JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 07 Issue 02
2024

*Research Article*
https://doi.org/10.56979/702/2024

# Dark Data in Accident Prediction: Using AdaBoost and Random Forest for Improved Accuracy

**Masroor Shah[1], Fazal Malik[1*], Muhammad Suliman[1], Noor Rahman[2], Irfan Ullah[1], Sana Ullah[1], Romaan Khan[3], and Salman Alam[4]**

[1]Department of Computer Science, Iqra National University Peshawar, Khyber Pakhtunkhwa (KPK), Pakistan.
[2]Fahad Bin Sultan University, Saudi Arabia.
[3]City University of Science and Information Technology Peshawar, KPK, Pakistan.
[4]COMSATS University Islamabad (CUI), Pakistan.
[*]Corresponding Author: Fazal Malik. Email: fazal.malik@inu.edu.pk

**Received:** February 28, 2024 **Accepted:** July 23, 2024 **Published:** September 01, 2024

**Abstract:** Dark data, or unused information included into routine activities, poses significant hurdles in the era of data-driven decision-making because of its volume and complexity. The goal of this publication is to increase the accuracy of accident prediction by proposing an efficient procedure for dark data extraction and analysis. Data extraction, classifier implementation, and performance evaluation are all done in a methodical manner by using AdaBoost and Random Forest classifiers. According to the results, the Random Forest classifier outperforms the AdaBoost classifier with an accuracy of 89.50%, compared to the former's 78.4%. These results highlight the potential of dark data to yield insightful information by demonstrating how well these classifiers improve accident prediction models. In addition to emphasizing the value of dark data for decision-makers and urban planners looking to improve prediction models and access hidden information, the study offers a methodology for using it. Our research highlights the increasing significance of dark data in enhancing decision-making procedures and forecast precision as data quantities increase.

**Keywords:** Big Data; Data Quality; Dark Data; Complexity of Dark Data; Accident Prediction.

## 1. Introduction

The introduction underscores data's fundamental role across diverse domains and delves into the challenges associated with big data, including quality, storage, security, and analytics. It highlights the concept of dark data, emphasizing its untapped potential within organizational datasets. The narrative points out the exponential growth and increasing complexity of both big and dark data, while the conclusion stresses the need for strategic interventions to fully harness the potential of big data for future applications and decision-making processes.

### 1.1. Data

Information is derived from data, which comes from a variety of sources including business, economic, social media, the Internet of Things (IoT), science, and sensor technologies. Each and every piece of data has intrinsically valuable information, and knowledge is created when this information is combined [1].

### 1.2. Big Data

Large amounts of complex data are produced in modern activities in a variety of fields, such as commerce, economics, social media, the Internet of Things, and science. What is usually referred to as big data is the aggregation and collecting of large and complicated data in everyday processes [1]. This dataset, which includes online, general, and conventional system data, is significantly impacted by sensor-derived, machine-generated, and social media data—especially from platforms like Facebook, Twitter, email, blogs, and other social media channels [2].

Large datasets, comprising unstructured and semi-structured data, voice recordings, videos, audios, and images, are produced both inside and outside of organizational boundaries as a result of advancements in data capture technologies. Large data management has historically proven difficult for companies because of limited technology, storage options, and expensive tool prices [3]. Traditional technology has significant challenges when it comes to critical concerns such data cleansing, analysis, processing, security, and information extraction from massive datasets [4]. But globally, new big data initiatives have opened up previously unheard-of prospects. In comparison to traditional technologies, they provide the newest frameworks, models, security measures, processing powers, storage capacities, and real-time analysis in addition to being more scalable, adaptable, and performance-focused [5].

1.3. Big Data Types

Figure 1 clarifies the various big data categories [6].

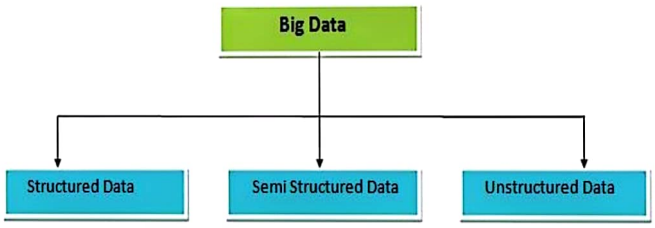


**Figure 1.** Big Data Types

*1.3.1. Organized Information*

The formal structure of data models connected to relational databases or other data tables is followed by structured data. Structured data is just information that has been given a predetermined format.

*1.3.2. Data that is Semi-Structured*

A variation of structured data that does not follow the formal structure of the data models connected to relational databases is called semi-structured data. It organizes hierarchical data into records and fields using tags or other semantic features, even in the absence of strict conformance.

*1.3.3. Data Without Structure*

Unstructured data is incompatible with conventional databases and does not have a clear internal organization. This is not the case with structured data that is kept in databases. Up to 80% of company data is classified as unstructured, and this percentage is increasing every year.

1.4. Big Data Technology's Features

Important features are critical to the investigation of big data technologies, as shown in Figure 2 [6].

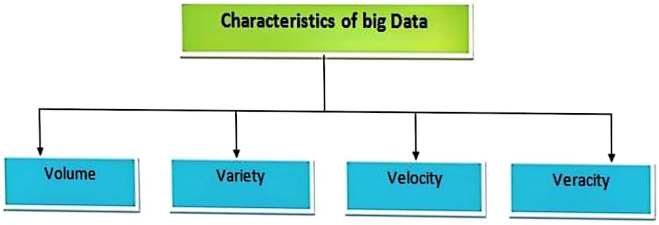


**Figure 2.** Characteristics of Big Data

*1.4.1. Capacity*

In the context of big data technologies, "volume" refers to the sheer amount of data, taking into account both the quantity and the storage needs. It is the ability to safely store data while doing daily tasks; data is measured in gigabytes (GB), terabytes (TB), megabytes (MB), and kilobytes (KB).

*1.4.2. 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.

*1.4.3. 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.

*1.4.4. Veracity*

Since there is a significant quantity of data involved, veracity emphasizes how crucial security is to big data systems. The shift to cloud storage highlights how important it is to protect data from outside attacks and unidentified entities.

*1.4.5. Value*

The most important feature of big data is its value, which comes after accuracy, speed, diversity, and volume. In an effort to maximize earnings and get a good return on investment, organizations analyze and use big data to their advantage. This emphasizes how big data may be strategically used to achieve business goals and financial success.

1.5. An 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 [7]:

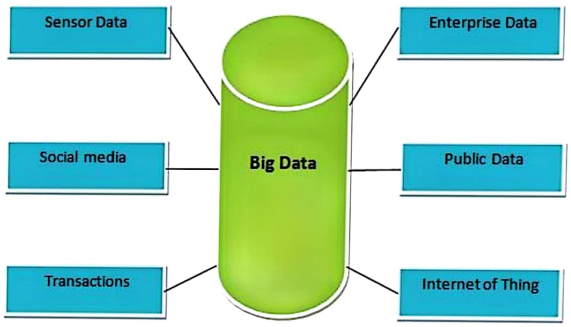


**Figure 3.** Big Data Examples

*1.5.1. Information from Sensors*

Information produced by equipment intended to recognize and react to events in the physical world is known as sensor data. 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.

*1.5.2. 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 [8].

*1.5.3. 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 [9].

*1.5.4. Enterprise Data*

Enterprise data is information that is shared by an organization's users in order to obtain critical advantages. Because of its extreme criticality, this data requires strict security protocols. Enterprise data management solutions are essential for maintaining communication between internal and external parties, protecting sensitive data, and making it easier to get data from a variety of applications.[10].

*1.5.5. Public Data*

Public data stands as a significant data source freely accessible to anyone, without legal restrictions on access, reuse, or redistribution. This includes government, national, local, and international datasets that contribute to diverse applications and analyses [8].

*1.5.6. Internet of Things (IoT)*

Due to developments in sensor and communication technologies, the Internet of Things (IoT) is a quickly growing subset of big data. A network of physical devices and objects having actuators, sensors,

and communication devices incorporated is referred to as the Internet of Things. Through network connection, these networked organizations gather and share data, facilitating an easy flow of information [11]. The variety of big data examples highlights the complexity of this sector, where efficient analysis and exploitation depend on a thorough grasp of different data kinds and sources

1.6. Challenges in Big Data Management

This scholarly discussion explores the complicated issues that are present in the field of big data, focusing on the difficulties that come with gathering, storing, processing, and using information. Six key issues are examined in the analysis: information discovery, storage, security, data analytics, and skills shortage. Every obstacle is examined in light of how it may affect organizations attempting to use big data for applications in the future [12].

*1.6.1. 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 [13].

*1.6.2. 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 [14].

*1.6.3. 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.

*1.6.4. 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.

*1.6.5. Data Analytics*

Big data analysis presents significant management, processing, and analytical issues due to its chaotic and disordered character. The complex characteristics of unstructured data make it difficult to derive significant insights, which hinder the efficient application of big data for future strategic objectives [46].

*1.6.6. Lack of Talent*

A major barrier to the efficient administration of large-scale data projects is the lack of qualified professionals, such as data scientists, analysts, and developers. Big data's volume requires knowledge of cutting-edge algorithms and methods, which makes it difficult for enterprises trying to understand the intricacies of big data analytics.

The difficulties posed by big data are complex and call for a range of strategic approaches, including strengthening information discovery strategies, improving data quality, finding effective storage solutions, putting strong security measures in place, developing highly qualified personnel, and more. For enterprises hoping to fully utilize big data for upcoming applications and decision-making procedures, addressing these obstacles is essential.

1.7. Dark Data

A subset of big data known as "dark data" is information that is gathered and processed on a regular basis within a company or organization, but is not used for critical future organizational gains. Even though some data is similar to big data in terms of danger and cost, certain data is nevertheless used in decision-making processes even though it isn't useful or valuable for corporate goals. Figure 4 shows the percentage of helpful versus dark data.

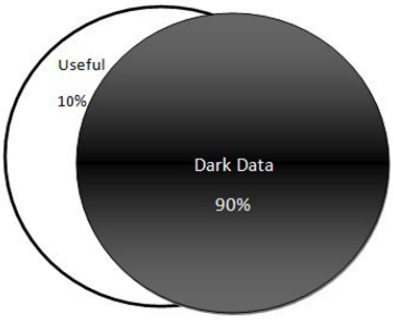


**Figure 4.** Dark Data Percentages Illustrated

Various sources, including traditional enterprises, social media, machines, 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 and is 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].

Corallo'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].

1.8. Overall Growth of Big and Dark Data

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.

*1.8.1. Organized Information*

Relational databases and other related tabular formats follow formal data models for structured data. Usually found in conventional database systems, it is distinguished by a certain data shape that adheres to preset structures.

*1.8.2. Semi-Structured Data*

Although semi-structured data has tags or markers that divide semantic parts, it differs from formal data models connected to relational databases. Although it does not strictly follow tabular patterns, it does enforce record and field hierarchies within the data

*1.8.3. Unstructured Data*

Unstructured data does not cleanly fit into typical data-bases and has an obvious internal structure. It is believed to make up much to 80% of corporate data, and its share is constantly rising. The absence of a recognizable structure in this kind of data presents difficulties.

*1.8.4. Data Warehouse*

An essential part of business intelligence is an enterprise data warehouse (EDW), sometimes known as a data warehouse. In order to facilitate reporting and data analysis, it acts as a central repository that unifies data from several sources.

*1.8.5. Social Media*

Massive volumes of data are produced by social media sites like Facebook and Twitter, including user interactions, emails, blogs, and other information. Using this data, analytics systems can better assess consumer sentiment, which helps with marketing and customer support initiatives.

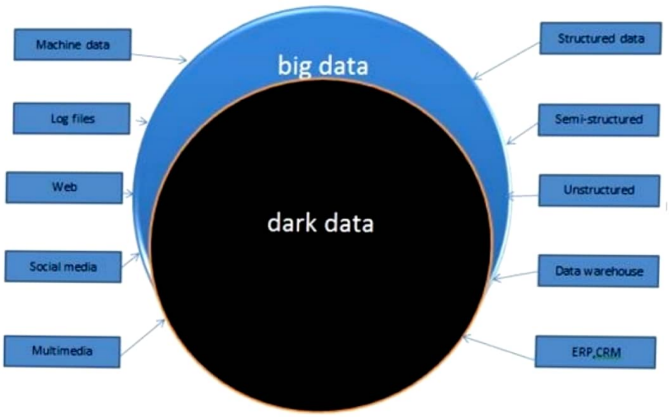


**Figure 5.** The Total Increase in Big and Dark Data

*1.8.6. Web Data*

Web data includes all information made available by the World Wide Web via online services. It facilitates the provision, sharing, dissemination, and consumption of information for a variety of purposes.

*1.8.7. Log File*

Events or communications between operating systems or various software components are recorded in a log file. This activity log is very helpful for maintaining records, tracking behavior, and making diagnoses.

*1.8.8. 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.

*1.8.9. 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.

*1.8.10. CRM and ERP*

Customer Relationship Management (CRM) and Enterprise Resource Planning (ERP) 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 era of data-driven decision-making, organizations are increasingly generating substantial data through daily transactions and advanced technologies. However, a significant challenge arises from dark data—untapped information that is not utilized in decision-making and often lacks structure. Addressing dark data involves developing effective policies and deploying advanced technologies to manage and comprehend its significance.

The usage of sensor technologies, machine-to-machine (M2M) connectivity, and the Internet of Things (IoT), which adds to the massive amounts of unstructured data, further accelerates the development of big data. This increase highlights the need for sophisticated computational tools and fast systems to handle large and varied datasets. Big data presents a number of difficulties, such as problems with quality, discovery, storage, security, and a lack of qualified personnel, all of which make it more difficult to draw useful conclusions from dark data.

In order to overcome the difficulties posed by dark data, this study presents an innovative extraction technique that makes use of Random Forest and AdaBoost classifiers. The study's particular contributions are as follows:

**Creation of an Optimal Extraction Method:** To improve the efficacy of prediction models, the study offers an optimum method for obtaining and exploiting dark data.

**Methodical Structure for Examination and Interpretation:** A methodical framework is created for the examination, application, and assessment of dark data extraction techniques. This framework offers a thorough method for handling the difficulties associated with dark data.

**Use of AdaBoost and Random Forest Classifiers:** In order to maximize accident prediction accuracy, the research makes use of AdaBoost and Random Forest classifiers. It has been shown that these classifiers considerably increase prediction accuracy when compared to conventional

**Enhanced Predictive Accuracy:** The accuracy of accident forecasts is significantly increased by using these sophisticated classifiers. This demonstrates how precise and useful insights from dark data may be transformed.

**Consequences for Urban Planners and Decision-Making:** The study emphasizes how crucial it is to use dark data to enhance prediction models, providing insightful information to help urban planners and decision-makers manage accident-prone areas.

This study advances the discipline by highlighting the crucial role that dark data plays in navigating the increasing amounts of data, establishing a precedent for using dark data to improve predictive accuracy and decision-making processes, and developing analytics capabilities.

The research paper will be divided into the following sections: Section 2 will provide a thorough analysis of the state-of-the-art methods currently in use; Section 3 will introduce the proposed framework and technique; Section 4 will present specific experimental results; and Section 5 will conclude with recommendations for further research.

## 2. Literature Review

The literature study explores the difficulties associated with dark data in big data, emphasizing the complications associated with its definition, management, and use. 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.

### 2.1. Review of the Unseen Dimensions of Dark Data

A subset of big data known as "dark data" includes information from a variety of industries, including social networking, commerce, economics, the Internet of Things, and the sciences. It comprises social media content, machine data, sensor data, general web data, and classic system data. 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 lacks 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 [20].

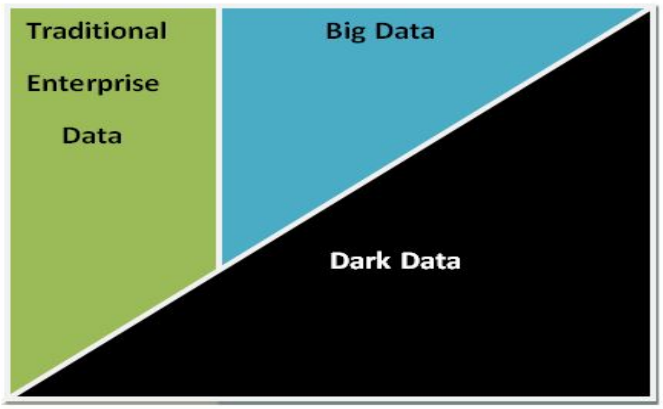


**Figure 6.** Growth of Dark Data

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 [21] explored the challenges and opportunities of dark data, focusing on data management to address performance issues. They highlighted the negative impact of central processing unit (CPU), random access memory (RAM), and data storage in knowledge (DISK) storage usage on tasks and emphasized the need for careful handling of dark data due to its potential risks and challenges.

The paper in reference [22] offered insights into the data analysis task execution and management procedures. The study determined the underlying reasons of failed tasks, including eviction, kill, and fail, which resulted in resource waste and application delay. Big data systems' fault tolerance, low latency, and application speed have all improved as a result of the analysis of unsuccessful job patterns. Semi-structured and unstructured dark data in business and industrial situations was the focus of the authors' work in [23]. They analyzed this data using business intelligence (BI) tools, and as illustrated in Figure 7, they suggested a BI strategy for obtaining useful information from dark data for the benefit of the organization.

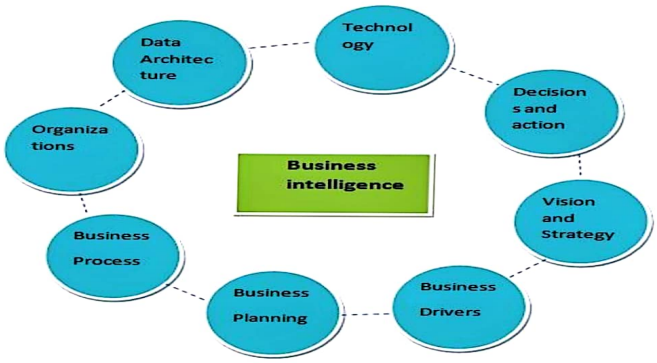


**Figure 7.** Business Intelligence (BI) Strategy

Corallo'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, Corallo 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.

In [24], the authors offer a textile application that uses both structured and unstructured data analysis to handle real-world problems. The application generates grid layout visualization, aiding in data management and problem-solving by extracting useful summaries from various data types. The researchers in [24] review unstructured data from social media platforms and applies online analytical processing (OLAP) technology to analyze numerical, textual, and elemental elements. The research contributes to the management and facilitation of unstructured data. The authors in [25] focus on designing big data techniques and algorithms to tackle current and future challenges, while the authors in [26] analyze unstructured data through various methods, reviewing applications such as text and audio analysis, video analytics, and social media analysis.

In [27], 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 support vector machine (SVM) and neural network models.

In order to make better decisions, the article in [28] highlights the significance of fusing knowledge, experience, and bravery with already-existing big data. It also raises important considerations regarding big data and its bad side. Although difficulties are acknowledged, the report emphasizes possible solutions that may be implemented well.

While research in [29] focuses on network infrastructure technology for computational usage in astronomy and cyber infrastructure for research and education (CY), [30] uses Hadoop processing and features for quick duplication removal to remove duplicates from datasets. The authors of paper [31] use APIs to gather textual data from the internet in order to study the behavior of global society. They then use

K-means cluster analysis to gain insights about national technological advancement, internet activity, and stressor responses.

The research study in [32] analyzes the disk spaces linked with big data and its ramifications, delving into big data research and highlighting the 4V model (velocity, volume, variety, and veracity).

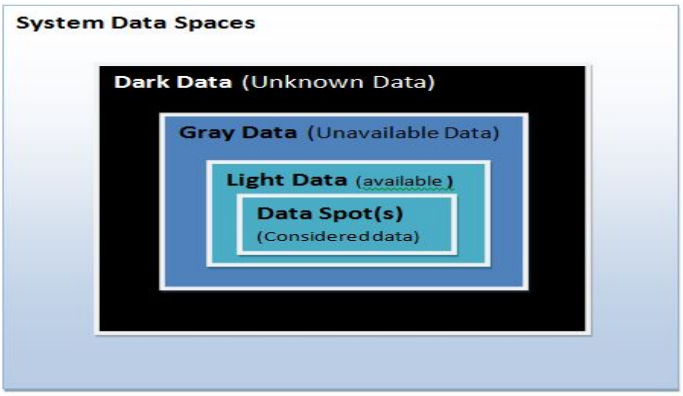


**Figure 8.** Big data's data space

The data space in the context of big data is depicted in Figure 8, which includes a variety of data types such light, gray, dark, and data spots. Researchers define dark data as unidentified and untrained data. Light data is easily accessible for use and processing, but gray data is only partially known and can be examined. Light data is subdivided into data spots. In [33], researchers shed light on dark data related to Northwestern Atlantic zooplankton from the 1970s to the 1980s, emphasizing the need to bring this data into public visibility. In [34], authors explore the importance and challenges of dark data and big data, emphasizing the opportunities they present for business analysis. In [35], authors review dark data, highlighting its significance and revealing that 90% of data exists in dark form. In [36], authors focus on data sharing, revealing that a substantial portion of data remains in the long tail, categorized as dark data. In [37], the government's future is reviewed 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 structured query language (SQL) queries to extract high-quality data from various domains [38].

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 [39]. 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 Deep LOCA-Lattice, utilizing basic models with good accuracy and revealing optimal performance parameters for breach size estimation in LOCA [40].

In order to address the issue of drunken driving-related traffic deaths, a study builds prediction models for early detection and policy creation. Traffic fatality data is used to construct a variety of supervised machine learning techniques, such as Random Forests, Decision Trees, Naïve Bayes, Logistic Regression, and Support Vector Machines [41].

Traffic control is improved by intelligent transportation systems, while problems like data sparsity and anomalous scenario observation are tackled by generative AI. In addition to reviewing the difficulties and suggesting future research areas for intelligent transportation systems, this paper delves into the applications of generative AI in traffic perception, prediction, simulation, and decision-making [42]. The University of British Columbia lab uses interdisciplinary work in computer science, statistics, and chemistry to address big data difficulties related to metabolomics. With the goal of offering metabolomics practitioners easily accessible solutions, our bioinformatics tools on GitHub address data processing, feature extraction, quantitative measurements, statistical analysis, and metabolite annotation [43].

## 3. 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.

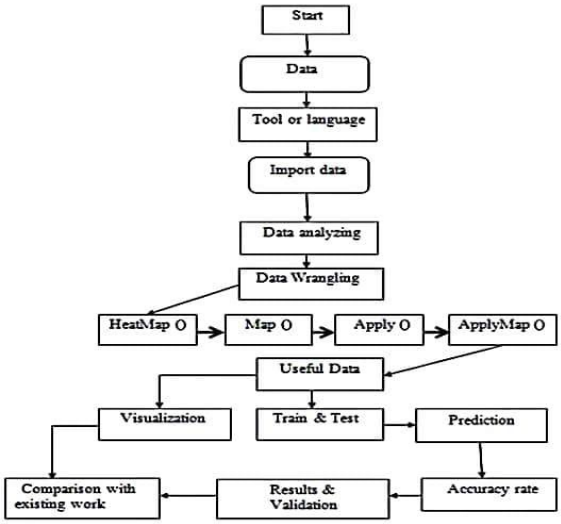


**Figure 9.** Block diagram of the proposed work

Python Jupyter Notebook is used in this study's data analysis, which includes data wrangling for cleaning and visualization to enable a more thorough comprehension of the dataset. There are multiple crucial processes and approaches in the system. Cleaning and displaying the dataset to identify underlying patterns and insights constitute the first phase. Making ensuring the data is ready for precise analysis and modeling is the goal of this step. Random Forest and AdaBoost classifiers are used in the study to forecast results. To enhance prediction performance, this entails dividing the dataset into subgroups for training and testing, defining pertinent variables, and using these classifiers. An extensive performance comparison is used to verify the classifiers' efficacy.

Performance and reliability metrics like accuracy, precision, recall, and F1 score are used to evaluate the predictions. The paper discusses the particular difficulties in gleaning information from dark data. It assesses how well machine learning classifiers do in forecasting results using this intricate kind of data. Algorithm 01 offers a step-by-step breakdown of the complete methodology, showing each step of the procedure from data preparation to assessment and prediction. The domains of dark data usage and predictive analytics both benefit greatly from this research. This study establishes a standard for using dark data to improve decision-making and prediction accuracy, highlighting the vital role that advanced analytics play in managing increasing amounts of data.

**Algorithm 01:** Algorithm for Optimized Extraction of Dark Data using AdaBoost and Random Forest

**Input:** Dark Data (UK accident information for urban and rural areas) from Kaggle

**Output:** Model Evaluation Results

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)

3.1. Data Acquisition

In our proposed research, we focus on analyzing dark data, which poses significant challenges for extracting information and making informed decisions. Initially, we conduct a thorough review of existing research. Subsequently, we select dark data, or otherwise unusable data, for detailed analysis. The data is sourced from Kaggle, a renowned data science community with an extensive repository of open-source data. Our dataset includes UK accident information for both urban and rural areas, amounting to 629 MB in a comma-separated values (CSV) file. Despite its richness, the dataset's complexity makes it difficult to analyze and derive valuable insights.

3.2. Tool or Application

Python is used in a Jupyter Notebook environment to analyze dark data. High-level programming languages like Python are chosen for their user-friendliness, effective coding capabilities, and capacity for large-scale data processing. An open-source, web-based tool called Jupyter Notebook improves this process by allowing real-time code execution, management of equations, integration of narrative prose, and visuals.

3.3. Import Data

Numerous libraries are used to import data into the Python Jupyter Notebook, including Math, NumPy, Matplotlib, Seaborn, and Pandas. These libraries are essential for reading, importing, and interpreting data.

3.4. Data Analyzing

Understanding the dataset requires a thorough understanding of data analysis. To improve understanding, details about the data are gathered, including the range of indices, total number of columns, and categories of data. According to Figure 10, the dataset, which represents accident data for UK cities and rural areas, has 32 columns and 1,917,274 rows. Data wrangling is used to clean the data, and then it is represented graphically for better 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 10.** Information of Data for analysis

3.5. Data Wrangling

Data wrangling, or data cleaning, involves addressing unnecessary values and null items, as shown in Figures 11 and 12. Heatmap(), Map(), Apply(), and ApplyMap() functions are employed for effective data cleaning.

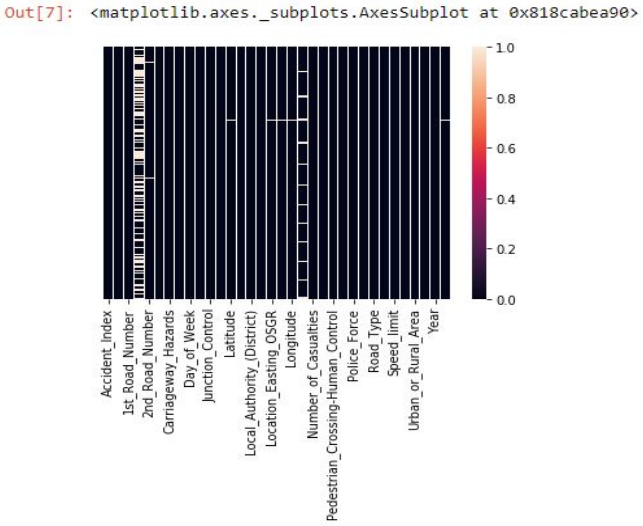


**Figure 11.** HeatMap for effective data cleaning

```
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 12.** Unnecessary and Null Values in Data

The data is successfully cleaned and unnecessary and null values are removed, resulting in a usable dataset as shown in Figures 13 and 14.

3.6. Data Visualization and Plotting

Plotting and visualizing data creates meaningful patterns and forms from the data, improving comprehension and emphasizing important information. As seen in Figure 15, visualization is used to highlight important data, such as the overall number of accidents in urban and rural areas and the severity of the incidents.

The visualizations provide valuable insights, revealing that a majority of accidents occur in urban areas. As seen in Figure 16, the dataset is further examined for accident severity, making the distinction between minor, major, and deadly accidents.

Decision-making may now be done with the data, demonstrating the possibility of obtaining valuable information from dark data that would otherwise appear useless. 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 in accident-prone areas.

3.7. Classifier using Random Forest

Using ensemble learning, the Random Forest Classifier is a potent machine learning technique that boosts prediction accuracy. This technique improves generalization on unknown data and reduces overfitting by combining several decision trees. A random subset of the data is used to train each decision tree in the forest, and the average of each tree's individual forecasts is used to determine the final prediction. This method makes use of the trees' diversity to lower variance and boost robustness.

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

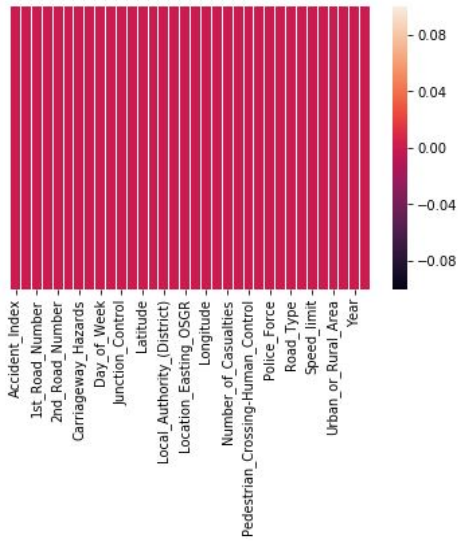


**Figure 13.** HeatMap of Clean Data

```
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 14.** Result of Clean Data

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

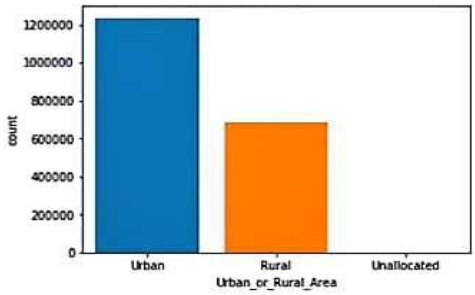


**Figure 15.** Visualization of Urban and Rural Accidents

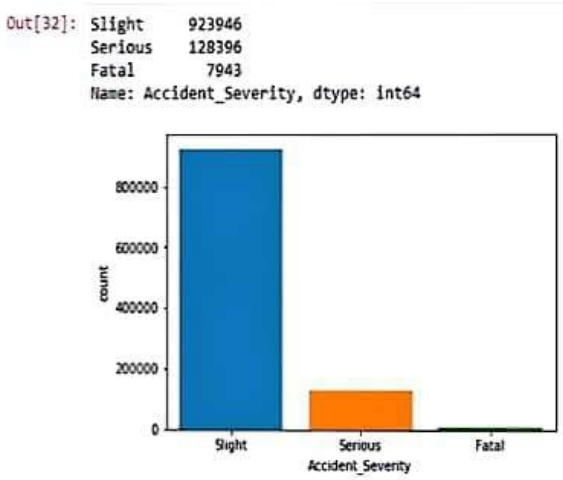


**Figure 16.** Visualization of Accident Severity

Random Forests can handle both classification and regression problems, and they work especially well with large, high-dimensional datasets. They also offer information on feature importance, which makes it possible to determine which variables in the dataset have the greatest influence.

Figure 17 shows the general design and operation of the Random Forest Classifier.

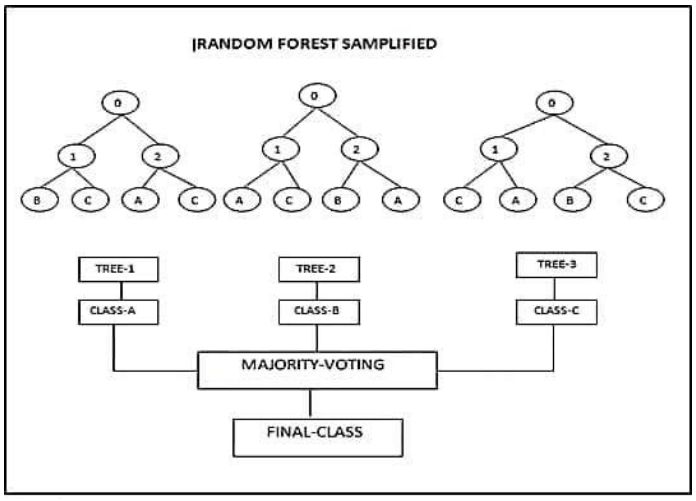


**Figure 17.** Overall view of Random Forest

3.8. The Classifier AdaBoost

Through the combination of their outputs, the AdaBoost classifier is an ensemble learning technique that improves weak learners to produce a strong prediction model. It is especially useful for examining sparse, incomplete, or unstructured dark data.

Adaptive Boosting, or AdaBoost, uses weak classifiers one after the other in a sequential fashion, fixing each other's mistakes. It improves overall model performance and accuracy by focusing more on challenging instances by modifying the weights of data points based on classification errors.

AdaBoost's iterative methodology aids in improving the model's predicted accuracy when studying dark data, such as missing accident reports or fragmented records. AdaBoost's efficiency is assessed by looking at how well it categorizes accident data.

3.9. Train and Test

A key component of machine learning is the training and testing process, which makes it possible to assess the performance of the model on untested data. The dataset is split into training and testing sets in order to evaluate the performance of the Random Forest and AdaBoost classifiers, which are used for prediction in this research work.

3.10. Dataset Splitting

The dataset is split into the training set and the testing set before going into the training and testing procedure. The machine learning model is trained on the training set, which enables it to identify patterns

and connections in the data. However, the testing set is only used to evaluate how well the model applies to fresh, untested data.

*3.10.1. Defining Independent (X) and Dependent (Y) Variables*

Typically, datasets in supervised machine learning are made up 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.'

- *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 independent variables.

- *Dependent Variable (Y)*

This variable is what the model seeks to forecast. In the circumstances of the investigation, a binary classification can apply (e.g., accident occurrence in urban or rural areas) or another relevant prediction task.

3.11. Utilizing the Random Forest Classifier

*3.11.1. Decision Trees in Random Forest*

Multiple decision trees are used in the Random Forest Classifier, an ensemble learning technique, to make predictions. Every decision tree generates its own predictions after being trained on a different subset of the dataset. The different forecasts are then combined to establish the final prediction, frequently using a voting system.

*3.11.2. Random Forest Classifier Training*

Giving the training set (X_train, Y_train) to the Random Forest Classifier is the first step in the training process. To create predictions, the classifier discovers patterns and relationships in the training set of data.

*3.11.3. Putting the Testing Set to the Test*

The Random Forest Classifier is trained and its performance is assessed by feeding it the testing set (X_test). Accuracy, precision, recall, and other performance measures are evaluated by comparing the classifier's predictions with the actual values (Y_test).

3.12. Utilizing the AdaBoost Classifier

*3.12.1. AdaBoost Boosting*

Adaptive Boosting, or AdaBoost, is an additional ensemble learning method. By assigning misclassified examples additional weight, it aims to improve the performance of weak learners—classifiers that outperform random chance by a small margin.

*3.12.2. AdaBoost Classifier Training*

The AdaBoost Classifier is learned on the training set (X_train, Y_train), just like the Random Forest. To enhance overall performance, the classifier modifies the weights of incorrectly categorized examples iteratively.

*3.12.3. Using the Testing Set for Testing*

Then, the predictions made by the AdaBoost classifier are assessed using the testing set (X_test). By contrasting the classifier's predictions with the actual values, the performance of the model is evaluated (Y_test).

The Random Forest and AdaBoost classifiers' ability to generalize to new data can be evaluated through the training and testing procedures. The efficacy and predictive capacity of the classifiers can be assessed, offering insights into how well they function in actual situations, by using distinct subsets of the dataset for training and testing.

3.13. Process for Training and Predicting Models

Using the Random Forest and AdaBoost classifiers, predictions are produced once the data has been divided into training and testing sets. In this procedure, the classifiers are trained on the training dataset so they can recognize patterns and relationships in the dataset. It is therefore possible to assess the performance of the trained models by using them to forecast results on the testing data. Comparing the expected and actual results in the testing set allows for the assessment of accuracy.

3.14. Evaluating Model Predictions and Performance Metrics

The Random Forest and AdaBoost classifiers are used to provide predictions on the testing set once they have been trained. This is an important step because it assesses how well the models apply to fresh and untested data. After analyzing the testing data, the classifiers use the patterns they discovered during training to predict the results.

The expected and actual results in the testing set are compared to see how accurate the forecasts are. A number of performance metrics must be calculated for this comparison, some examples of which could include:

- **Accuracy**: The percentage of cases out of all instances that were accurately predicted.
- **Precision**: The percentage of actual positive forecasts among all positive forecasts.
- **Recall**: The percentage of real positive instances that are accurate positive predictions.
- **F1 Score:** A single statistic that balances recall and precision, calculated as the harmonic mean of the two.

These metrics offer a thorough grasp of the model's performance inside the framework of the particular predictive task—is it regression, classification, or another kind of analysis.

3.15. Comparative Performance Evaluation

A detailed assessment and comparison of the AdaBoost and Random Forest classifiers' performance against other programs or tools are required for the validation of the findings. This comparison aims to determine the effectiveness and reliability of the proposed models.

Several aspects are considered during this evaluation:

- **Predictive Accuracy**: The ability of the models to correctly predict outcomes.
- **Robustness**: The models' consistency and reliability when applied to different subsets of data or under varying conditions.
- **Efficiency**: The computational resources and time required to train and deploy the models, as well as their scalability to larger datasets.

By rigorously assessing these factors, the study ensures that the proposed models not only perform well in terms of accuracy but also demonstrate practical applicability and robustness in real-world scenarios. The comparative analysis helps to highlight the strengths and weaknesses of the AdaBoost and Random Forest classifiers, providing valuable insights for further improvements and applications.

3.16. Confusion Matrix

To get the prediction accuracy score, a confusion matrix is employed. To create all of the data labels in the form of a matrix, as seen in Figure 18, we first import the confusion matrix module from Sklearn, matric.

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

**Figure 18.** Confusion Matrix Data Label

Figure 19 now displays the confusion matrix for the Random Forest classifier as a heatmap created with the Seaborn library.

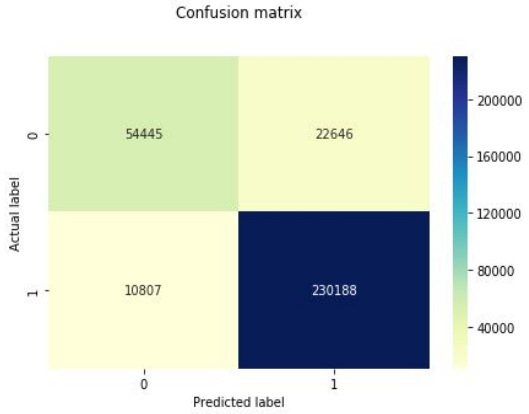


**Figure 19.** Confusion matrix for Random Classifier

The confusion matrix is an essential tool for assessing how well classification algorithms perform on the UK accident dark dataset. It is displayed as a heatmap in Figure 19 using the Seaborn library. By contrasting the model's predictions with the actual labels, this matrix offers a thorough analysis of classification accuracy and error patterns.

As seen in Figure 19, our suggested model produces the confusion matrix. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are among the total count of actual and anticipated labels that are included in it. These constituents are considered crucial for calculating diverse performance indicators and characterizing the predicted accuracy of the model.

- *Confusion Matrix Components*

**True Positive (TP):** Accurate positive predictions, totaling 54445 instances.

**True Negative (TN):** Accurate predictions of negative cases, amounting to 230188 instances.

**False Positive (FP):** Erroneous positive predictions, occurring 10807 times.

**False Negative (FN):** Inaccurate negative predictions, comprising 22646 cases.

By utilizing these matrix elements, precise calculations of our model's precision, recall, and overall accuracy are facilitated. This comprehensive analysis of the confusion matrix provides a thorough evaluation of our proposed AdaBoost model's effectiveness in classification tasks, guiding further optimizations for enhanced performance.

3.17. Performance Measures

Relevant indicators are constructed in order to assess how well the classification models work for accidents in both urban and rural areas:

True positive (TP) and true negative (TN) cases are taken into account while calculating accuracy, which is the total percentage of correctly classified cases.

$$Accuracy(AC) = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Conversely, precision quantifies the percentage of true positive predictions for precision

$$Precision\ (PR) = \frac{TP}{TP+FP} \quad (2)$$

The percentage of true positive predictions among all actual cases is quantified by recall

$$Recall\ (RE) = \frac{TP}{TP+FN} \quad (3)$$

A balanced assessment of both metrics is given by the F1-score, which is a harmonic mean of recall and precision.

$$F1\ Score = 2 \times \frac{PR \times RE}{PR+RE} \quad (4)$$

Thus, the confusion matrix provides a thorough assessment of how well the suggested model predicts accident data in both urban and rural locations. The analysis's insights, which focus on regions with high rates of false positives or false negatives, aid in the model's improvement and forecast accuracy.

## 4. Results and Discussion

The study uses Python and Jupyter Notebook to reuse accident dark data from the United Kingdom (UK) for the results and discussion part. AdaBoost and Random Forest classifiers achieve high pattern recognition precision, as evidenced by the 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.

4.1. Predicted Results for Urban and Rural Areas

In this section, we present the predicted results of our model for urban and rural areas using dark data from UK accident information. The results are illustrated in Figures 20 and 21, highlighting the frequency of the accuracy rate and the total number of predicted accidents in both urban and rural settings.

*4.1.1. Frequency of Accuracy Rate*

Figure 20 illustrates the frequency of the accuracy rate of our model's predictions. This figure demonstrates how often our model achieves different levels of accuracy in predicting accidents. One important metric that aids in our comprehension of the performance and dependability of our model is the accuracy rate.

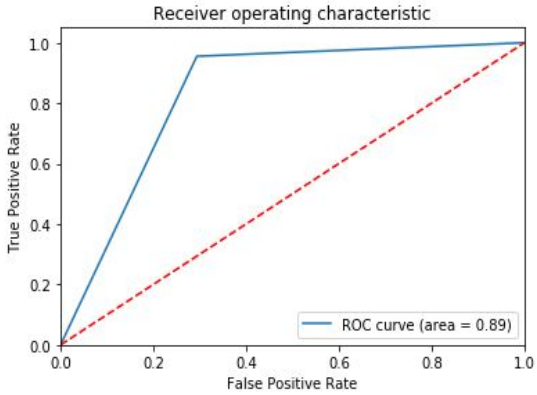


**Figure 20.** Frequency of Accuracy Rate

*4.1.2. Predicted Outcomes for Urban and Rural Regions*

The ultimate projected findings for both urban and rural areas are displayed in Figure 21. Our model predicts that there will be 804,516 overall accidents in urban areas and 255,769 in rural areas. This significant difference indicates that the model predicts a higher number of accidents in urban areas compared to rural areas.

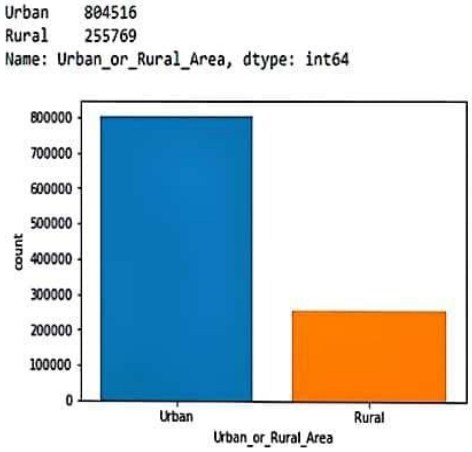


**Figure 21.** Predicted Results

This disparity in predicted accidents between urban and rural areas highlights that urban areas are more prone to accidents, as per our model's predictions. The higher number of predicted accidents in urban areas suggests that our model has a higher accuracy rate in predicting accidents in these regions. The results underline the importance of focusing on urban areas for accident prevention and safety measures.

*4.1.3. Analysis and Implications*

The analysis of the predicted results indicates several key implications:

- **Higher Risk in Urban Areas**: The higher number of predicted accidents in urban areas suggests a greater risk of accidents in these regions. This aligns with the common understanding that urban areas, due to thier higher traffic density and more complex road networks, tend to have a higher incidence of accidents.
- **Model Accuracy**: The model's higher accuracy in predicting urban accidents demonstrates its effectiveness in analyzing accident data in densely populated and complex environments. This could be due to the availability of more comprehensive data and identifiable patterns in urban settings.
- **Focus on Urban Safety Measures**: Given the higher predicted accident rate in urban areas, it is crucial for policymakers and urban planners to prioritize safety measures in these regions. This could include improving road infrastructure, implementing stricter traffic regulations, and enhancing public awareness campaigns.
- **Rural Area Insights**: While the predicted number of accidents in rural areas is lower, it is still significant. Rural areas often face challenges such as longer emergency response times and less

robust infrastructure, which can exacerbate the impact of accidents. Therefore, targeted interventions in rural areas are also necessary.

The predicted outcomes show how various approaches are required in urban and rural locations for accident management and prevention. Our model's insights can be used to guide specific safety measures, which will ultimately lower accident rates and improve public safety.

4.2. Cross-Validation and Classification Report

The module, its featured classes, and all of the functions for the train and test datasets were validated using a cross-validation test from the Sklearn package. To provide classification reports, they fit data into the random forest classifier and validate it. The prediction findings are accurate if the total average classification report matches the prediction score. The classification report is as follows. Table 1 displays the average classification result using cross-validation, which is 89.50%. This indicates that the prediction was accurate in the end, with an accuracy percentage of 89.50%.

**Table 1.** Classifiers' Results Comparisons Using Cross-Validation

| Classifiers | Precision (%) | Recall (%) | F-Score (%) | Accuracy (%) |
|---|---|---|---|---|
| AdaBoost | 86 | 62 | 72 | 78.4 |
| Random Forest | 83.44 | 70.63 | 76.5 | 89.50 |

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

The emerging idea of using dark data—information that was previously unexplored and frequently disregarded—in data-driven decision-making offers a chance to improve prediction models and increase 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.

4.4. Classifier Performance Evaluation

An extensive evaluation is conducted on two well-known classifiers: the Random Forest and AdaBoost classifiers, based on precision, recall, F-Score, and total accuracy. While recall evaluates the capacity to capture all positive occurrences, precision measures the accuracy of positive predictions. By balancing precision and recall, the F-Score offers a fair assessment of a classifier's effectiveness. The classifier's total accuracy indicates how well it can categorize cases throughout the whole dataset.

4.5. AdaBoost Classifier Evaluation

The AdaBoost classifier exhibits a precision of 86%, indicating 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. Although there is still opportunity for growth in terms of catching all positive events, the classifier's total accuracy of 78.4% highlights its skill in accurate classification.

4.6. Evaluation of Random Forest Classifiers

On the other hand, the Random Forest classifier exhibits an 83.44% precision, indicating a noteworthy percentage of successfully detected positive cases among the projected positives. Recall measure of 70.63% is impressive as it indicates a careful capture of genuine positive cases, demonstrating the sensitivity of the classifier. The final F-Score of 76.5% shows that recall and precision have been harmoniously balanced. The classifier is a strong candidate in properly predicting accident occurrences in both urban and rural environments, as evidenced by its notable accuracy of 89.50%.

4.7. Contribution to the Field

This study advances the paradigm of data use by being the first to apply dark data to incident prediction models, which makes a substantial contribution to the area. 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 enter a data-rich era, utilizing the latent insights found in dark data becomes a critical tactic for improving decision-making procedures and improving the precision of prediction models across various industries.

## 5. Conclusion and future work

Conclusively, this study offers an exhaustive investigation of the complex terrain of big data, stressing its fundamental function in several fields and exposing obstacles related to data integrity, storage, safety, and analysis. 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 and 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 86% precision in pattern recognition for AdaBoost and 89.50% accuracy for the Random Forest classifier, this study pioneers dark data applications 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.

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. Feng Shi†, Liming Xia†, Fei Shan†, Dijia Wu, Ying Wei, Huan Yuan, Huiting Jiang, Yaozong Gao, & He Sui, Dinggang Shen. (2020). Large-Scale Screening of COVID-19 from Community Acquired Pneumonia using Infection Size-Aware Classification.
2. Aditya Kakde, Nitin Arora, Durgansh Sharma, & Subhashchander Sharma. (2020). Multi-spectral classification and recognition of breast cancer and pneumonia. Polish Journal of Medical Physics and Engineering, 26(1), doi: 10.2478/pjmpe-2020-0001.
3. Hussain, A., Khan, A., Yar, H., & Khan, N. F. (2019). Efficient Deep learning Approach for Classification of Pneumonia using Resources Constraint Devices in Healthcare. In The 5th International Conference on Next Generation Computing 2019.
4. Jung, S.-m., Kinoshita, R., Thompson, R. N., Linton, N. M., Yang, Y., Akhmetzhanov, A. R., & Nishiura, H. (2020). Epidemiological Identification of A Novel Pathogen in Real Time: Analysis of the Atypical Pneumonia Outbreak in Wuhan, China, 2019–2020. Journal of Clinical Medicine, 9(3), 637. Doi: 10.3390/jcm9030637.
5. Pereira, R. M., Bertolini, D., Teixeira, L. O., & Silla Jr, C. N. (2020). Covid-19 identification in chest x-ray images on flat and hierarchical classification scenarios. arXiv preprint arXiv:2004.05835.
6. Loey, M., Smarandache, F., & Khalifa, N. M. (2020). Within the Lack of COVID-19 Benchmark Dataset: A Novel GAN with Deep Transfer Learning for Coronavirus Detection in Chest X-ray Images.
7. Khalifa, N. M., Smarandache, F., & Loey, M. (2020). COVID-19 Chest X-ray Images Diagnosis: A Neutrosophic and Deep Transfer Learning Approach. doi:10.20944/preprints202004.0371.v1.
8. Janizek, J. D., Erion, G., & Lee, S.-I. (2020). An Adversarial Approach for the Robust Classification of Pneumonia from Chest Radiographs. arXiv preprint arXiv:2001.04051.
9. Chowdhury, M. E. H., Rahman, T., Khandakar, A., Mazhar, R., Abdul Kadir, M., Bin Mahbub, Z., Islam, K. R., Khan, M. S., Iqbal, A., Al-Emadi, N., & IbneReaz, M. B. (2020). Can AI help in screening Viral and COVID-19 pneumonia? Department of Electrical Engineering, Qatar University, Doha-2713, Qatar.
10. Rahman, T., Chowdhury, M. E. H., Khandakar, A., Islam, K. R., Islam, K. F., Mahbub, Z. B., Kadir, M. A., & Kashem, S. (2020). Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection using Chest X-ray. Applied Sciences, 10, x. doi: FOR PEER REVIEW.
11. Zhou, M., Chen, Y., Yang, D., Xu, Y., Yao, W., Huang, J., Jin, X., Pan, Z., Tan, J., Wang, L., Xia, Y., Zou, L., Xu, X., Wei, J., Guan, M., Feng, J., Zhang, H., & Qu, J. (2020). Improved deep learning model for differentiating novel coronavirus pneumonia and influenza pneumonia. doi: https://doi.org/10.1101/2020.03.24.20043117.
12. Rehman, A., Naz, S., Khan, A., Zaib, A., & Razzak, I. (2020). Improving Coronavirus (COVID-19) Diagnosis using Deep Transfer Learning. doi: https://doi.org/10.1101/2020.04.10.035717.
13. Alshammari, M. K., Alotaibi, M. A., AlOtaibi, A. S., Alosaime, H. T., Aljuaid, M. A., Alshehri, B. M., ... & Alotaibi, A. A. (2023). Prevalence and Etiology of Community- and Hospital-Acquired Pneumonia in Saudi Arabia and Their Antimicrobial Susceptibility Patterns: A Systematic Review. Medicina, 59(4), 760.
14. File Jr, T. M., & Ramirez, J. A. (2023). Community-acquired pneumonia. New England Journal of Medicine, 389(7), 632-641.
15. Chouhan, V., Singh, S. K., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., Damaševi˘cius, R., & de Albuquerque, V. H. C. (2020). A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images. Applied Sciences, 10, 559. doi:10.3390/app10020559.
16. Zhang, Y.-J., Zhao, P., & Zhou, Z.-H. (2020). Exploratory Machine Learning with Unknown Unknowns. arXiv preprint arXiv:2002.01605.
17. Varshni, D., Nijhawan, R., Thakral, K., Mittal, A., & Agarwal, L. (2019). Pneumonia Detection Using CNN based Feature Extraction. IEEE.
18. EL ASNAOUI, K., CHAWKI, Y., & IDRI, A. (2020). Automated Methods for Detection and Classification Pneumonia based on X-Ray Images Using Deep Learning. Complex system engineering and human system, Mohammed VI Polytechnic University Benguerir, Morocco. IEEE Conference.
19. Khalifa, N. E. M., Taha, M. H. N., Hassanien, A. E., & Elghamrawy, S. (2020). Detection of Coronavirus (COVID-19) Associated Pneumonia based on Generative Adversarial Networks and a Fine-Tuned Deep Transfer Learning Model using Chest X-ray Dataset. Computers and Artificial Intelligence, Cairo University, Giza, Egypt.
20. Neili, Z., Fezari, M., & Redjati, A. (2020). ELM and K-nn machine learning in classification of Breath sounds signals. International Journal of Electrical and Computer Engineering (IJECE), 10(4), 3528-3536. doi: 10.11591/ijece.v10i4.pp3528-3536.

21. Islam, K., Wijewickrema, S., Collins, A., & O'Leary, S. (2020). A Deep Transfer Learning Framework for Pneumonia Detection from Chest X-ray Images. DOI: 10.5220/0008927002860293.
22. Verma, G., & Prakash, S. (2020). Pneumonia Classification using Deep Learning in Healthcare. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9(4). ISSN: 2278-3075.
23. Ma, H. R., Deng, B. Y., Liu, J., Jiang, P., Xu, Y. L., Song, X. Y., ... & Fu, W. (2023). Lung ultrasound to diagnose infectious pneumonia of newborns: A prospective multicenter study. Pediatric Pulmonology, 58(1), 122-129.
24. Ilyas, M., Rehman, H., & Nait-ali, A. (2020). Detection of Covid-19 From Chest X-ray Images Using Artificial Intelligence: An Early Review. arXiv preprint arXiv:2004.05436.
25. Evans, E. D., Duvallet, C., Chu, N. D., Oberst, M. K., Murphy, M. A., Rockafellow, I., ... & Alm, E. J. (2020). Predicting human health from biofluid-based metabolomics using machine learning. doi: https://doi.org/10.1101/2020.01.29.20019471.
26. Nagamounika, R., Sri Harshitha, C. N. S. V., Tejaswi, A., Lavanya, K., & Lakshmi, P. R. S. M. (2020). Prediction of pneumonia disease by using deep convolutional neural networks. Journal of Emerging Sciences, ISSN NO: 0377-9254.
27. Liebenlito, M., Irene, Y., & Abdul Hamid. (2020). Classification of Tuberculosis and Pneumonia in Human Lung Based on Chest X-Ray Image. Indonesian Journal of Pure and Applied Mathematics, 2(1), 24–32. doi: 10.15408/inprime.v2i1.14545.
28. Yao, S., Chen, Y., Tian, X., Jiang, R., & Ma, S. (2020). An Improved Algorithm for Detecting Pneumonia Based on YOLOv3. Applied Sciences, 10, 1818. doi:10.3390/app10051818.
29. Ansari, N., Faizabadi, A. R., Motakabber, S. M. A., Ibrahimy, M. I., & Lampur, K. (2020). Effective Pneumonia Detection using Res Net based Transfer Learning. January - February 2020, ISSN: 0193-4120, Pages 15146–15153.
30. Verma, G., & Prakash, S. (2020). Pneumonia Classification using Deep Learning in Healthcare. International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-9 Issue-4, February 2020.
31. Hidayatullah, R. C., & Violina, S. (2020). Convolutional Neural Network Architecture and Data Augmentation for Pneumonia Classification from Chest X-Rays Images. International Journal of Innovative Science and Research Technology, ISSN No: 2456-2165, 2020.
32. Chadha, R. (2019). A Novel Approach for Detecting Pneumonia in Machine Learning. International Journal of Trend in Innovative Research (IJTIIR), www.ijtiir.com | Volume-1 | Issue-4.
33. Chumbita, M., Cillóniz, C., Puerta-Alcalde, P., Moreno-García, E., Sanjuan, G., Garcia-Pouton, N., ... & Garcia-Vidal, C. (2020). Can Artificial Intelligence Improve the Management of Pneumonia. J. Clin. Med., 9, 248. doi:10.3390/jcm9010248.
34. Ebner, L., Christodoulidis, S., Stathopoulou, T., Geiser, T., Stalder, O., Limacher, A., Heverhagen, J. T., Mougiakakou, S. G., & Christe, A. (2020). Meta-analysis of the radiological and clinical features of Usual Interstitial Pneumonia (UIP) and Nonspecific Interstitial Pneumonia (NSIP). PLOS ONE, 15(1), e0226084. https://doi.org/10.1371/journal.pone.0226084.
35. Shuaib, K. M., Shahid, S. A., Javed, A. A., & Zaid, M. (2020). Pneumonia Detection through X-Ray Using Deep Learning. IOSR Journal of Computer Engineering (IOSR-JCE), 22(1), 08-11. https://doi.org/10.9790/0661-2201040811.
36. Reddy, R., Teja, G. S. R., Tej, D. S., & Babu, P. V. (2020). Prediction of Pneumonia using deep learning. International Journal of Advance Research, Ideas and Innovations in Technology, ISSN: 2454-132X.
37. Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q., Cao, K., Liu, D., Wang, G., Xu, Q., Fang, X., Zhang, S., Xia, J., & Xia, J. (2020). Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT. Department of radiology China.
38. Liao, Y.-H., Wang, Z.-C., Zhang, F.-G., Abbod, M. F., Shih, C.-H., & Shieh, J.-S. (2019). Machine Learning Methods Applied to Predict Ventilator-Associated Pneumonia with Pseudomonas aeruginosa Infection via Sensor Array of Electronic Nose in Intensive Care Unit. Sensors, 19, 1866. https://doi.org/10.3390/s19081866.
39. Setiawan, W., & Damayanti, F. (2019). Layers Modification of Convolutional Neural Network for Pneumonia Detection. Journal of Physics: Conference Series, 1477(5), 052055. IOP Publishing. https://doi.org/10.1088/1742-6596/1477/5/052055.
40. Pierre, J. F., Akbilgic, O., Smallwood, H., Cao, X., Fitzpatrick, E. A., Pena, S., Furmanek, S. P., Ramirez, J. A., & Jonsson, C. B. (2020). Discovery and Predictive Modeling of Urine Microbiome, Metabolite and Cytokine Biomarkers in Hospitalized Patients with Community Acquired Pneumonia. bioRxiv preprint. doi: https://doi.org/10.1101/2020.03.05.979427.

41. Urey, D. Y., Saul, C. J., & Taktakoglu, C. D. (2019). Early Diagnosis of Pneumonia with Deep Learning. arXiv:1904.00937v1 [cs.CV].
42. Wang, K., & Shi, Y. (2023). Discussion on the syndrome differentiation treatment of pneumonia and cough in children. MEDS Chinese Medicine, 5(8), 121-126.
43. Yu, X., Zhang, S., Xu, J., Huang, Y., Luo, H., Huang, C., ... & Wang, X. (2023). Nomogram using CT radiomics features for differentiation of pneumonia-type invasive mucinous adenocarcinoma and pneumonia: multicenter development and external validation study. American Journal of Roentgenology, 220(2), 224-234.
44. Javed, R., & Khan, A. H. (2022). A Systematic Analysis for Cardiovascular Disease Classification Using Deep Learning. Journal of Computing & Biomedical Informatics, 3(02), 32-41.