characteristic of big data is its value. Organizations engage in analyzing and leveraging big data to derive benefits, aiming for high profits and a favorable return on investment. This underscores the strategic utilization of big data in achieving organizational objectives and financial success.

E. Example of Big Data

Big Data, as illustrated in Figure 3, encompasses various data types and sources, each presenting unique challenges and opportunities for analysis. The following examples highlight significant domains contributing to the vast landscape of Big Data [9]:

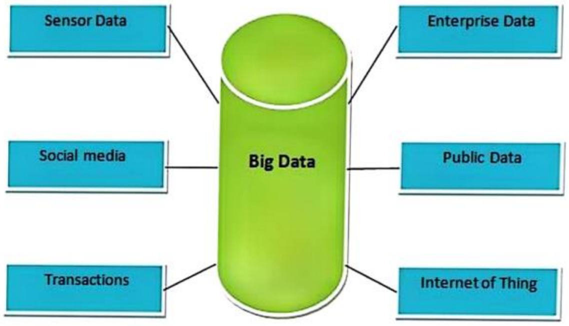


Figure 3. *Examples of Big Data*

1. Sensor Data

Sensor data constitutes information generated by devices designed to detect and respond to stimuli within the physical environment. The evolution of data capture devices has resulted in the collection of extensive datasets, encompassing unstructured, semi-structured, and multimedia data such as voice recordings, videos, audios, and graphics. This data originates both within and outside organizational boundaries, requiring advanced analytics to derive meaningful insights

2. Social Media

Social media platforms, including but not limited to Facebook, Twitter, email, and blogs, generate substantial volumes of data. This data serves as a valuable resource for social media analytics tools and applications. Common applications include leveraging customer sentiment analysis to enhance business marketing strategies and improve customer service activities [9].

3. *Transactions*

Routine machine activities contribute to the generation and storage of transactional information. This category encompasses embedded system data, records of purchases, transaction logs, mobile phone records, and other forms of transactional data. These datasets are integral to understanding and optimizing various business processes [10].

4. *Enterprise Data*

Enterprise data refers to information shared among the users of an organization to derive essential benefits. This data is highly critical and necessitates stringent security measures. Enterprise data management systems play a pivotal role in ensuring data security, managing internal and external communication, and facilitating data retrieval from various applications [11].

5. *Public Data*

Public data stands as a significant data source freely accessible to anyone, without legal restrictions on access, reuse, and redistribution. This includes government, national, local, and international datasets that contribute to diverse applications and analyses [9].

6. *Internet of Things (IoT)*

The Internet of Things (IoT) represents a rapidly expanding category within Big Data, driven by advancements in communication and sensor technologies. IoT encompasses a network of physical devices and items embedded with actuators, sensors, and communication devices. These interconnected entities collect and exchange data, fostering a seamless flow of information through network connectivity [12]. The diverse examples of Big Data underscore the multifaceted nature of this field, where a comprehensive understanding of various data types and sources is crucial for effective analysis and utilization.

F. *Challenges in Big Data Management*

This academic discourse delves into the multifaceted challenges inherent in the realm of big data, emphasizing the complexities associated with its collection, storage, processing, and utilization. The analysis focuses on six pivotal challenges: Data Quality, Information Discovery, Storage, Security, Data Analytics, and the Lack of Talents. Each challenge is scrutinized within the context of its implications for organizations striving to harness the potential of big data for future applications [15].

1. *Data Quality*

The veracity of big data emerges as a paramount challenge, characterized by the pervasive issues of data dirtiness, inconsistency, and disorderliness. Notably, substantial financial resources are expended annually by various nations to rectify and enhance the quality of data, aiming to facilitate the nuanced analysis and extraction of qualitative insights [44].

2. *Information Discovery*

The intricate task of information discovery is impeded by the presence of untidy and disorganized data sets. The extraction of meaningful insights from such data necessitates a profound understanding of data management, design, and computational methodologies. Organizations grapple with the complexities of transforming raw data into actionable information [45].

3. *Storage*

The sheer magnitude and intricacy of data generated across diverse fields pose a significant conundrum regarding storage. Organizations encounter challenges in managing the colossal volumes of data, necessitating sophisticated high-capacity storage systems. The associated costs, both in terms of infrastructure and expert personnel, amplify the complexities of data storage management.

4. *Security*

Ensuring the security of big data represents a formidable challenge, encompassing the safeguarding of data against unauthorized access and corruption throughout its lifecycle. The multifaceted nature of big data exacerbates the complexity of implementing robust security measures, demanding vigilant strategies to protect sensitive information.

5. *Data Analytics*

Analyzing the disorderly and chaotic nature of big data poses formidable challenges in terms of its management, analysis, and processing. The intricate nature of messy data complicates the extraction of meaningful insights, thereby impeding the effective utilization of big data for future strategic purposes [46].

6. *Lack of Talents*

A critical impediment in the effective management of big data projects is the scarcity of skilled professionals, including developers, analysts, and data scientists. The voluminous nature of big data necessitates expertise in novel algorithms and techniques, presenting a considerable challenge for organizations aiming to navigate the complexities of big data analytics.

The challenges associated with big data are multifaceted and require strategic interventions encompassing data quality improvement, enhanced information discovery techniques, efficient storage solutions, robust security measures, advanced data analytics methodologies, and the cultivation of a skilled workforce. Addressing these challenges is imperative for organizations seeking to harness the full potential of big data for future applications and decision-making processes.

G. *Dark Data*

Dark data, a subset of big data, refers to information within a business or organization that is collected and processed during daily transactions but remains unusable for future and essential organizational benefits. Despite sharing similarities with big data in terms of cost and risk, certain data lacks value and utility for organizational objectives but is still employed in decision-making processes. The proportion of useful versus dark data is illustrated in Figure 4.

Various sources, including traditional enterprise, social media, machine, and sensor devices, contribute complex data, with up to 80% falling under the category of dark data. This implies that only 10% of the data is deemed useful, while the remaining 90% is classified as dark data [15]. Dark data comprises a substantial portion of big data that is neither beneficial nor usable for significant purposes, often concealed in cloud and machine storage. Analyzing this unexplored data proves challenging in the real world, and its value is crucial for future opportunities. The associated costs are notably high, and the expanding volume of big data necessitates innovative algorithms and techniques [16].

Dealing with big data and its dark side poses significant challenges for businesses and organizations. Identifying information or value from raw or semi-structured data remains an obstacle. The majority of businesses aim to retain valuable data, as extracting useful information from dark data can be beneficial. Considering the prevalence of unstructured data within organizations, efficient mining to derive valuable patterns becomes pivotal for making informed future decisions. Prior research has yielded techniques, algorithms, tools, and software solutions for addressing dark data, aiming to extract valuable information and mitigate the expensive and risky nature of data storage [17-18].

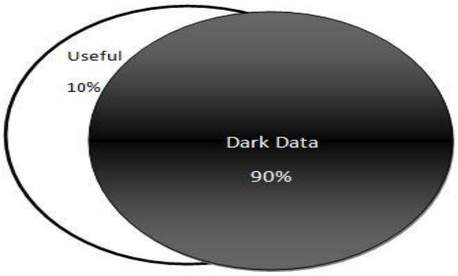


Figure 4. *Examples of Dark Data Percentage*

Gartner's research underscores the growing complexity of dark data as a challenging issue for organizations. Accessing, recovering, and discovering information from dark data, including files, emails, backup files, and structured and unstructured data, has become increasingly difficult. A substantial 70% of data within organizations and business areas is identified as dark data, imposing significant costs and risks [19].

H. Overall Growth of Big and Dark Data

The Figure 5 illustrates the comprehensive growth trajectory of both Big and Dark data. These two categories encapsulate the burgeoning volume of data in contemporary information ecosystems. The subsequent sections delineate the nature and characteristics of various data types contributing to this expansive growth.

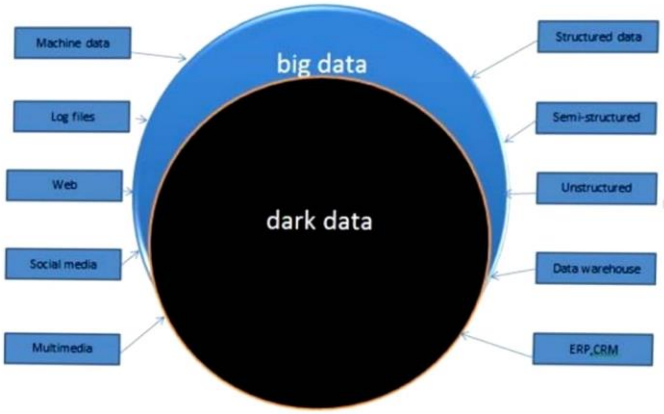


Figure 5. *Overall Growth of Big and Dark Data*

I. Structured Data

Structured data adheres to formal data models associated with relational databases or similar tabular formats. It is characterized by a specific data shape conforming to predefined structures, often found in traditional database systems.

J. Semi-Structured Data

Semi-structured data deviates from formal data models associated with relational databases but includes tags or markers that segregate semantic elements. Despite lacking strict adherence to tabular structures, it enforces hierarchies of records and fields within the data.

K. Unstructured Data

Unstructured data lacks a discernible internal structure and does not neatly fit into traditional databases. It constitutes a substantial portion of business data, estimated to be up to 80%, and is continually increasing. This type of data poses challenges due to its lack of identifiable structure.

L. Data Warehouse

A data warehouse, or enterprise data warehouse (EDW), is a pivotal component of business intelligence. It serves as a centralized repository that integrates data from diverse sources, facilitating reporting and data analysis.

M. Social Media

Social media platforms such as Facebook and Twitter generate vast amounts of data, including user interactions, emails, blogs, and other content. Analytics tools leverage this data to understand customer sentiment, supporting marketing and customer service activities.

N. Web Data

Web data encompasses information provided by the World Wide Web through internet services. It serves various applications, facilitating the consumption, provision, sharing, and publication of information.

O. Log File

A log file records events or messages between operating systems or different software components. This chronicle of activities is particularly valuable for diagnosing issues, monitoring behavior, and maintaining records.

P. Multimedia Data

Multimedia data databases consist of diverse media types, including sequential data, videos, audios, animations, text, images, and graphics. This compilation presents a multifaceted approach to data representation.

Q. Machine Data

Machine data emanates from the activities of computerized or machine processes, encompassing embedded systems, mobile phones, and other networked devices. It involves the generation and storage of information by physical devices equipped with sensors, actuators, and electronic components.

R. ERP and CRM

Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM) systems are pivotal in managing and securing vast amounts of data. These systems facilitate data retrieval from both internal and external sources, supporting seamless communication and application integration.

In the realm of daily business transactions, organizations generate substantial data for future utility. However, the challenge extends beyond mere storage concerns to the realm of dark data, which is characterized not only by its non-utilization in decision-making but also by its inherent lack of structure. Addressing dark data necessitates the formulation of policies and the deployment of technologies to comprehend and manage its significance effectively.

Organizations and industries increasingly employ sensor technologies, Machine to Machine (M2M) communication, and the Internet of Things (IoT) for real-time data processing, contributing to the proliferation of big data. The prevalence of unstructured data, exceeding the volume of structured data, underscores the need for high-speed systems and advanced computing tools to handle this diverse and voluminous data landscape.

II. Literature Review

The literature review delves into challenges of dark data in big data, focusing on its definition, management, and utilization complexities. Dark data, an untapped information source, presents significant challenges due to its inherent complexity. Proposed solutions, including reduplication methods, are discussed, emphasizing the impact of dark data on tasks, data analysis, and decision-making. The subsequent research introduces an advanced extraction method with AdaBoost and Random Forest classifiers, showcasing transformative potential for precise predictions and latent insights.

A. Review of Unseen Dimensions of Dark Data

Dark data, a subset of big data, encompasses information generated in various fields such as business, economic, social networking, the Internet of Things, and sciences. It includes traditional system data, general web data, machine data, sensor data, and social media content. Dark data poses challenges as it consists of information stored in daily transactions but remains unused for future benefits, representing a costly and risky aspect of big data. Some data lack value and usability for decision-making. Researchers have strived to use dark data for better decisions; with studies indicating that up to 80% of data may be classified as dark data as shown in Figure 6. This subset mainly exists within the structure of big data, and traditional enterprise databases have lower chances of generating dark data [14, 15].

The researcher faced challenges utilizing dark data due to the complexity and size of big data. To address issues such as system load, speed, and power consumption, they identified techniques, including reduplications. Duplication was identified as a cause of increased system load, complexity, and processing speed. The solution involved separating duplicated data as metadata and storing datasets in cloud storage, followed by real-time reduplication during post-processing.

The authors in [16] explored challenges and opportunities of dark data, focusing on data management to address performance issues. They highlighted the negative impact of CPU, RAM, and DISK storage usage on tasks and emphasized the need for careful handling of dark data due to its potential risks and challenges.

Article in [17] provided insights into task execution and management processes in data analysis. The research identified root causes of unsuccessful tasks, such as fail, kill, and eviction, leading to resource waste and application slowdown. Analyzing patterns of unsuccessful jobs contributed to improving application performance, low latency, and fault tolerance in big data systems.

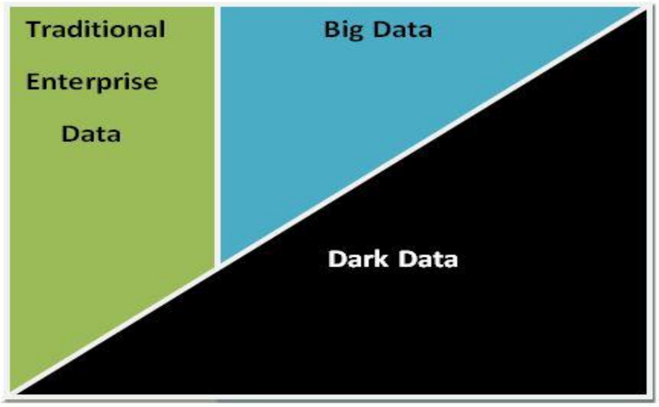


Figure 6. *Growth of Dark Data*

The authors in [18] worked on semi-structured and unstructured dark data in industrial and business contexts. They used Business Intelligence (BI) software to analyze this data, proposing a BI strategy for extracting valuable information from dark data for organizational benefit, as shown in Figure 7.

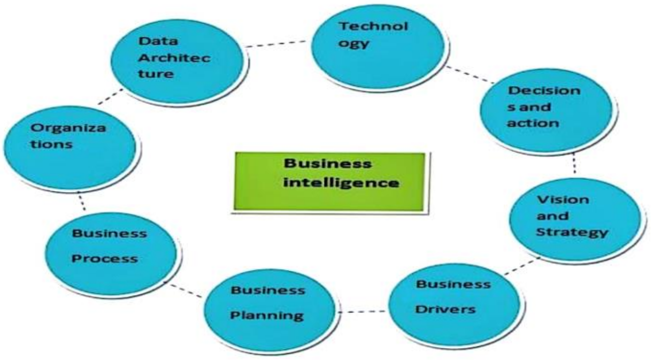


Figure 7. *Business Intelligence (BI) Strategy*

Gartner's assessment in [19] underscores the challenge organizations face in harnessing email and file-level analytics tools for dark data. Despite long neglect, dark data proves vital for enhancing business decisions. Introducing the Simpana analytic tool, Gartner leverages email and file-level data to extract valuable insights, promoting a better understanding of data assets and offering cost reduction, risk mitigation, and operational simplicity.

The authors in [20], present a Textile application addressing real-world issues through the analysis of both structured and unstructured data. The application generates grid layout visualization, aiding in data management and problem-solving by extracting useful summaries from various data types. The [21] reviews unstructured data from social media platforms and applies OLAP technology to analyze numerical, textual, and elemental elements. The research contributes to the management and facilitation of unstructured data. The authors in [22] focuses on designing big data techniques and algorithms to tackle current and future challenges, while authors in [23] analyzes unstructured data through various methods, reviewing applications such as text and audio analysis, video analytics, and social media analysis.

In [24], the authors develop a practical system for analyzing dark data in PDF files, achieving a 75% average accuracy rate using morphological analysis and classification reports based on SVM and neural network models. The article in [25] poses critical questions about big data and its dark side, emphasizing the importance of combining existing big data with knowledge, experience, and courage for better decision-making. The paper acknowledges challenges but underscores potential solutions through effective execution. Research in [26] centers on CY Verse network infrastructure technology for computable use in astronomy, while [27] removes duplication from datasets using Hadoop processing and features for fast duplication removal. In article [28] the authors collect textual data from the internet using APIs to understand global society's behavior, employing K-Means cluster analysis for insights into countries' internet activities, national technology evolution, and responses to stressors.

Research study in [29], delves into big data research, emphasizing the 4V model (Velocity, Volume, Variety, and Veracity), and explore the disk spaces associated with big data and their implications.

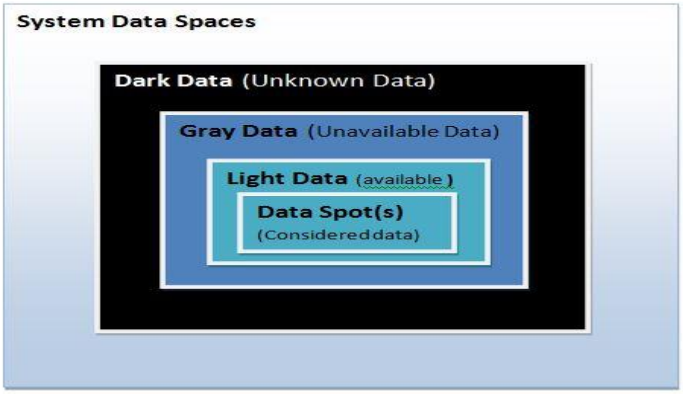


Figure 8. *Data Space in Big Data*

Figure 8 illustrates the data space within the realm of big data, encompassing various types such as dark data, gray data, light data, and data spots. Dark data, as defined by researchers, represents unknown and unqualified data. Gray data is partially known and can be analyzed, while light data is readily available for use and processing. Data spots are subsets of light data. In [30], researchers shed light on dark data related to Northwestern Atlantic zooplankton from the 1970s to 1980s, emphasizing the need to bring this data into public visibility. In [31] authors explore the importance and challenges of dark data and big data, emphasizing the opportunities they present for business analysis. In [32] authors review dark data, highlighting its significance and revealing that 90% of data exists in dark form. In [33] authors focus on data sharing, revealing that a substantial portion of data remains in the long tail, categorized as dark data. In [34] a reviews the government's future based on big data, dark data, smart data, and open data, emphasizing the challenges and importance for organizations. An approach introduces a system called deep dive for dark data, employing SQL queries to extract high-quality data from various domains [35]. A method provides insights into dark data, its challenges, and opportunities, emphasizing machine learning technology and proposing a methodology for utilizing dark data in the IoT domain [36] authors. An approach focuses on accurate Loss-of-Coolant Accident (LOCA) diagnosis in nuclear power plants. It challenges the assumption that complex models yield better performance, emphasizing the impact of dataset construction. The study introduces DeepLOCA-Lattice, utilizing basic models with good accuracy, revealing optimal performance parameters for breach size estimation in LOCA [40].

A study uses machine learning to address drunk driving-related road fatalities, building prediction models for early detection and policy development. Various supervised machine learning algorithms, including Random Forests, Decision Tree, Naïve Bayes, Logistic Regression, and Support Vector Machine, are implemented on traffic fatalities data [41].

Intelligent transportation systems enhance traffic management, and generative AI addresses issues like data sparsity and abnormal scenario observation. This review explores generative AI's applications in traffic perception, prediction, simulation, and decision-making, summarizing challenges and proposing future research directions for intelligent transportation systems [42].The University of British Columbia lab addresses metabolomics big data challenges through interdisciplinary work in chemistry, computer science, and statistics. Our bioinformatics tools on GitHub tackle data processing, feature extraction, quantitative measurements, statistical analysis, and metabolite annotation, aiming to provide accessible solutions for metabolomics practitioners [43].

Big data poses challenges in quality, discovery, storage, security, analytics, and a talent shortage. Extracting useful information from dark data is a significant challenge due to veracity issues, vast data volumes, and a lack of skilled professionals, hindering effective utilization.

In this research article we have made an attempt to tackle dark data challenges by introducing an advanced extraction method. The AdaBoost and Random Forest classifiers are used to optimize the accuracy in accident prediction. Using dark data unlocks transformative potential, providing precise predictions for decision-makers and urban planners, establishing a precedent for revealing latent insights in escalating data volumes.

The paper is organized as follows: research methodology (Section 3), results and discussion (Section 4) and conclusion and future work (Section 5).

III. METHODOLOGY

The research employs a comprehensive stepwise methodology as shown in Figure 9, starting with an analysis of existing dark data research, followed by proposing an extraction method, implementing predictive techniques, and evaluating results. Python Jupyter Notebook is used for data analysis, involving data wrangling for cleaning and visualization for enhanced understanding. Prediction utilizes Random Forest and AdaBoost classifiers with dataset splitting and variable definition. Results are validated through performance comparison, assessing accuracy, precision, recall, and F1 score. The study addresses challenges in dark data insight extraction and evaluates machine learning classifier performance in outcome prediction. The stepwise whole process of the methodology is also sown in Algorithm 1.

A. Data

In our proposed research work, we select ignorable data (Dark Data), which poses challenges for extracting information and making informed decisions.

B. Tool or Application

The analysis of dark data is performed using Python Jupyter Notebook, leveraging the high-level programming capabilities of Python. Jupyter is an open-source web-based application.

C. Jupyter Notebook

Jupyter Notebook serves as an open-source web application enabling the creation and sharing of documents containing real-time code, equations, visualizations, and narrative text.

D. Data Wrangling

Data wrangling involves the cleaning of data, addressing issues such as unnecessary values and null items. Key functions include HeatMap(), Map(), Apply(), and ApplyMap() for efficient data cleaning.

E. Visualization and Plotting

Data plotting and visualization transform data into useful patterns and shapes, enhancing overall understanding and highlighting valuable information.

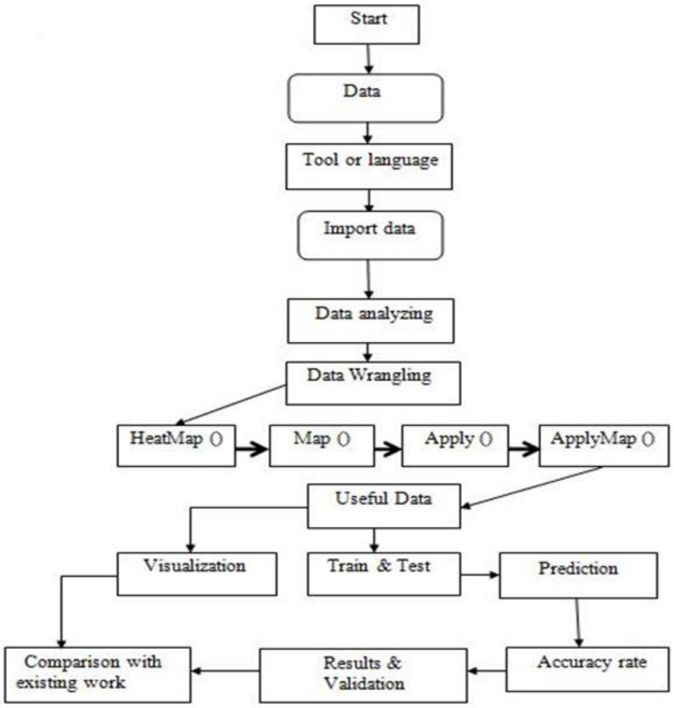
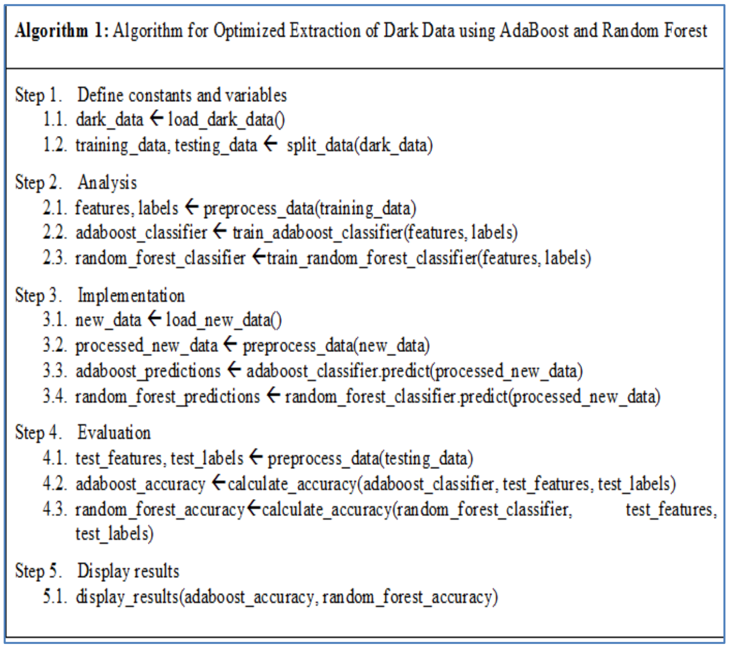


Figure 9. *Block diagram of the proposed work*



Algorithm 1: Algorithm for Optimized Extraction of Dark Data using AdaBoost and Random Forest

```
Step 1.  Define constants and variables
    1.1. dark_data ← load_dark_data()
    1.2. training_data, testing_data ← split_data(dark_data)
Step 2.  Analysis
    2.1. features, labels ← preprocess_data(training_data)
    2.2. adaboost_classifier ← train_adaboost_classifier(features, labels)
    2.3. random_forest_classifier ←train_random_forest_classifier(features, labels)
Step 3.  Implementation
    3.1. new_data ← load_new_data()
    3.2. processed_new_data ← preprocess_data(new_data)
    3.3. adaboost_predictions ← adaboost_classifier.predict(processed_new_data)
    3.4. random_forest_predictions ← random_forest_classifier.predict(processed_new_data)
Step 4.  Evaluation
    4.1. test_features, test_labels ← preprocess_data(testing_data)
    4.2. adaboost_accuracy ←calculate_accuracy(adaboost_classifier, test_features, test_labels)
    4.3. random_forest_accuracy←calculate_accuracy(random_forest_classifier, test_features,
         test_labels)
Step 5.  Display results
    5.1. display_results(adaboost_accuracy, random_forest_accuracy)
```

F. Random Forest Classifier

A powerful machine learning classification algorithm, the Random Forest Classifier, employs ensemble learning to generate predictions by aggregating multiple decision trees as shown in Figure 10.

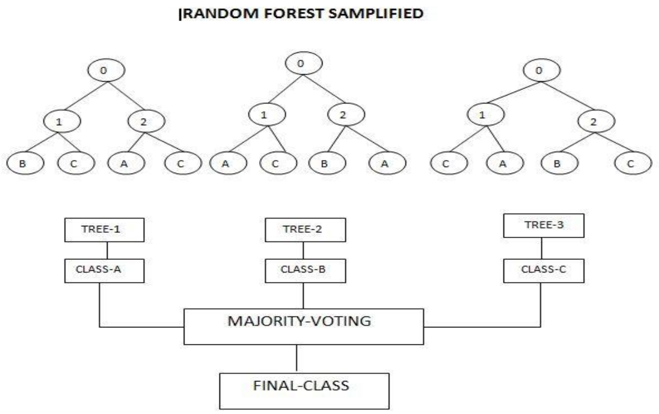


Figure 10. *Overall view of Random Forest*

G. The AdaBoost Classifier

The AdaBoost classifier, a robust ensemble learning technique, is a focal point for analysis. Known for boosting the performance of weak learners, it is systematically evaluated for effectiveness and predictive accuracy in classifying accident information in urban and rural areas.

H. Train and Test

The process of training and testing is a fundamental step in machine learning, enabling the evaluation of model performance on unseen data. In the context of the research study, the Random Forest Classifier and AdaBoost Classifier are employed for prediction, and the dataset is divided into training and testing sets to assess the effectiveness of these classifiers.

1. *Dataset Splitting*

Before delving into the training and testing process, the dataset is divided into two subsets: the training set and the testing set. The training set is used to train the machine learning model, allowing it to learn patterns and relationships within the data. The testing set, on the other hand, is reserved for assessing how well the model generalizes to new, unseen data.

2. *Defining Independent (X) and Dependent (Y) Variables:*

In the context of supervised machine learning, datasets are typically composed of independent variables (features) denoted as 'X' and a dependent variable (target or label) denoted as 'Y.' The goal is to predict 'Y' based on the patterns observed in 'X.'

a) Independent Variables (X)

These are the features or attributes in the dataset that the model uses for prediction. In the context of the research study, features related to dark data or other relevant parameters may be considered as independent variables.

b) Dependent Variable (Y)

This is the variable, the model aims to predict. In the context of the research study, it could be a binary classification (e.g., accident occurrence in urban or rural areas) or another relevant prediction task.

3. *Utilizing the Random Forest Classifier*

a) Decision Trees in Random Forest

The Random Forest Classifier is an ensemble learning method that utilizes multiple decision trees for prediction. Each decision tree is trained on a random subset of the dataset and makes independent predictions. The final prediction is then determined by aggregating the individual predictions, often through a voting mechanism.

b) Training the Random Forest Classifier

The training process involves providing the Random Forest Classifier with the training set (X_train, Y_train). The classifier learns the patterns and relationships within the training data to make predictions.

c) Testing with the Testing Set

The testing set (X_test) is then fed into the trained Random Forest Classifier to evaluate its performance. The classifier's predictions are compared against the actual values (Y_test) to assess accuracy, precision, recall, and other performance metrics.

4. *Utilizing the AdaBoost Classifier*

a) Boosting in AdaBoost

AdaBoost, short for Adaptive Boosting, is another ensemble learning technique. It focuses on improving the performance of weak learners (classifiers that perform slightly better than random chance) by giving more weight to misclassified instances.

b) Training the AdaBoost Classifier

Similar to the Random Forest, the AdaBoost Classifier is trained on the training set (X_train, Y_train). The classifier iteratively adjusts the weights of misclassified instances to improve overall performance.

c) Testing with the Testing Set

The testing set (X_test) is then used to evaluate the AdaBoost Classifier's predictions. The classifier's performance is assessed by comparing its predictions with the actual values (Y_test).

The training and testing process allows assessing how well the Random Forest and AdaBoost classifiers generalize to unseen data. By utilizing different subsets of the dataset for training and testing, the effectiveness and predictive power of the classifiers can be evaluated, providing insights into their performance on real-world scenarios.

I. Prediction

After splitting the data and utilizing the Random Forest and AdaBoost classifiers, predictions are generated, and accuracy is assessed.

J. Result Validation

The results of our proposed research study are evaluated and validated, comparing the results of the AdaBoost classifiers and random forest classifier against existing tools or applications for performance assessment.

K. Prediction on the testing set

After splitting the data and utilizing the Random Forest and AdaBoost classifiers, predictions are generated, and accuracy is assessed. Once the models are trained on the data, they are used to make predictions on the testing set. The classifiers assign predicted outcomes to the previously unseen instances based on the patterns they learned during the training phase.

The accuracy of the predictions is assessed by comparing the predicted outcomes with the actual outcomes in the testing set. This involves calculating metrics such as accuracy, precision, recall, and F1 score, depending on the nature of the predictive task (classification, regression, etc.).

L. Result Validation

The results of our proposed research study are evaluated and validated, comparing the results of the AdaBoost classifiers and Random Forest classifier against existing tools or applications for performance assessment.

The performance of the AdaBoost classifiers and Random Forest classifier is rigorously evaluated. This involves assessing their predictive accuracy, robustness, and efficiency in handling the specific characteristics of the dataset.

IV. Results and discussion

In the results and discussion section, the study repurposes UK accident dark data with Python and Jupyter Notebook. AdaBoost and Random Forest classifiers achieve high pattern recognition precision, evidenced by confusion matrix and performance measures. Pioneering dark data application in accident prediction, the research emphasizes transformative potential for refining models and advancing decision-making by leveraging untapped datasets.

A. *Data Set*

In the initial phase of our proposed work, existing research is thoroughly analyzed through a research study. We then select ignore data (dark data) or unusable data for in-depth analysis. The data is sourced from Kaggle, a prominent data scientist community, offering a vast repository of open-source data. Our dataset comprises UK accident information for both urban and rural areas, with a substantial volume of 629 MB in a CSV file. Although rich in information, the dataset is challenging to analyze and extract valuable insights due to its complex nature.

B. *Tool or Language*

Python, a high-level programming language, is employed for analyzing dark data using Jupyter Notebook, an open-source web-based application. Python is chosen for its ease of use, shorter coding time, and the capability to handle unlimited data processing. Jupyter Notebook further facilitates real-time code execution, equation handling, visualizations, and narrative text creation as shown in Figure 11.

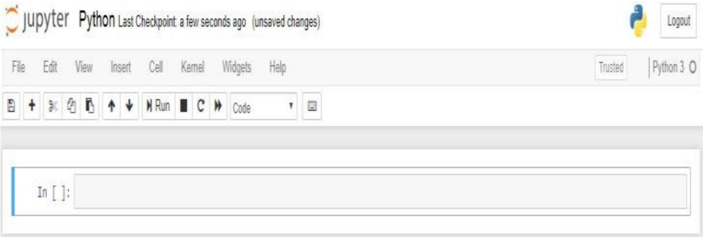
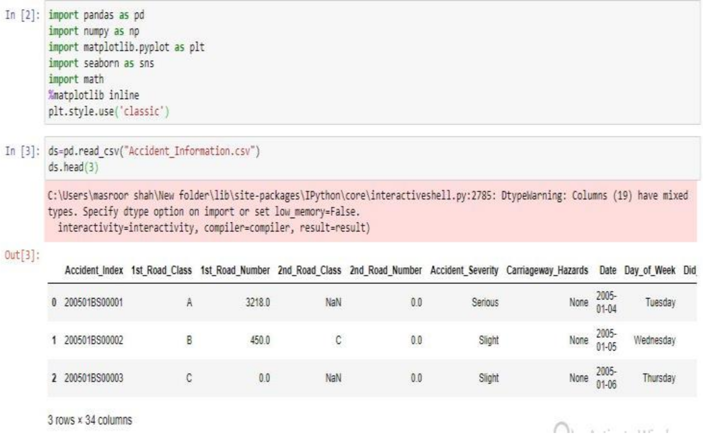


Figure 11. *Python Jupyter Notebook*

C. *Import Data*

Data is imported into Python Jupyter Notebook using various libraries, including pandas, NumPy, matplotlib, seaborn, and math. These libraries play a crucial role in importing, reading, and analyzing the data as shown Figure 12.



```
In [2]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        import math
        %matplotlib inline
        plt.style.use('classic')

In [3]: ds=pd.read_csv("Accident_Information.csv")
        ds.head(3)

        C:\Users\masroor shah\New folder\lib\site-packages\IPython\core\interactiveshell.py:2785: DtypeWarning: Columns (19) have mixed
        types. Specify dtype option on import or set low_memory=False.
          interactivity=interactivity, compiler=compiler, result=result)

Out[3]:
```

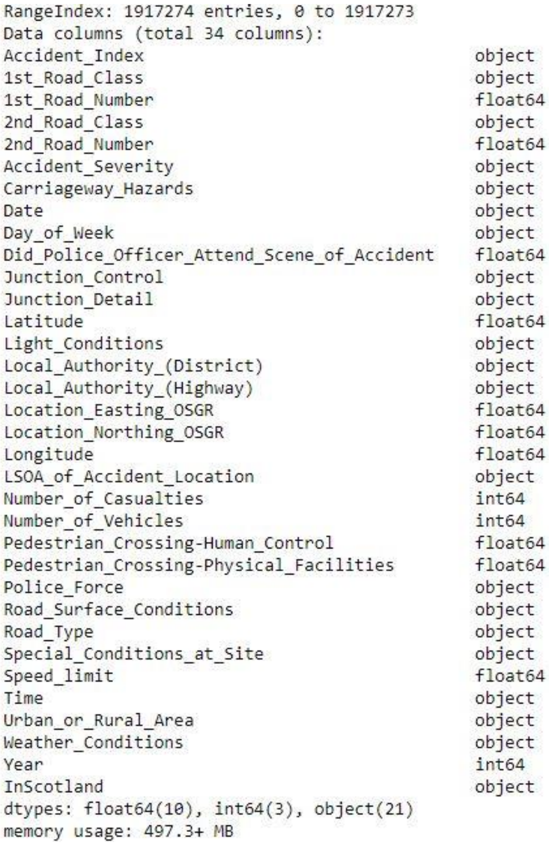
	Accident_Index	1st_Road_Class	1st_Road_Number	2nd_Road_Class	2nd_Road_Number	Accident_Severity	Carriageway_Hazards	Date	Day_of_Week	Did
0	200501BS00001	A	3218.0	NaN	0.0	Serious	None	2005-01-04	Tuesday	
1	200501BS00002	B	450.0	C	0.0	Slight	None	2005-01-05	Wednesday	
2	200501BS00003	C	0.0	NaN	0.0	Slight	None	2005-01-06	Thursday	

3 rows × 34 columns

Figure 12. *Importing and Read Data*

D. *Data Analyzing*

Data analysis is a pivotal phase for understanding the dataset. Information about the data, such as the range of indices, total number of columns, and data types, is obtained to enhance comprehension.



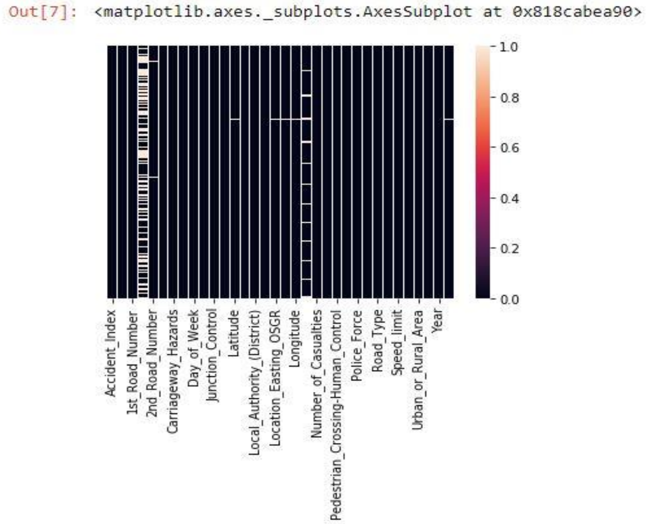
```
RangeIndex: 1917274 entries, 0 to 1917273
Data columns (total 34 columns):
Accident_Index                                object
1st_Road_Class                                object
1st_Road_Number                               float64
2nd_Road_Class                                object
2nd_Road_Number                               float64
Accident_Severity                             object
Carriageway_Hazards                           object
Date                                          object
Day_of_Week                                   object
Did_Police_Officer_Attend_Scene_of_Accident   float64
Junction_Control                              object
Junction_Detail                               object
Latitude                                      float64
Light_Conditions                              object
Local_Authority_(District)                    object
Local_Authority_(Highway)                     object
Location_Easting_OSGR                         float64
Location_Northing_OSGR                        float64
Longitude                                     float64
LSOA_of_Accident_Location                     object
Number_of_Casualties                          int64
Number_of_Vehicles                            int64
Pedestrian_Crossing-Human_Control             float64
Pedestrian_Crossing-Physical_Facilities       float64
Police_Force                                  object
Road_Surface_Conditions                       object
Road_Type                                     object
Special_Conditions_at_Site                    object
Speed_limit                                   float64
Time                                          object
Urban_or_Rural_Area                           object
Weather_Conditions                            object
Year                                          int64
InScotland                                    object
dtypes: float64(10), int64(3), object(21)
memory usage: 497.3+ MB
```

Figure 13. *Information of Data for analysis*

The dataset, representing UK accident information in urban and rural areas, consists of 32 columns and 1,917,274 rows as shown in Figure 13. Data wrangling is performed to clean the data, followed by graphical representation for improved understanding.

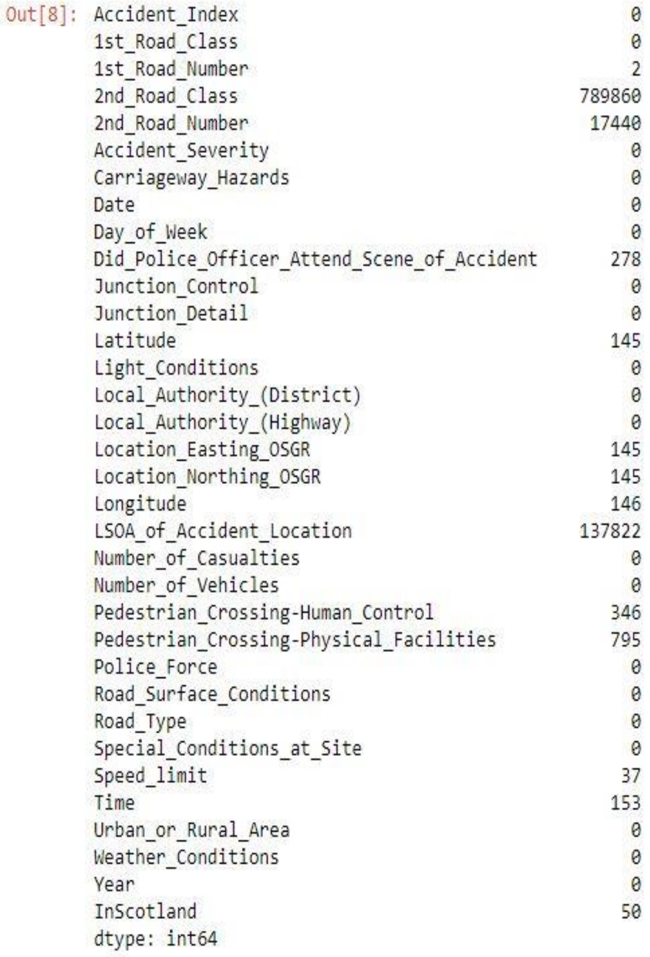
E. *Data Wrangling*

Data wrangling, or data cleaning, involves addressing unnecessary values and null items as shown in Figure 14 and Figure 15. Heatmap(), Map(), Apply(), and ApplyMap() functions are employed for effective data cleaning.



```
Out[7]: <matplotlib.axes._subplots.AxesSubplot at 0x818cabea90>
```

Figure 14. *HeatMap for effective data cleaning*



```
Out[8]: Accident_Index                                      0
        1st_Road_Class                                      0
        1st_Road_Number                                     2
        2nd_Road_Class                                 789860
        2nd_Road_Number                                 17440
        Accident_Severity                                   0
        Carriageway_Hazards                                 0
        Date                                                0
        Day_of_Week                                         0
        Did_Police_Officer_Attend_Scene_of_Accident       278
        Junction_Control                                    0
        Junction_Detail                                     0
        Latitude                                          145
        Light_Conditions                                    0
        Local_Authority_(District)                          0
        Local_Authority_(Highway)                           0
        Location_Easting_OSGR                             145
        Location_Northing_OSGR                            145
        Longitude                                         146
        LSOA_of_Accident_Location                      137822
        Number_of_Casualties                                0
        Number_of_Vehicles                                  0
        Pedestrian_Crossing-Human_Control                 346
        Pedestrian_Crossing-Physical_Facilities           795
        Police_Force                                        0
        Road_Surface_Conditions                             0
        Road_Type                                           0
        Special_Conditions_at_Site                          0
        Speed_limit                                        37
        Time                                              153
        Urban_or_Rural_Area                                 0
        Weather_Conditions                                  0
        Year                                                0
        InScotland                                         50
        dtype: int64
```

Figure 15. *Unnecessary and Null Values in Data*

```
Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x81ff63da20>
```

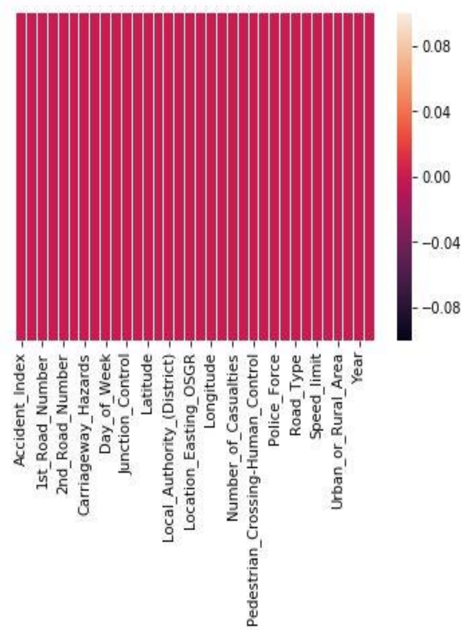
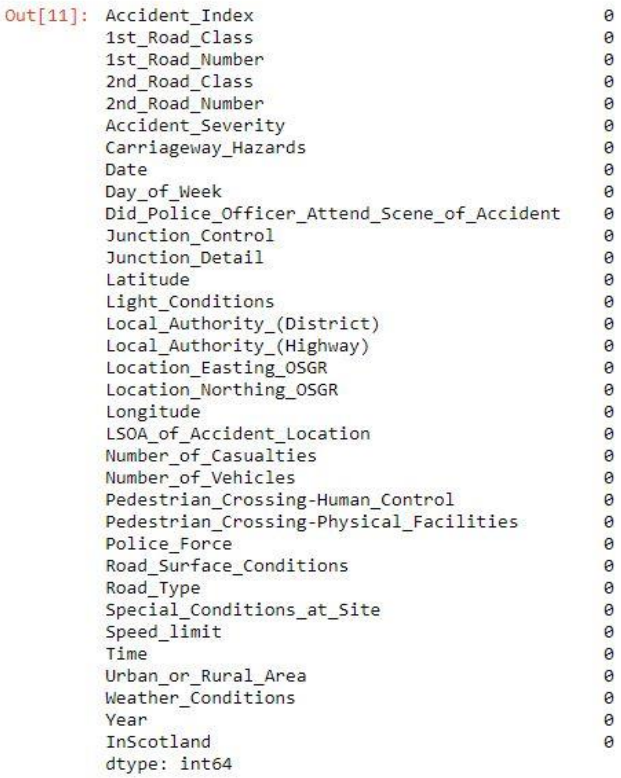


Figure 16. *HeatMap of Clean Data*



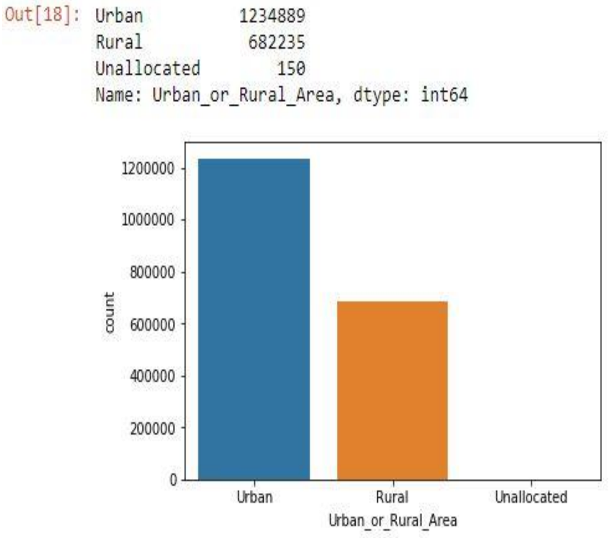
```
Out[11]: Accident_Index                                 0
         1st_Road_Class                                 0
         1st_Road_Number                                0
         2nd_Road_Class                                 0
         2nd_Road_Number                                0
         Accident_Severity                              0
         Carriageway_Hazards                            0
         Date                                           0
         Day_of_Week                                    0
         Did_Police_Officer_Attend_Scene_of_Accident    0
         Junction_Control                               0
         Junction_Detail                                0
         Latitude                                       0
         Light_Conditions                               0
         Local_Authority_(District)                     0
         Local_Authority_(Highway)                      0
         Location_Easting_OSGR                          0
         Location_Northing_OSGR                         0
         Longitude                                      0
         LSOA_of_Accident_Location                      0
         Number_of_Casualties                           0
         Number_of_Vehicles                             0
         Pedestrian_Crossing-Human_Control              0
         Pedestrian_Crossing-Physical_Facilities        0
         Police_Force                                   0
         Road_Surface_Conditions                        0
         Road_Type                                      0
         Special_Conditions_at_Site                     0
         Speed_limit                                    0
         Time                                           0
         Urban_or_Rural_Area                            0
         Weather_Conditions                             0
         Year                                           0
         InScotland                                     0
         dtype: int64
```

Figure 17. *Result of Clean Data*

The data is successfully cleaned and unnecessary and null values are removed, resulting in a usable dataset as shown in Figure 16 and Figure 17.

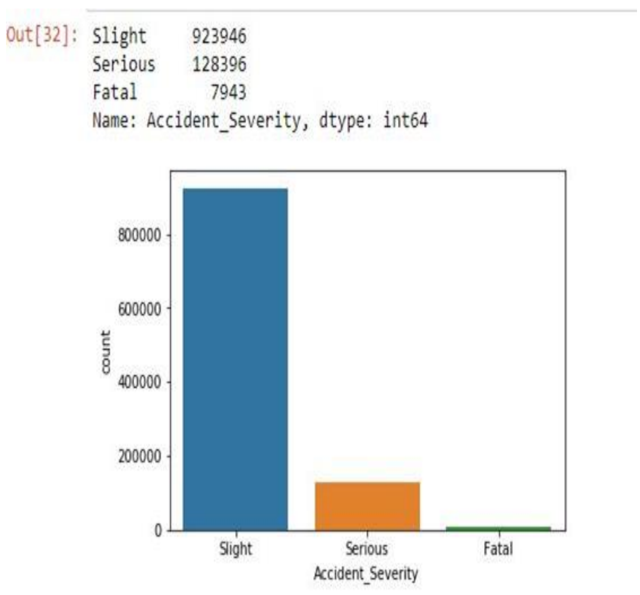
F. Data Visualization

Data plotting and visualization aid in transforming data into useful patterns and shapes. Visualization is performed to highlight useful information, such as the total number of accidents in urban and rural areas and accident severity as shown in Figure 18.



```
Out[18]: Urban          1234889
         Rural           682235
         Unallocated        150
         Name: Urban_or_Rural_Area, dtype: int64
```

Figure 18. *Visualization of Urban and Rural Accidents*



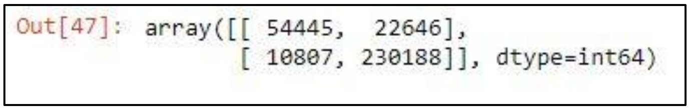
```
Out[32]: Slight     923946
         Serious    128396
         Fatal        7943
         Name: Accident_Severity, dtype: int64
```

Figure 19. *Visualization of Accident Severity*

The visualizations provide valuable insights, revealing that a majority of accidents occur in urban areas. The dataset is further analyzed for accident severity, distinguishing between slight, serious, and fatal accidents as shown in Figure 19.
The data is now useful for decision-making, showcasing the potential of extracting meaningful information from seemingly unusable dark data. The focus shifts to predicting the average accuracy rate of accidents in urban or rural areas based on UK accident information, aiding in decision-making for accident-prone areas.

G. Confusion Matrix

Confusion matrix is used for to calculate the prediction accuracy score. So, we first import confusion matrix library from Sklearn, matric to generate the all data labels in form of matrix as shown in Figure 20.



```
Out[47]: array([[ 54445,  22646],
                [ 10807, 230188]], dtype=int64)
```

Figure 20. *Data Label of Confusion Matrix*

Now using seaborn library to present the confusion matrix in form of heatmap is shown in Figure 21.

H. Performance Measures

In evaluating the performance of the classification models for accidents in urban and rural areas, relevant metrics are established:
The formulae used for calculation are as follows:

Accuracy (AC) = (TP + TN) / (TP + TN + FP + FN)……..... (1)
Precision (PR) = TP / (TP + FP)…………………………….. (2)
Recall (RE) = TP / (TP + FN)…………………..……..…….. (3)
F1 Score = 2 × (PP × RE) / (PR + RE)…………..………..…. (4)

Accuracy: Measures the overall correctness of predictions.
Precision: Evaluates the accuracy of positive predictions.
Recall: Assesses the ability to capture all relevant instances.
F1-Score: Combines precision and recall into a single value, offering a comprehensive performance evaluation.
The average accuracy score is determined through a comprehensive analysis of accuracy, recovery rates, and F1 scores. Precision accuracy and recall, pivotal in assessing model performance, represent the percentage of relevant results and the proportion of correctly classified relevant results, respectively. The average precision and recall scores encapsulated in the F1 score provide a holistic measure of the model's effectiveness. These calculations are derived from the confusion matrix.

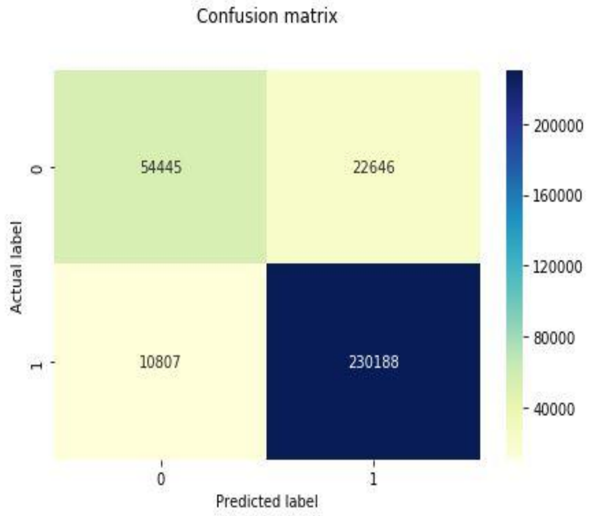


Figure 21. *Confusion matrix*

I. Predicted Results

The final result of our prediction and their values is show in Figure 22.

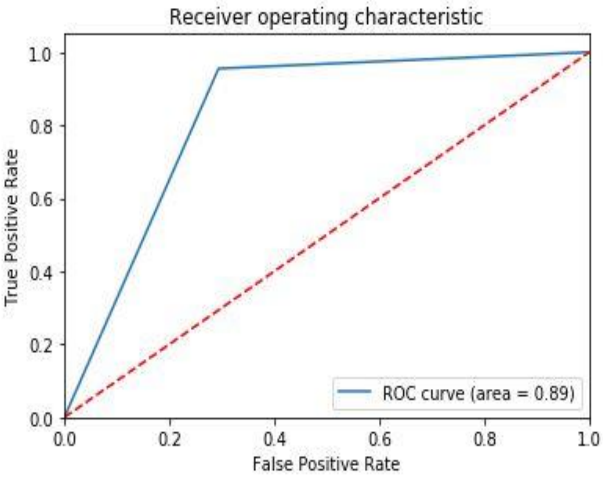
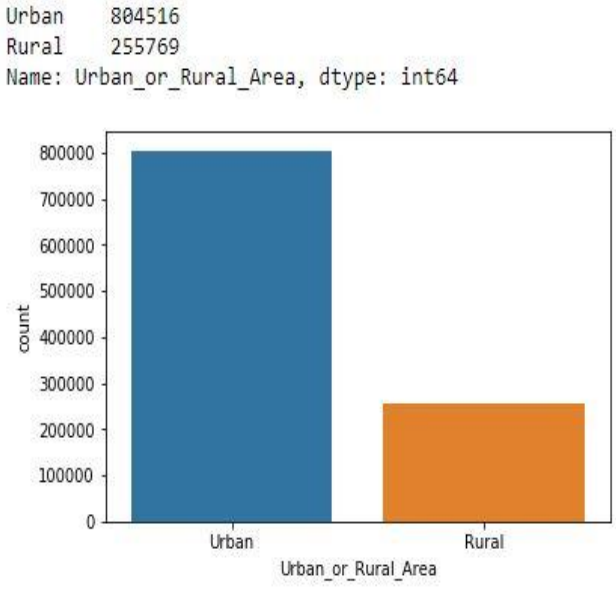


Figure 22. *Frequency of Accuracy Rate*



```
Urban    804516
Rural    255769
Name: Urban_or_Rural_Area, dtype: int64
```

Figure 23. *Predicted Results*

The Figure 23 shows the final results for prediction of Urban and Rural area. That show the total number of accidents in Urban is 804516 and Rural is 255769. It is means that accident accuracy is high in Urban than Rural. Our prediction is shows that Urban held large accidents rather than Rural.

J. Result Validation

For result validation we used cross validation test from Sklearn library to validate the module there factored, classes and all function for train and test dataset. They validate the random forest classifier and fit data into classifier to generate classification reports. If total average classification report equal to prediction score, then results of prediction are accurate. The following is the classification report. By cross validation the average classification result is 89.48% as shown in Table 1. So that show the final accuracy rate 89.48% of prediction is right.

Table 1. *Classifies results comparisons*

Classifiers	Precision (%)	Recall (%)	F-Score (%)	Accuracy (%)
AdaBoost classifier	86	62	72	78.4
Random Forest classifier	71	83.44	76	89.48

1. Unlocking hidden insights using dark data for enhanced accident prediction

In the realm of data-driven decision-making, the burgeoning concept of harnessing dark data, previously untapped and often overlooked information, presents an opportunity for refining predictive models and elevating accuracy rates. This research delves into the utilization of unused data, colloquially known as dark data, to extract valuable insights, thereby contributing to the optimization of accident prediction models. The focal objective is to enhance decision-making processes and achieve a more accurate prediction of accident occurrences in urban and rural settings.

2. Classifier Performance Evaluation

Two prominent classifiers, namely the AdaBoost and Random Forest classifiers, undergo an in-depth evaluation based on precision, recall, F-Score, and overall accuracy. The precision metric gauges the accuracy of positive predictions, while recall assesses the ability to capture all positive instances. F-Score harmonizes precision and recall, providing a balanced measure of a classifier's performance. The overall accuracy denotes the classifier's proficiency in accurately classifying instances across the entire dataset.

3. AdaBoost Classifier Evaluation

The AdaBoost classifier exhibits a precision of 86%, indicating a substantial accuracy in identifying positive instances among the predicted positives as shown in Table 1. However, a relatively lower recall of 62% suggests a proportionately higher likelihood of missing true positive instances. This trade-off is reflected in the F-Score of 72%, portraying a balanced performance. The overall accuracy, standing at 78.4%, underscores the classifier's proficiency in accurate classification, albeit with room for improvement in capturing all positive instances.

4. Random Forest Classifier Evaluation

Conversely, the Random Forest classifier showcases a precision of 71%, denoting a significant proportion of correctly identified positive instances among the predicted positives. Impressively, the recall metric at 83.44% signifies a meticulous capture of true positive instances, highlighting the classifier's sensitivity. The resulting F-Score of 76% reflects a harmonious balance between precision and recall. Notably, the classifier achieves a noteworthy accuracy of 89.48%, positioning it as a formidable contender in accurately predicting accident occurrences in both urban and rural settings.

5. Contribution to the Field

This research significantly contributes to the field by pioneering the application of dark data in accident prediction models, thereby advancing the paradigm of data utilization. The incorporation of previously unexplored information enhances the granularity and comprehensiveness of predictive models, consequently elevating accuracy rates.

The improvement from an accuracy rate of 78.4% to 89.48% underscores the efficacy of leveraging dark data. The substantial enhancement in accuracy, translating to a more precise prediction of accidents in urban and rural contexts, holds paramount implications for decision-makers and urban planners alike. This research not only optimizes existing predictive models but also sets a precedent for future endeavors seeking to unlock latent potential in unutilized datasets.

This research underscores the transformative potential of utilizing dark data for refining accident prediction models. The precision, recall, and accuracy metrics serve as quantitative evidence of the tangible benefits derived from incorporating previously unexplored information. As we venture into an era of data abundance, harnessing the latent insights within dark data emerges as a pivotal strategy for enhancing decision-making processes and advancing the accuracy of predictive models in diverse domains.

V. Conclusion and future work

A. Conclusion

In conclusion, this research article provides a comprehensive exploration of big data's intricate landscape, emphasizing its foundational role across diverse domains and highlighting challenges in data quality, storage, security, and analytics. The study introduces and underscores the untapped potential of dark data within organizational datasets, acknowledging its escalating complexity. Recognizing the imperative need for strategic interventions, the research navigates the complexities of dark data, examining its definition, management, and proposing optimized solutions. The introduced extraction method, utilizing AdaBoost and Random Forest classifiers, showcases transformative potential for precise predictions and uncovering latent insights. With an 86% precision in pattern recognition for AdaBoost and 89.48% accuracy for the Random Forest classifier, this study pioneers dark data application in accident prediction, setting a precedent for leveraging untapped datasets. The proposed approach encompasses analysis, implementation, and evaluation phases, addresses challenges in extracting insights from dark data, emphasizing transformative potential for refining predictive models and advancing decision-making processes.

B. Future Work

Future directions include extending the application of optimized extraction methods to various domains, fostering interdisciplinary collaboration, and exploring ethical considerations in harnessing the full potential of big and dark data.

References

[1] Oussous, A., Benjelloun, F.-Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. Journal of King Saud University – Computer and Information Sciences, 30. https://doi.org/[DOI]

[2] Liu, Y. (2018). Research on the application of big data in academic libraries. IEEE International Conference on Information Technology and Business Science. https://doi.org/10.1109/ICITBS. IEEE2018

[3] Khan, M., Wu, X., Xu, X., & Dou, W. (2017). Big Data Challenges and Opportunities in the Hype of Industry 4.0. SAC Symposium Big Data Networking Track, IEEE ICC 2017.

[4] Cao, R., & Gao, J. (2018). Research on Reliability Evaluation of Big Data System. 3rd IEEE International Conference on Cloud Computing and Big Data Analysis 2018.

[5] Xu, Z. (2018). Research on Enterprise Knowledge Unified Retrieval Based On Industrial Big Data. Sixth International Conference on Advanced Cloud and Big Data 2018.

[6] Kumar, S., & Singh, M. (2019). Big Data Analytics for Healthcare Industry: Impact, Applications, and Tools. International Journal of Big Data Management, 2(1), 48–57. https://doi.org/10.26599/BDMA.2018.9020031

[7] L'heureux, A. (2017). Machine Learning with Big Data: Challenges and Approaches. IEEE Access. https://doi.org/10.1109/ACCESS.2696365.2017

[8] Lv, Z., Song, H., Basanta-Val, P., Steed, A., & Jo, M. (2017). Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics. IEEE Transactions on Industrial Informatics, 13(4), 1891-1899. doi: 10.1109/TII.2017.2650204

[9] Shobanadevi, A., & Maragatham, G. (2017). Data Mining Techniques for IoT and Big Data – A Survey. Proceedings of the International Conference on Intelligent Sustainable Systems. https://doi.org/10.1109/ICISS. IEEE 2017

[10] Perikos, I., & Hatzilygeroudis, I. (2018). A Framework for Analyzing Big Social Data and Modelling Emotions in Social Media. Fourth International Conference on Big Data Computing Service and Applications, IEEE 2018.

[11] Ali, A. R. (2018). Real-Time Big Data Warehousing and Analysis Framework. In 3rd International Conference on Big Data Analysis. IEEE.

[12] Sezer, O. B., Dogdu, E., & Ozbayoglu, A. M. (2017). Context Aware Computing, Learning and Big Data in Internet of Things: A Survey. IEEE. DOI 10.1109/JIOT.2017.2773600.

[13] Katal, A., Wazid, M., & Goudar, R. H. (2013). Big Data: Issues, Challenges, Tools and Good Practices. IEEE. ISBN 978-1-4799-0192-0.

[14] Costa, C., & Zeinalipour-Yazti, D. (2018). Telco Big Data: Current State & Future Directions. In 19th IEEE International Conference on Mobile Data Management.

[15] Trivedi, B., & Gokulnath, K. (2017). Research on dark data analysis to reduce complexity in big data. International Education & Research Journal, 3(5).

[16] Cafarella, M., Ilyas, I. F., Kornacker, M., Kraska, T., & Ré, C. (2016). Dark Data: Are We Solving the Right Problems? In IEEE. ISBN 978-1-5090-2020-1.

[17] Rosa, A., Chen, L. Y., & Binder, W. (2015). Understanding the Dark Side of Big Data Clusters: An Analysis beyond Failures. In IFIP International Conference on Dependable Systems and Networks.

[18] Mwiti Kevin, N., Munyi Wanyaga, F., Kibaara, D., Dinda, W. A., & Ngatia, J. K. (2016). Dark data: Business Analytical tools and Facilities for illuminating dark data. Scientific Research Journal (SCIRJ), 4(4).

[19] Gartner. (2014). Turning dark data into smart data: File analysis innovation delivers an understanding unstructured dark data. Innovation Insight.

[20] Pandey, A. V., & Bertini, E. (2016). TextTile: An Interactive Visualization Tool for Seamless Exploratory Analysis of Structured Data and Unstructured Text. IEEE. DOI 10.1109/TVCG..2598447.

[21] Dabbechi, H., Haddar, N., Ben Abdallah, M., & Haddar, K. (2017). A Unified Multidimensional Data Model From Social Networks For Unstructured Data Analysis. In ACS 14th International Conference on Computer Systems and Applications. IEEE.

[22] Khan, M., Wu, X. T., Xu, X. L., & Dou, W. (2017). Big Data Challenges and Opportunities in the Hype of Industry 4.0. In State Key Laboratory for Novel Software Technology. Nanjing University, Nanjing, P.R. China. IEEE.

[23] Tanwar, M., Duggal, R., Khatri, S. K. (2015). Unravelling Unstructured Data: A Wealth of Information in Big Data. In Amity Institute of Information Technology. Amity University Uttar Pradesh, Noida, India. IEEE.

[24] Sato, S., Kayahara, A., & Imai, S. I. (2017). Unstructured Data Treatment for Big Data Solutions. In Panasonic Corporation. IEEE.

[25] Håkonsson, and Carroll, T. (2016). Is there a dark side of Big Data. Journal of Organization Design, 5(5). DOI 10.1186/s41469-016-0007-5.

[26] Heidorn, P. B., Stahlman, G. R., & Steffen, J. (2018). Astrolabe: Curating, Linking and Computing Astronomy's Dark Data. University of Arizona School of Information. American Astronomical Society.

[27] Zhang, D., Liao, C., Yan, W., Ta, R., & Zheng, W. (2017). Data Deduplication based on Hadoop. In Department of Computer Science School of Information Science and Engineering. Xiamen University. IEEE.

[28] Kolesnichenko, O., Yakovleva, D., Zhurenkov, O., Smorodin, G., Mazelis, L., & Kolesnichenko, Y. (2016). Text Big Data Analytics: exploring API opportunity. In Internet as Global storage – how to get the situation awareness from Dark Data. Security Analysis Bulletin. Moscow, Russia.

[29] Lugmayr, A., Stockleben, B., Scheib, C., & Mailaparampil, M. (2017). Cognitive big data: survey and review on big data research and its implications. What is really new in big data? http://dx.doi.org/10.1108/JKM-07-2016-0307.

[30] Wiebe, P. H., & Dickson, M. (2015). Bringing dark data into the light: A case study of the recovery of Northwestern Atlantic zooplankton data collected in the 1970s and 1980s. GeoResJ, 6.

[31] Gašpar, D., & Mabić, M. (2018). Light up the value of dark data. In 21st International Research/Expert Conference "Trends in the Development of Machinery and Associated Technology" TMT 2018, Karlovy Vary, Czech Republic, September 18-22, 2018.

[32] Saxena, S. (2018). Dark data and its future prospects. International Journal of Engineering Technology Science and Research (IJETSR), 5(1), ISSN 2394 – 3386.

[33] Johnson, J. N., Hanson, K. A., Jones, C. A., et al. (2018). Data sharing in neurosurgery and neurology journals. Cureus, 10(5), e2680. DOI: 10.7759/cureus.2680.

[34] Priyadarshy, S. (2018). Big data, smart data, dark data and open data: eGovernment of the future.

[35] Zhang, C., Shin, J., Ré, C., Cafarella, M. (2016). Extracting databases from dark data with DeepDive. Proceedings of the conference held in Palo Alto, CA, USA, June 26-July 01, 2016.

[36] Trajanov, D., Stojanov, R., Zdraveski, V., Kocarev, L. (2018). Dark data in Internet of Things (IoT): Challenges and opportunities. Proceedings of the 7th Small Systems Simulation Symposium 2018, Niš, Serbia, February 12-14, 2018.

[37] Kolesnichenko, O. (2016). How to get the situation awareness from dark data. Security Analysis Moscow.

[38] Sookhak, M., Yu, F. R., & Zomaya, A. Y. (2017). Auditing big data storage in cloud computing using divide and conquer tables. IEEE Transactions on Parallel and Distributed Systems, DOI: 10.1109/TPDS.2017.2784423.

[39] Shukla, S., & Mehta, R. (2017). Review on artificial bee colony algorithm on big data to find out required data sources. Information Technology department, Parul Institute of Engineering and Technology, Vadodara, E-ISSN No: 2454-9916, Volume 3, Issue 5, May 2017.

[40] Xiao, X., Qi, B., Liang, J., Tong, J., Deng, Q., & Chen, P. (2023). Enhancing LOCA Breach Size Diagnosis with Fundamental Deep Learning Models and Optimized Dataset Construction. Energies, 17(1), 159.

[41] Alam, M. J., Akter, N. J., Srabony, M. S. M. A., Aziz, A., & Ronak, N. A. (2023). Traffic Fatality Prediction Using Machine Learning Algorithms: Performance Analysis and Comparison Study (Doctoral dissertation, Sonargaon University (SU)).

[42] Yan, H., & Li, Y. (2023). A Survey of Generative AI for Intelligent Transportation Systems. arXiv preprint arXiv:2312.08248.

[43] Guo, J., Yu, H., Xing, S., & Huan, T. (2022). Addressing big data challenges in mass spectrometry-based metabolomics. Chemical Communications, 58(72), 9979-9990.

[44] Rajawat, A. S., Bedi, P., Goyal, S. B., Kautish, S., Xihua, Z., Aljuaid, H., & Mohamed, A. W. (2022). Dark web data classification using neural network. Computational Intelligence and Neuroscience, 2022.

[45] Brogaard, J., & Pan, J. (2022). Dark pool trading and information acquisition. The Review of Financial Studies, 35(5), 2625-2666.

[46] Aaen, J., Nielsen, J. A., & Carugati, A. (2022). The dark side of data ecosystems: A longitudinal study of the DAMD project. European Journal of Information Systems, 31(3), 288-312.

Authors

First Author – Masroor Shah, Department of Computer Science, Iqra National University, Peshawar, KPK, Pakistan,
Second & Correspondence Author – Dr. Fazal Malik, Department of Computer Science, Iqra National University, Peshawar, KPK, Pakistan,
Third Author – Abd Ur Rub, School of Electronics and Information, Northwestern Polytechnical University, Xi'an Shaanxi, China,
Fourth Author – Muhammad Suliman, Department of Computer Science, Iqra National University Peshawar, KPK, Pakistan,
Fifth Author – Irfan ullah, Department of Computer Science, Iqra National University, Peshawar, KPK, Pakistan,
Sixth Author – Sana Ullah, Department of Computer Science, Iqra National University, Peshawar, KPK, Pakistan,
Seventh Author – Romaan Khan, City University of Science and Information Technology Peshawar, KPK, Pakistan,
Eighth Author – Salman Alam, COMSATS University Islamabad (CUI), Islamabad, Pakistan,