VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2024, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 12, Number 3, July-Septemeber 2024 pp: 82-97

Revolutionizing Network Intelligence: Innovative Data Mining and Learning Approaches for Knowledge Management in Next-Generation Networks

Daud Khan [1], **Haseeb khan** [2], **Muhammad Abrar Khan** [1], **Waqas Ahmad** [1*], **Zain Shaukat** [1], **Shahab Ul Islam** [3]

[1]Department of Computer Science Iqra National University Peshawar , Pakistan; [2]Kyungpook National university , Deague south korea ; [3]Department of Engineering Parthenope, University of Naples, Italy

Keywords: *Graph theory; shortest path; time complexity; heap data structures; Dijkstra's algorithm.*

Journal Info:
Submitted:
August 12, 2024,
Accepted:
August 20, 2024
Published:
August 30, 2024

Abstract The Information and Communication (ICT) industry, a global giant among service sectors, is known for both its massive scale and its unforgiving demands. Here, downtime is unacceptable, requiring constant high availability – often at the stringent Sigma Six standard. Redundancy is a common solution, but it comes at a cost. To meet these demands proactively, the ability to predict load and growth becomes crucial. This project aims to develop a prototype, or proof of concept, that utilizes data mining to provide early warnings and growth forecasts for the ICT industry with good accuracy. Big data is key to making discoveries in any data analysis project. Normally, this data comes from real-time system logs. However, for this initial test, I used a dataset called MIT Reality Mining. This dataset is useful because real-world companies, especially in the tech industry (ICT), are often hesitant to share their current information. By using MIT Reality Mining, I could still find trends and potential reasons behind them in the ICT industry. It's important to remember that this is a limited functionality prototype. While it can serve as a guideline for Telcos looking to implement data warehouses, the actual implementation details will need to adapt to the specific needs of each industry.

***Correspondence author email address:** Waqas.ahmad@inu.edu.pk
DOI: 10.21015/vtse.v12i3.1882

1 Introduction

The rise of the Information Age, fueled by the merging of computers and communication, has created a society hungry for knowledge. But most of this knowledge exists as raw data - like unprocessed ingredients. Data itself is just recorded facts, whereas information is the meaning we extract from those facts, the patterns hidden beneath the surface.[1].Imagine vast databases brimming with valuable, undiscovered insights. Data mining is the technique for unlocking

these hidden gems. It's like sifting through a mountain of information to find the gold nuggets – previously unknown and potentially game-changing knowledge. This process combines artificial intelligence, machine learning, statistics, and database systems. The goal of data mining is to transform this raw data into a human-friendly format, making it usable and valuable. It involves a range of activities, including database and data management, data cleaning and modeling, selecting the right metrics, managing complexity, refining the discovered patterns, visualizing the results, and keeping the system updated with new information [2].

Traditionally, data mining involved manually or semi-automatically sifting through massive datasets to uncover hidden patterns.These patterns could be groups of similar data points (clustering), unusual data points (anomalies), or connections between data points (association rules). This concept of "mining" for valuable information from data has been around for centuries. Early techniques for finding patterns included methods like Bayes' theorem and regression analysis. However, the rise of computers and the ever-growing amount of data available have made data collection, storage, and analysis much more powerful.As datasets have grown in size and complexity, manual analysis has become less feasible. Data mining utilizes a variety of algorithms inspired by computer science, such as neural networks, clustering, genetic algorithms, decision trees, and support vector machines, to automate this process of uncovering hidden patterns[3].

Essentially, data mining is the application of these algorithms to large datasets to reveal hidden insights and relationships.It bridges the gap between applied statistics, artificial intelligence, and database management. By taking advantage of how data is stored and organized in databases, data mining allows these powerful algorithms to be applied efficiently to massive datasets. Data mining has found wide application in various scientific and engineering fields, including genetics, medicine, electrical engineering, and education[4].Data mining empowers users to analyze data from various angles and categorize it, revealing hidden connections. As illustrated in the figure below, data mining involves a series of steps. The first step is data warehousing, where information from separate databases is combined. This process ensures data is collected and organized chronologically, free of errors and duplicates. Data warehouses are particularly useful for tasks requiring frequent updates and adjustments. They essentially make information readily available for those involved in decision-making. These warehouses typically function as read-only databases, providing historical data for analysis. After warehousing, data is carefully selected, transformed (formatted), and then mined for hidden patterns. The insights gained from this process are then incorporated back into the system, creating a continuous cycle of data analysis and discovery.

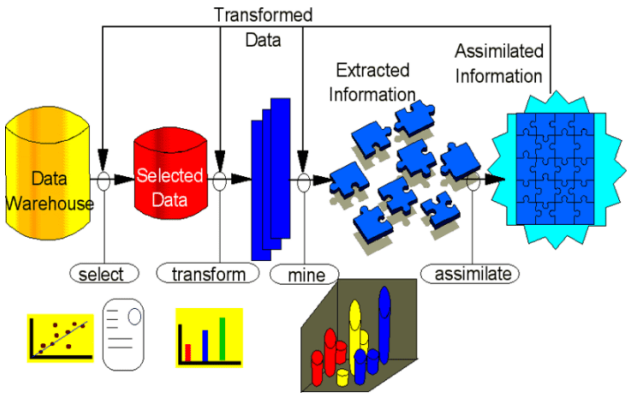

Figure 1. Data Mining [4]

.

The vast amount of data generated by communication networks is a treasure trove for researchers studying network behavior. This data comes in all sizes, from small-scale studies to massive datasets. It's crucial for not just operating and maintaining the networks that make up the internet, but also for analyzing, simulating, and emulating their behavior.

Data mining techniques, like artificial neural networks and decision trees, can be used to analyze this data and predict future trends and network behavior. This valuable information helps businesses make informed decisions. Additionally, data mining can answer complex questions in industries where

traditional methods were too slow.

Even better, these data mining methods work with existing software and hardware, getting the most out of the information we already possess.[5]

2 Scope and Objectives

The scope of this study is to develop a prototype or proof of concept that leverages data mining techniques to provide early warnings and growth forecasts specifically tailored for the Information and Communication Technology (ICT) industry. This prototype aims to address the critical need for high availability and predictive capabilities within the ICT sector by utilizing big data, particularly from real-time system logs. However, the primary objective of this study is to create a prototype that can accurately predict load and growth trends in the ICT industry using data mining techniques. This involves:

- Demonstrating the feasibility of using data mining for early warning systems and growth forecasting in the ICT sector.
- Utilizing the MIT Reality Mining dataset to identify trends and potential reasons behind them, serving as a proxy for real-world data.
- Providing a guideline for telecommunications companies (Telcos) to implement data warehouses tailored to their specific needs, despite the limitations of the prototype.
- Ensuring that the prototype adheres to high standards of availability and reliability, such as the Sigma Six standard, and addresses the costs associated with redundancy.

3 STATE-OF-THE-ART

Data mining, sometimes called knowledge discovery, is the process of analyzing large amounts of information from various angles. This analysis helps uncover hidden patterns and extract valuable insights. Businesses can use this knowledge to make informed decisions, such as increasing profits, reducing costs, or both. Data mining acts like a computer-assisted tool for sifting through data and extracting its meaning. It allows businesses to predict future trends and answer complex questions that were previously too time-consuming to tackle. Imagine having a tool that can uncover hidden connections between your customer data. Data mining software does just that! It goes beyond simply rearranging information; it acts like a detective, revealing previously unknown relationships among data points. This allows businesses to identify customers with similar interests, a valuable asset for targeted marketing campaigns and personalized customer experiences.

In[6]Data mining, though still a young field, is already making waves in a wide range of industries, including retail, finance, healthcare, manufacturing, and aerospace. Companies are using data mining techniques to unlock the power of their historical data (chronological information). Think of it as sifting through a warehouse full of information. Data mining uses pattern recognition, statistics, and math to identify hidden gems: important trends, relationships, and anomalies that might otherwise be missed [7].Data mining technology goes beyond just analyzing data; it's a powerful tool for uncovering new business opportunities. By using computer algorithms, data mining automatically identifies trends and patterns in massive datasets. This allows companies to answer questions that previously required extensive testing, directly from the information they already have.In [8] Data mining isn't just about analyzing data; it's a powerful tool for making predictions. Take targeted advertising, for example. Data mining can analyze past marketing campaigns to identify which audiences are most likely to respond to future campaigns. This predictive power extends beyond marketing. It can be used to forecast financial risks like bankruptcy or loan defaults. Data mining can also help identify groups of people who are likely to react similarly to specific situations[9].One of data mining's superpowers is automatically finding hidden patterns in vast datasets. Imagine sifting through mountains of data and uncovering previously unknown trends and relationships.

In [10],the classification of extraordinary information documentation that might be causing data errors and mistakes requires further exploration and

analysis. This involves association rule learning (dependency modeling).In [11]Data mining isn't just about finding patterns; it can also reveal hidden connections between different pieces of information. Imagine a supermarket that tracks customer purchases. Data mining can help them discover which products are often bought together, like peanut butter and jelly. This knowledge is gold! They can use it to target advertising campaigns, optimize product placement on shelves, and boost sales. Another powerful technique is called clustering. This involves grouping similar data points together. For instance, a bank might use data mining to cluster customers based on their spending habits. This allows them to target specific groups with relevant marketing campaigns and manage risk more effectively[12].Imagine you have a pile of incoming emails. Data mining with classification can help sort them automatically! It assigns new data points to predefined categories, like labeling emails as "spam" or "important." This technique is like a detective looking for the best fit. It finds a formula that most accurately predicts future values based on existing data. For example, a company might use regression to forecast future sales based on past trends. Data mining can condense massive datasets into easy-to-understand summaries. It helps you see the bigger picture by identifying key trends and generating reports you can use for better decision-making.

3.1 How does data mining work?

The ever-growing field of big data can feel overwhelming, with complex algorithms and models. Data mining bridges the gap between the two.It uses software to analyze relationships and patterns hidden within large datasets, unlocking valuable customer insights.These software tools often rely on various logical schemes, including statistics, machine learning, and neural networks.Here's how data mining utilizes some common association types.This involves sorting data points into predefined categories.For example, a coffee chain might use data mining to classify customers based on their purchase history.This allows them to identify loyal customers and tailor promotions to their preferences.In essence, data mining helps businesses make sense of big data by uncovering valuable customer insights that can be used to improve sales and marketing efforts[13].

3.2 Data mining techniques

Data mining uses various techniques to uncover hidden patterns in data. Here are some of the most common ones:

3.2.1 Grouping

This technique clusters data points with similar characteristics[14].For example, a retail store might group customers based on their shopping habits to identify market segments and tailor marketing campaigns.

3.2.2 Association

This method helps discover relationships between different pieces of data.Imagine a grocery store analyzing customer purchases to see what items are often bought together, like bread and milk.This knowledge can be used to optimize product placement on shelves[15].

3.2.3 Sequential Patterns

This technique predicts future behaviour based on past sequences. For instance, an online retailer might use sequential patterns to predict the probability of a customer buying a backpack based on their past purchases.

3.3 Data Mining Algorithms

Here's a deeper dive into some specific data mining algorithms:

3.3.1 Genetic Algorithms

Inspired by natural selection, this technique mimics evolution to find the best solution to a problem. Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as mutation, selection, and crossover (also called recombination). This algorithm imitates the Darwinian idea that Nature is the best optimizer.

In the manuscript on Genetic Algorithms, include:

1. Basic Principles: Description of selection, crossover, mutation, and fitness evaluation.

2. Applications: Examples of GAs in optimization, machine learning, scheduling, and design.
3. Implementation Parameters: Population size, crossover rate, mutation rate, selection methods, and termination criteria.

3.3.2 Decision Trees

Imagine a flowchart with questions leading to different outcomes. Decision trees use a tree-like structure with questions at each branch to classify data points. For example, a bank might use a decision tree to assess a customer's loan eligibility based on income, credit score, and other factors[16]. (CART and CHAID are two specific types of decision trees with different approaches to splitting data.)

3.3.3 Rule Induction

This technique identifies "if-then" rules from data. Imagine a set of rules like "if a customer buys a camera, then they are more likely to buy a memory card." These rules can be used for targeted marketing or product recommendations.

3.3.4 Data Visualization

This involves creating charts and graphs to visually represent complex data relationships. These visuals help identify trends and patterns that might be difficult to spot in raw data. By using these techniques, data mining empowers businesses to make data-driven decisions and gain valuable insights from their information.

4 TECHNICAL SECTION

4.1 Data set

To conduct our data mining analysis, we utilized Weka, a freely available and open-source software specifically designed for data mining tasks.The data set was obtained from http://reality.mediamit.edu/download.php named Single subject.Then this dataset was imported in MySQL.After that it was imported in .csv files.These.csv files were then used in Weka for analyzing and determining the variations using data mining techniques. Here are the steps to import any file into Weka and analyze it.This tutorial is conducted on Linux, an open-source operating system: Download the latest version of Weka from http://prdownloads.sourceforge.net/weka/weka-3-6-6.zip This version is suitable for the Linux environment.Unzip the downloaded file.Open the terminal and Navigate to the folder where you unzipped the file.Run the following command: java -jar weka.jar

4.2 The reality mining dataset

The Reality Mining project is the largest mobile phone study ever conducted in academia.It collects an unprecedented amount of data on participants' daily lives, creating a massive dataset that will be freely available to researchers worldwide.This data will include information on over 50,000 hours of human activity, equivalent to roughly 60 years of continuous data.According to David Lazer, a renowned social network researcher at Harvard, this project has the potential to revolutionize the field of social network research.[17].

4.3 Data types

The Reality Mining project involves 100 participants using Nokia 6600 smartphones.These phones have custom software installed, designed by the University of Helsinki, that runs in the background and collects data.The MIT Media Lab forms the bulk of the participants, making up 75 percent of the group. The remaining 25 participants are external students from the nearby MIT Sloan Business School.

The group at the Media Lab includes 20 master's students from outside MIT and 5 MIT freshmen.The data collected includes: Call logs (who you called and when), Nearby Bluetooth devices (identifies other participants you're near), Cell tower IDs (tracks your location), App usage (what apps you use and for how long), Phone status (charging, idle).This nine-month study is expected to generate data equivalent to roughly 500,000 hours of information on the participants' location, communication patterns, and app usage.After the study is complete, a public, anonymized version of the dataset will be released for other researchers to use.

Table 1. Data mining algorithm table

Algorithm	Proposer Provider	Description	Applicability	Pros (features)	Cons
Decisions Trees	Microsoft SQL, Server Analysis	Classification algorithm provided by Microsoft SQL Server Analysis Services Used in analytical modelling of discrete and continuous attributes. Makes a model that generates a chain of nodes in a tree	Applies in investigative analysis and prediction also used for Exploration	Quick implementation. Simple to understand. Performance is better in terms of large data. Generates Symbolic description. High dimensional is no longer a problem	Enormous search space Easily not expressible Optimal decisions are made that doesn't guarantee to return globally optimal tree
ID3	Ross Quinlan	Classification algorithm builds a decision tree from a fixed set of instances. The resulting tree is used for the classification of future samples. Requires attribute value description, predefined and discrete classes.	Applied in inductive methods	Easily understandable and good for generating rules.	Not able to grip non-numeric data easily. Experiences over fitting, Might be quite large. Reduction is required
K-means algorithm	Stuart Lloyd	Simple iterative method. One of the popular clustering algorithms. Partitions a given data set into users specified number of clusters *k*. The algorithm iterates between two steps, named data assignment and relocation of means till convergence.	Particularly applied when using heuristics Applied in computer vision. Often used as a pre processing step for other algorithms.	Scalable, easily understandable-able. Simple can be easily modified to deal with streaming data.	It will falter, whenever the data is not well described. Converges when the assignments do not change. Quite sensitive to initial centroid location Convergence is only to a local optimum.

4.4 Phone usage statistics

The Reality Mining project offered a unique opportunity to study how people use their phones for an extended period.By tracking phone usage for 100 participants, the researchers gained insights into both user habits and phone design. For example,the study revealed that 35 percent of participants used the clock function regularly,primarily for setting alarms and hitting snooze. Interestingly,this frequently used function required ten keystrokes to access from the default settings. This suggests that commonly used features might benefit from easier accessibility.The study also found that certain features,like the alarm clock, were used more frequently at home compared to work environments.The text below the passage likely includes a chart displaying this data.Perhaps most surprisingly, regardless of the participants' technical expertise, even complex phone features weren't used extensively.In fact, the simple pre-installed game "Snake" was used just as much as the more sophisticated Media Player application [18].

While studying app usage offers valuable insights,the Reality Mining project underscores the phone's primary function: communication.The data reveals a clear dominance of voice calls.A whopping 81 percent of phone interactions involved placing or answering calls [19].Interestingly, voice calls weren't the only communication method. Text messages

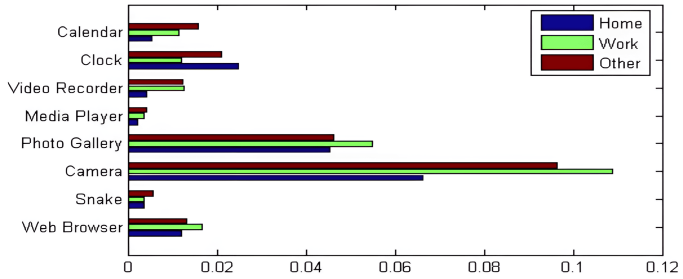

Figure 2. Aggregate application usage in co ntext

accounted for 5 percent of interactions, and emails made up another 13 percent.By understanding users' communication routines, phone designs could be optimized. For example, frequently used features could be placed in more prominent locations.This could improve overall phone usability.As we'll explore later, these improvements can be further enhanced by considering a user's social network.

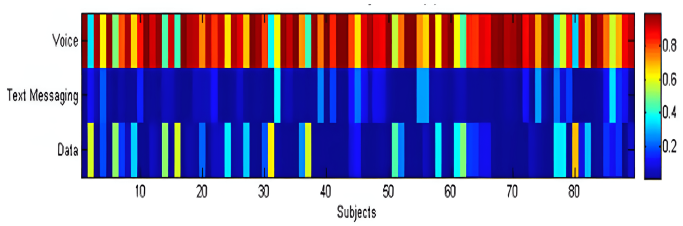

Figure 3. Communication usage patterns

4.5 Data characterization and validation

The Reality Mining project aimed to collect data continuously, but there were occasional interruptions.These interruptions, fortunately, didn't affect critical information.A small program was installed to automatically start data collection upon phone startup and to keep checking if it's running. Ideally, this program would ensure continuous data collection [20].It also guaranteed data collection began immediately after turning on the phone. However, while this program aimed to be active most of the time the phone was on, the collected data wasn't entirely error-free.The next section discusses three ways errors might occur: data corruption, software malfunction, and, most importantly, user error.

4.6 Data corruption

The Reality Mining project initially stored all collected data on a phone's flash memory card, which has a limited lifespan. Unfortunately, early versions of the data collection software repeatedly wrote to the same memory locations[21].This caused the cards to fail after about a month of data collection, resulting in complete data loss. We then modified the software to temporarily store data in the phone's RAM before transferring it to the flash memory card.Thankfully, this change prevented further complete data loss.However, ten cards failed before the issue was identified, leading to data loss for four Media Lab students and six Sloan students for the months of August and September[22].

4.7 Bluetooth errors

This study aims to see how accurate phone data is for mapping social connections. However, there's a hurdle to consider: Bluetooth range.Bluetooth typically reaches only 10 meters, and walls can further limit it [23].This means people who aren't truly close might be mistakenly logged as being together.Another challenge is the data collection interval. Since information is only collected every five minutes, brief interactions might be entirely missed. There's also a small chance that a nearby phone might not be detected during a scan. This could be caused by two factors.A minor issue with the phone's Bluetooth function, called the "BTServer," can crash about once every three days. This has a minimal impact on overall data accuracy.The data collection process itself might miss a device during a scan. However,it's important to note that this study benefits from both phones actively searching for each other.This significantly reduces the chances of missing a connection due to these limitations. Here's a positive aspect of using phone data in this study: since both participants' phones are actively searching for each other,the chance of missing a connection due to a software crash or detection issue is very low–less than 1 in 1,000 scans.Our research at MIT suggests that these limitations have a minimal impact on identifying strong social connections.Even with some "background noise" from occasional errors, the data can still accurately capture close relationships based on frequent communication and Bluetooth proximity. However, it's important to consider that university communities might not be representative of the broader population [24].

Errors and inaccuracies might be more significant in other social settings.If further studies show that from errors the level of "background noise" is very high, there are ways to improve the accuracy of the data. One approach leverages the time information in the Bluetooth ID (BTID) logs.For example, someone briefly walking by another person would likely enter and leave the record at different times than members of a group actually interacting. Similar logic can be applied to identify other unusual patterns in the data.

4.7.1 Human-induced errors

This dataset includes two main types of human errors caused by phone usage: Phone turned off in which Users may intentionally turn off their phones or let the battery die.Our research suggests that, on average, participants reported letting their batteries die about 2.5 times per month. Additionally, one-fifth of the participants regularly turned their phones off in specific situations, like classes, movies, or most commonly, while sleeping [25].To minimize data loss when phones are turned off, the project timestamps the event right before the battery dies and pauses data collection. Phone restarts create a new times-

tamped record. However, another challenge emerges: misplaced phones. Our study revealed varying forgetfulness: 30 percent of participants never misplace their phones, 40 percent experience it once a month, and the remaining 30 percent misplace them once a week [26].

Distinguishing between a phone being on but forgotten and simply left at home or work is a significant hurdle.To address this, we developed a 'forgotten phone' classifier.This classifier relies on a combination of factors: location remaining constant for a period, phone being charged, and inactivity regarding calls, texts, and alarms. When tested on a portion of the dataset with labeled data, the classifier successfully identified days when the phone was forgotten. However, it also misclassified a day when a participant stayed home sick. Filtering out days when the phone is off ensures we only analyze data when the phone is with the user. However, this approach might discard valuable information from days when the phone is simply turned off. A more complex challenge is determining if a user briefly leaves their workplace without their phone. This seems to be a common occurrence for many participants, and there's no single clear way to definitively categorize this behavior.The good news is, as discussed in the evaluation section of the research, these occasional separations don't significantly impact the strong correlation between physical proximity and self-reported communication.Building on the findings from the association analysis, frequent physical proximity within the workplace does offer some insight. However, the real value lies in uncovering external interactions between participants. In these external settings, users are less likely to leave their phones behind.

4.8 Missing data

Since we know who participated in the study and when data was collected, we can identify missing information. There are two main reasons for missing data, as discussed earlier: data corruption and phones being turned off.The good news is that we have usable data for about 85.3 percent of the total time the phones were active.Only a small portion, around 5 percent, is due to data errors.The bulk of the missing 14.7 percent can be attributed to around 20 percent of participants powering off their phones at night[27].

4.9 Surveys and diaries vs. phone data

To validate the accuracy of our data in capturing social network dynamics, we asked students using Nokia 6600 phones to complete online surveys about their social interactions and collaborators throughout the day [28]. By comparing survey responses with the collected data, we found strong correlations between the frequency of self-reported communication and the number of logged Bluetooth IDs (correlation coefficient R=0.78, p-value=0.003).This suggests that more frequent Bluetooth connections aligned with participants reporting more communication. We also found strong correlations self-reported interactions between two people (dyadic data) and the corresponding data on physical proximity (dyadic immediacy data) (correlation coefficient R=0.74, p-value=0.0001).This indicates that people who reported interacting more often were also physically closer more frequently according to Bluetooth data.To ensure data accuracy, a smaller group of participants meticulously recorded their movements for several months. Analysis confirmed the information regarding physical closeness (immediacy) and location to be reliable, with the only gaps occurring when phones were powered off [29].

4.10 Research design and methodology

The Reality Mining project has three main goals:

Develop technology and algorithms to understand, model, and potentially influence human behavior.

Sense user behavior using phone sensors that collect data on location, proximity, communication, and phone usage.

Build models based on data collected from a pilot study involving 100 participants over eight months, representing roughly 500,000 hours (equivalent to 60 years) of human activity.

The study participants:

70 people from the MIT Media Lab 30 external students from the nearby MIT Sloan Business School Future plans:

Develop algorithms to create improved models of social networks. Explore ways to use proximity-based

prompts to influence real-world social interactions.

4.11 Continuous Bluetooth scanning

The fact that most modern smartphones have built-in short-range RF networks, like Bluetooth or Wi-Fi Direct, presents a promising opportunity for example, This research leverages two features of modern phones to pinpoint location and activity. Cellular networks (like GSM) and Bluetooth. The Cell towers transmit signals to phones, and based on the nearest tower a phone connects to, its general location can be estimated. This is similar to how cell phone calls work. While Bluetooth is a shorter-range wireless technology used for connecting devices like phones and laptops within a few meters. Reality Mining utilizes Bluetooth in a new way.

4.11.1 How Bluetooth helps with location tracking

Every Bluetooth device can "see" other Bluetooth devices nearby. When a phone detects another Bluetooth device, it collects some information: Like a fingerprint, every device has a unique BTID. Users can give their devices names (e.g., "Tony's Nokia").

4.11.2 Device type

A code indicating the type of device (phone, laptop, etc.).While Bluetooth adoption was slow initially, it's now common in phones. This makes BTID data valuable for understanding how people interact with each other based on their proximity [30].

While Bluetooth was originally designed to connect phones to headsets and laptops, it has an unexpected benefit. Bluetooth devices can detect other nearby Bluetooth devices. This research takes advantage of this by using a software program called Blue Aware. Blue Aware runs silently in the background on compatible phones and collects data whenever the phone is on. It detects and logs the unique identifiers (BTIDs) of other Bluetooth devices nearby.It timestamps these encounters to create a record of physical proximity.

This is similar to the Jabberwocky project [31], but with a key difference. Jabberwocky focused on desktop computers, while Blue Aware leverages the always-on nature of Bluetooth in phones to continuously collect data. The study referenced in [31] continuously transmitted newly discovered Bluetooth IDs, which could drain a phone's battery in about 18 hours.

While constant monitoring provides richer data, most users expect their phones to last longer than that. To address this, Blue Aware: Searches for nearby devices every five minutes: This strikes a balance between data collection and battery life, ensuring standby times exceeding 36 hours for most phones. Alerts users at startup: A notification informs users that Blue Aware is running in the background. Provides a user interface: Users can see the data being collected, choose to remove specific data points, or disable logging entirely. While Blue Aware runs on phones, Bluedar is a separate device designed for public spaces.

Bluedar continuously searches for nearby Bluetooth devices. It transmits the discovered Bluetooth IDs (BTIDs) wirelessly to a server over a Wi-Fi network (802.11b). Bluedar is a Bluetooth beacon built using a Class 2 Bluetooth chipset. It connects to the internet via an 802.11b wireless bridge housed in a discreet box. The system is controlled remotely through a web server. Bluedar's Class 2 Bluetooth chipset has a wider range than phones, typically detecting devices within 25 meters. The project is currently exploring the use of Bluedar data to develop a "proximity-based preface package" (the purpose of this package is not explained in the provided text).

4.12 Cell tower probability distributions

Many researchers have explored using cell tower IDs to pinpoint user location. For example, Lausanne ET AL. proposed a method for estimating positions based on cell tower data. However, there are significant challenges. Cell phones can detect towers from miles away, leading to imprecise location estimates. In cities, phones can be in range of many towers simultaneously, making it difficult to determine the exact location. Including signal strength data can improve accuracy, but signal can be distorted by reflections (multi path deformation). Even at the same location, phones can connect to different towers based on factors like signal strength and network traffic.

The Reality Mining project addressed these limitations by using Bluetooth data. Time spent in one loca-

tion by analyzing how long a phone stays connected to a tower, we can estimate the likelihood of it being the true location. We only considered cell towers identified when a stationary Bluetooth device was also detected, ensuring the user was within a 10-meter radius. This approach combines cell tower data with Bluetooth verification for a more accurate understanding of user location.

The study found limitations in using cell tower data to pinpoint user location within a 10-meter radius. Users in the same office (users 2 and 4) had similar cell tower patterns despite spending different amounts of time there. Users in a different area (users 1 and 5) had fewer unique cell towers than those in the office. User 3, in a separate office, had a mix of cell tower patterns from the other two groups. While cell tower mapping techniques are improving, they may not always meet the high accuracy needs of some location-based applications. The project explored using stationary Bluetooth device IDs as an additional location indicator. This approach significantly improved user localization, especially indoors. Buildings with weak cell signals often have many stationary Bluetooth devices (like computers). Overall, participants lacked cell reception 6 percent of the time but were still within range of a Bluetooth device or another phone 21 percent and 29 percent of the time, respectively. The researchers expect Bluetooth coverage to increase as the technology becomes more widespread in devices[32].

4.13 Privacy Implications

Concerns about privacy are understandable when studying data from real people. However, this research was conducted ethically with informed consent from all 100 participants. The project envisions a future where phones have more powerful processors and can analyze data directly on the device. In this scenario, insights could be generated in real-time without needing to send data elsewhere. Unfortunately, current phone technology can't run the complex models needed for such on-device analysis. Therefore, this research focuses on demonstrating the potential of phone-collected data, not creating a deployable system outside a research setting.

5 EXPERIMENTS

We have selected an open-source tool for applying data mining techniques due to its extensive functionality. This tool allows us to extract data for testing or to perform abstract-level analysis by quickly developing clusters from the available data. It is particularly effective for analyzing large datasets in a short amount of time. For this tutorial, we will use Weka version 3.6.6 to demonstrate how to analyze data using this tool. For Weka, the JDK needs to be installed on your machine. We have installed the latest version of the JDK for this purpose. The installation process for Weka is straightforward. Simply go to http://sourceforge.net/projects/weka/files/weka-3-6-windows-jre/3.6.6/weka-3-6 6jre.exe/download and download the .exe file. This installer will automatically install Weka along with the latest version of the JDK.

I have not installed JDK on my machine so I am using this version of Weka. Alternatively, a version of Weka is available that does not include the JDK installation.

If you already have the JDK installed on your machine, you can download this version from http://sourceforge.net/projects/weka/files

/weka-3-6- windows/3.6.6/weka-3-6.exe/download . However, we will proceed with the .exe file that includes the JDK. Double-click on the .exe file to start the installation process. Simply press the NEXT button. And then after reading the terms and conditions click on —I agree. Then on the next screen from the drop-down list select the —full and make sure that —Associate files| and —install jre should be checked(for jdk installation). And press —NEXT. Now select the folder where you want to save it by clicking —Browse button. I have selected the default which was given and gone to the next screen. the installation process begins when you click "Install." A progress bar window will open, followed by a command prompt for the JDK installation. Press "Install" to proceed with the JDK installation. Once the JDK installation is complete, click "Finish" to close the window. The Weka installation will resume; click "Next" to proceed and then "Finish" to close the installation window. To start Weka, go to the Start menu and select Weka. When Weka opens for

the first time, a pop-up screen will appear. Choose "Explorer" from this screen to access the data exploration functionality. Here, you can analyze the type of data and determine how to extract the necessary information from the available data.

Figure 4. Weka Pop Up Screen

When you first open WEKA, a pop-up screen will appear. Select "Explorer" from this screen. On the next screen, choose "Open file" and select a .csv file, which we have obtained from MySQL.The data was originally in .sql format, and we converted it to .csv format to use it in WEKA for analysis.I used another open-source tool, WAMP server, to utilize its MySQL.The file was imported into MySQL and then exported as a .csv file. WEKA supports .csv and .arff file formats.Now, we have several .csv files derived from the original .sql file, as it contained multiple tables with millions of records. The challenge faced was that WEKA does not have the capacity to load large files into memory.

Our initial dataset files were likely too large for efficient analysis. We reduced their size (unspecified method) before loading them. While the results were acceptable, Weka offers a wider range of possibilities. As mentioned earlier, Weka supports both .arff and .csv file formats. Here are the specific files we obtained after processing the .sql schema: activityspan.csv, cellspan.csv, callspan.csv (used in the example below), cellname.csv, celltower.csv, coverspan.csv, device.csv, devicespan.csv, person.csv, phonenumber.csv, singlesubject.csv.

We can use Weka to explore and analyze this prepared data. Weka offers various algorithms for data

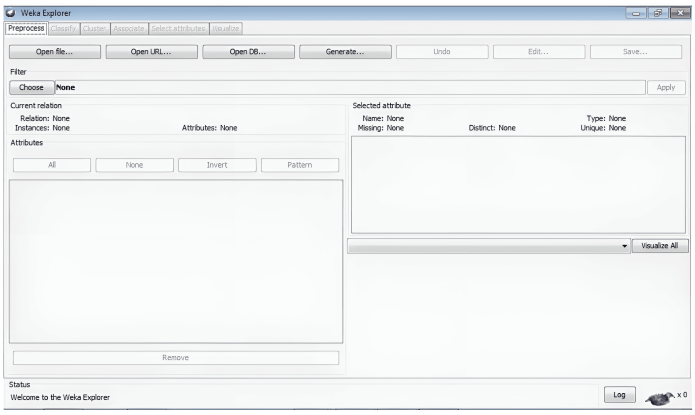

Figure 5. Weka Explorer Screen

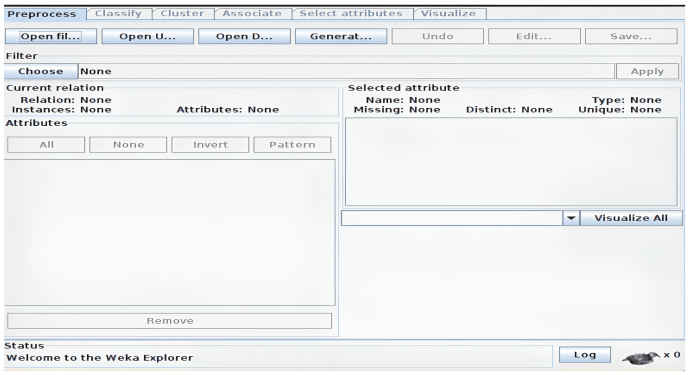

Figure 6. Weka Open file screen

manipulation and visualization. These algorithms help us extract specific information and present it in a way that's easier to understand. This example focuses on the callspan.csv file, which contains call data for individuals. It shows the call start and end times. By selecting attributes from the left-hand side in Weka's interface, we can see how their values affect the data displayed on the right.

In the figure above we have chosen the clustering algorithm (EM-I 100 –N –I –M 1,0E-6 –S 100), this is chosen by clicking the choose button from the cluster tab. Then select the option —Use training set and select the option Store clusters for visualizations and then click start to start visualization. Here, we can observe the specific times and duration during which customers are using a particular network service.

The figure above illustrates the data in cluster form, where the data is denser. Empty slots indicate

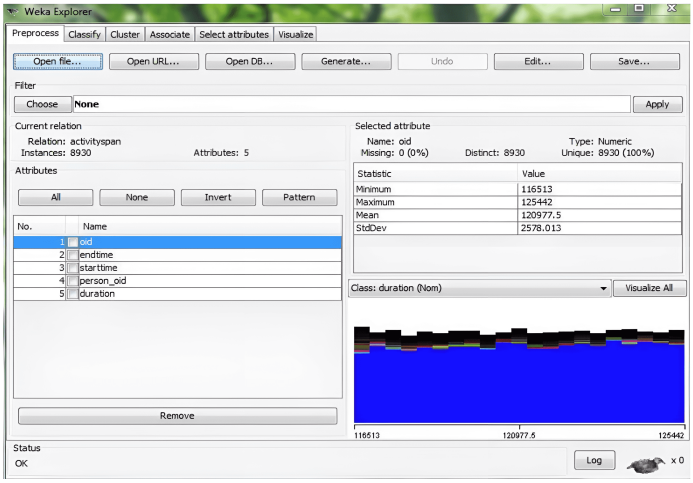

Figure 7. Weka choose filter screen

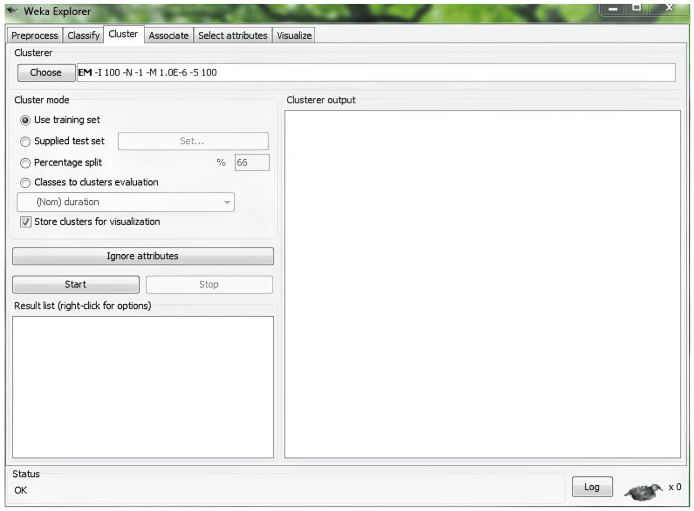

Figure 8. Weka Cluster tab Screen

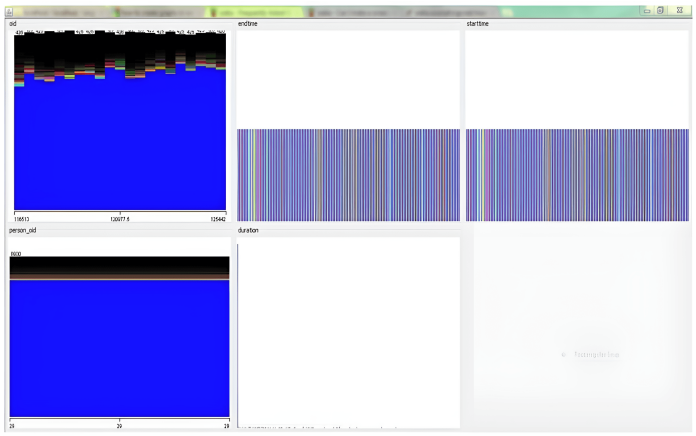

Figure 9. End Time Start Time Screen

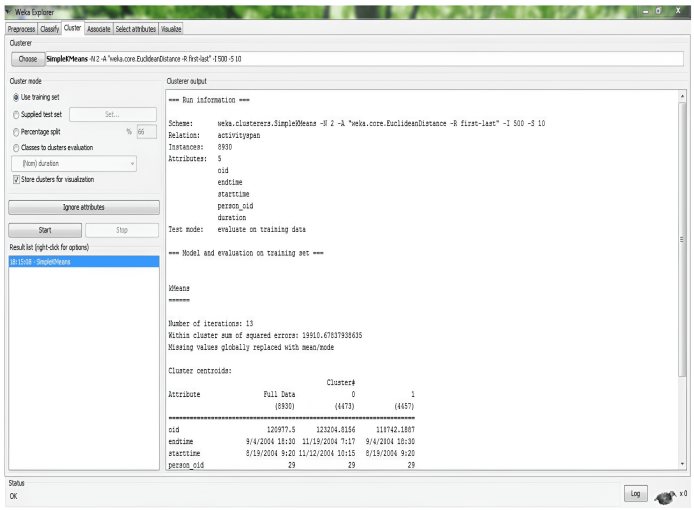

Figure 10. Cluster Output Screen

errors; however, the given data has a very low error rate. Most of the data is organized into well-defined clusters. When data is highly organized with no gaps, it is considered dense. Conversely, empty areas indicate errors in the data. Some data may be scattered and not part of any cluster. When making decisions, we focus on clusters where the maximum number of customers are concentrated.

This is an example of applying an algorithm to analyze our data and aid in decision-making. The details of the K-Means algorithm were discussed in previous sections. Essentially, it provides a summary: if a certain condition is met, then a specific result follows.

The following is a refined summary of the results obtained by applying the K-Means algorithm to our data. Different clusters have been formed according to this algorithm. Errors (empty values) are filled by taking the mean or mode of the data.

This analysis focuses on the closeness of the data for a single attribute, although it applies to the entire dataset. It may be other attributes. We have analyzed this attribute along with others in the dataset to understand the conditions under which data forms clusters. This observation holds true for all attributes except for person ID.

In the figure above, different headers are displayed, including person ID, end time, and duration. Person ID

```
kMeans
======

Number of iterations: 13
Within cluster sum of squared errors: 19910.67837938635
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute               Full Data             0                1
                          (8930)          (4473)           (4457)
=============================================================================
oid                     120977.5      123204.8156      118742.1887
endtime          9/4/2004 18:30  11/19/2004 7:17   9/4/2004 18:30
starttime        8/19/2004 9:20 11/12/2004 10:15   8/19/2004 9:20
person_oid                  29               29               29
duration               0:00:00          0:00:00          0:00:00


Clustered Instances

0       4473 ( 50%)
1       4457 ( 50%)
```

Figure 11. Data Screen

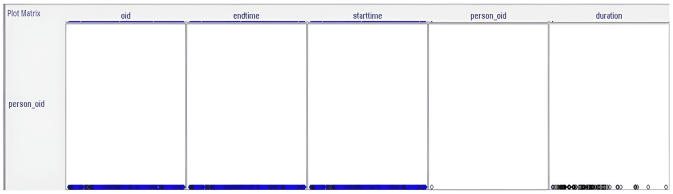

Figure 12. Plot matrix screen

serves as a unique identifier for each individual, similar to a primary key in databases, ensuring each user has a distinct and non-duplicated identifier. OID represents the unique call ID assigned to each call, ensuring each call has a specific identifier. Start time indicates when a call is initiated between users, while end time denotes when the call is terminated.

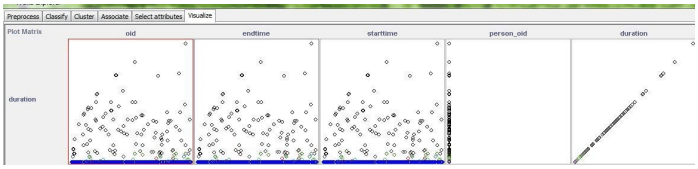

Figure 13. Visualize screen

Based on this duration of the call has been calculated initially, the data was dispersed and inconsistent, lacking uniformity and stability.

However, the graphs indicate a steady increase in usage over time. As previously mentioned, the "old" represents the call ID, "end time" represents the call's

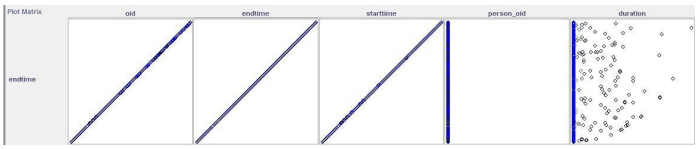

Figure 14. Visualize screen 2

end time, and "start time" represents the call's start time. The linear patterns in the graphs suggest an increase in usage.

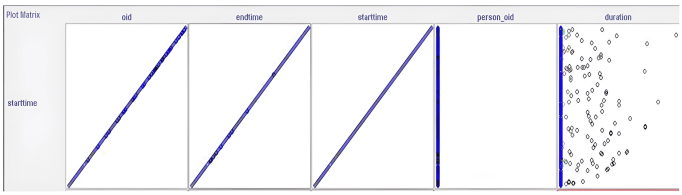

Figure 15. Visualize screen 3

Similar to earlier observations, these graphs confirm a gradual and progressive rise in usage.The graphs with linear trends depict correlations among the attributes within that particular dataset. Observing the graphs, the abundance of blue and red points indicates a high volume of customers at those times. When correlated with time and other factors, it shows that there are many callers during those periods. Additionally, the numerous scattered points suggest that these callers are not regular. From a management perspective, this information helps in making informed decisions. For instance, if a telecommunications company plans to launch a new package to attract more callers, it can focus on areas of interest identified from the clusters. Furthermore, if the company needs to back up its data, it can determine the times when service utilization is minimal to minimize disruption. By analyzing clusters and scattered points, management can decide the optimal time for backups. Utilizing such techniques allows the company to make quick and efficient decisions to enhance its services. Manually searching through millions of records is time-consuming and impractical in today's business environment. While we have used caller timings as an example, similar analyses can be conducted on other datasets as needed. This approach provides a

high-level overview of the data, and for more detailed insights and critical analysis, further exploration can be conducted. By clicking on these clusters we will go in more depth and we will be clear more about those points and their specs.

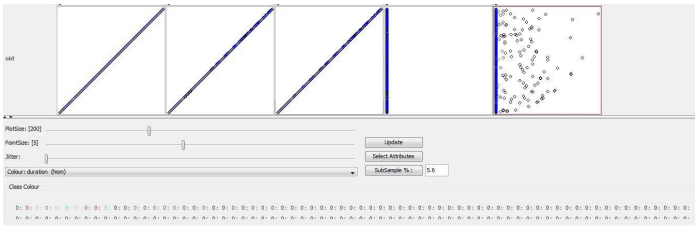

Figure 16. Visualize screen 4

The graph above indicates that the data exhibits uniformity without any variation or jitter, suggesting a consistent pattern or similarity.

6 CONCLUSION

Extracting and analyzing information from various perspectives to derive useful insights is known as data mining. Using different data mining techniques, we analyzed the MIT Reality Mining dataset, which was sourced from MIT and provided in SQL format. This dataset contains usage data for one hundred users, including their call timings and durations. We selected an open-source tool to import the datasets and perform the analysis. The data was converted to .csv format and then to.arff format, which is compatible with Weka. These converted files were then analyzed in Weka to identify variations and disparities using data mining techniques. By applying different data mining methods, we were able to clean the dataset of various errors, observe the increase and growth in usage, and track the gradual and steady increase in data, which was initially scattered and inconsistent. Different graphs were drawn to illustrate the data scattering, gradually increasing usage, and variations in data. This level of analysis would not have been possible without data mining techniques, which are essential for extracting useful information from large databases and making quick decisions. Using the MIT Reality Mining dataset and various data mining techniques, we found that the Weka tool was very effective in highlighting the key features and characteristics of the telecom industry. Specifically, the Weka tool helped us distinguish the major aspects of the telecom industry. Our analysis focused on user concerns, such as how long users stay connected to the network, the duration of calls, and the length of time consumers use the allocated bandwidth.

6.1 Future Work

We leveraged the data mining capabilities of Weka to uncover key user concerns within the telecom industry. We analyzed how long users connect to the network, maintain calls, and utilize the allocated bandwidth. In the future, our work can be expanded to evaluate different types of networks under various scenarios. This will enable companies to make informed decisions quickly. In this section (and others as necessary), we detail how we compared the software, with relevant subsections. Additionally, we may need to explain why certain software was not included in our analysis and how it could be incorporated in future studies.

Author Contributions

Daud Khan: advances network intelligence by introducing cutting-edge data mining and learning techniques. His work enhances knowledge management and optimizes network performance, providing practical solutions for modern network challenges **Waqas Ahmad , Haseeb khan**: Data Creation, Writing-Original draft preparation. **Daud Khan , Muhammad Abrar Khan**: Visualization, Investigation. **Daud Khan, Zain shaukat**: Supervision.: **Waqas Ahmad, Shahab Ul Islam**: Software, Validation. **Zain shaukat, Shahab Ul Islam**: Writing- Reviewing and Editing.

Compliance with Ethical Standards

It is declared that all authors don't have any conflict of interest.Furthermore, informed consent was obtained from all individual participants included in the study.

References

[1] Gupta, Manoj Kumar and Chandra, Pravin, "A comprehensive survey of data mining," *International Journal of Information Technology*, p 12,1243–1257,2020.

[2] Liu, Kunpeng and Fu, Yanjie and Wu, Le and Li, Xiaolin and Aggarwal, Charu and Xiong, Hui, "Automated feature selection: A reinforcement learning perspective,"

IEEE Transactions on Knowledge and Data Engineering, p. 1063112272–2284, 35, 2021.

[3] Dogan, Alican and Birant, Derya, "Machine learning and data mining in manufacturing," *Expert Systems with Applications*, vol. 166,p 114060,2021.

[4] Padmanaban, K and Senthil Kumar, AM and Azath, H and Velmurugan, AK and Subbiah, Murugan, "Hybrid data mining technique based breast cancer prediction," *Subbiah, Murugan, booktitle= AIP Conference Proceedings* vol. 2523, 2023.

[5] Sadeghi, Sanaz and Soltanmohammadlou, Nazi and Nasirzadeh, Farnad, "Applications of wireless sensor networks to improve occupational safety and health in underground mines," *Journal of safety research*, p 8-22,83,2022.

[6] Wu, Wen-Tao and Li, Yuan-Jie and Feng, Ao-Zi and Li, Li and Huang, Tao and Xu, An-Ding and Lyu, Jun, "Data mining in clinical big data: the frequently used databases, steps, and methodological models," *Military Medical Research*, p 8,1–12,2021.

[7] Jassim, Mustafa Abdalrassual and Abdulwahid, Sarah N, "IOP conference series: materials science and engineering," *IOP Publishing*, p 1090,012053,2021.

[8] Shu, Xiaoling and Ye, Yiwan, "Knowledge Discovery: Methods from data mining and machine learning," *Social Science Research*, p 110,102817,2023.

[9] Abdallah, Emad E and Otoom, Ahmed Fawzi and others, "Intrusion detection systems using supervised machine learning techniques: a survey," *Procedia Computer Science*, p 201,205–212,2022.

[10] Zhong, Yong and Chen, Liang and Dan, Changlin and Rezaeipanah, Amin, "A systematic survey of data mining and big data analysis in internet of things," *International Journal of Information Technology*, p 78,18405–18453,2022.

[11] Sunhare, Priyank and Chowdhary, Rameez R and Chattopadhyay, Manju K, "Internet of things and data mining: An application oriented survey," *Journal of King Saud University-Computer and Information Sciences*, p 34,3569–3590,2022.

[12] Wu, Wen-Tao and Li, Yuan-Jie and Feng, Ao-Zi and Li, Li and Huang, Tao and Xu, An-Ding and Lyu, Jun, "Data mining in clinical big data: the frequently used databases, steps, and methodological models," *Military Medical Research*, p 8,1-12,2021.

[13] Fadelelmoula, Ashraf Ahmed, "Exploiting Cloud Computing and Web Services to Achieve Data Consistency, Availability, and Partition Tolerance in the Large-Scale Pervasive Systems," *International Journal of Interactive Mobile Technologies*, p 15,15,2021.

[14] Saeed, Nasir and Ahmad, Waqas and Bhatti, Dost Muhammad Saqib, "Localization of vehicular ad-hoc networks with RSS based distance estimation," *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, p 15,1-6,2018.

[15] Shu, Xiaoling and Ye, Yiwan, "Knowledge Discovery: Methods from data mining and machine learning," *Social Science Research*, p 110,102817,2023.

[16] Abd Elaziz, Mohamed and Abualigah, Laith and Attiya, Ibrahim, "Advanced optimization technique for scheduling IoT tasks in cloud-fog computing environments," *Future Generation Computer Systems*, p 12,142–154,2021.

[17] Ahmad, Waqas and Husnain, Ghassan and Ahmed, Sheeraz and Aadil, Farhan and Lim, Sangsoon and others, "Received signal strength-based localization for vehicle distance estimation in vehicular ad hoc networks (VANETs)," *Journal of Sensors*, p 2023 ,2023.

[18] Abd Elaziz, Mohamed and Abualigah, Laith and Attiya, Ibrahim, "Predicting Students' Performance Employing Educational Data Mining Techniques, Machine Learning, and Learning Analytics," *International Conference on Communication, Networks and Computing*, p 12,166–177,2022.

[19] Al-Hawari, Assem and Najadat, Hassan and Shatnawi, Raed, "Classification of application reviews into software maintenance tasks using data mining techniques," *Software Quality Journal*, p 30,667–703,2021.

[20] Amanowicz, Marek and Jankowski, Damian, "Detection and classification of malicious flows in software-defined networks using data mining techniques," *Amanowicz, Marek and Jankowski, Damian*, p 21,2972,2021.

[21] Ahmad, Waqas and Ahmed, Sheeraz and Sheeraz, Najia and Khan, Ayub and Ishtiaq, Atif and Saba, Malka, "Localization Error Computation for RSSI Based Positioning

System in VANETs," *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)*, p 12,1–6,2020.

[22] Hu, Lun and Pan, Xiangyu and Tang, Zehai and Luo, Xin, "A fast fuzzy clustering algorithm for complex networks via a generalized momentum method," *IEEE Transactions on Fuzzy Systems*, p 21,3473–3485,2021.

[23] Amanowicz, Marek and Jankowski, Damian, "Detection and classification of malicious flows in software-defined networks using data mining techniques," *Amanowicz, Marek and Jankowski, Damian*, p 21,2972,2021.

[24] Ullah, Tahz and Hussnain, Engr Ghasssan and Ahmad, Waqas and Sikander, Gulbadan and Ashfaq, Muniba, "An efficient machine learning based multiclass cyber attacks classification and prediction," *The Sciencetech*, p 4,,2023.

[25] Aksan, Fachrizal and Jasiński, Michał and Sikorski, Tomasz and Kaczorowska, Dominika and Rezmer, Jacek and Suresh, Vishnu and Leonowicz, Zbigniew and Kostyaw, "Clustering methods for power quality measurements in virtual power plant," *Energies*, p 14,5902,2021.

[26] Menegazzo, Jeferson and Von Wangenheim, Aldo, "Road surface type classification based on inertial sensors and machine learning: A comparison between classical and deep machine learning approaches for multi-contextual real-world scenarios," *Computing*, p 103,2143–2170,2021.

[27] Khan, Majid and Husnain, Ghassan and Ahmad, Waqas and Shaukat, Zain and Jan, Latif and Haq, Ihtisham Ul and Islam, Shahab Ul and Ishtiaq, Atif, "Performance evaluation of Machine Learning models to predict heart attack," *Machine Graphics and Vision*, p 32,99–114,2023.

[28] Kim, Hae Reong and Sung, MinDong and Park, Ji Ae and Jeong, Kyeongseob and Kim, Ho Heon and Lee, Suehyun and Park, Yu Rang, "Analyzing adverse drug reaction using statistical and machine learning methods: A systematic review," *Medicine*, p 101,e29387,2022.

[29] Khan, Rizwan and Jan, Latif and Khan, Shahid and Zafar, Mohammad Haseeb and Ahmad, Waqas and Husnain, Ghassan, "An effective algorithm in uplink massive MIMO systems for pilot decontamination," *Results in Engineering*, p 101873,2024.

[30] Tiwari, Neelu and Singh, Naveen Kumar and Singh, Rajni and Rameshwar, Rudra, "Identifying potential churners through predictive analysis: evaluation using pro-active-attrition management logistic regression," *International Journal of Technology Transfer and Commercialisation*, p 18,439–461,2021.

[31] Edastama, Primasatria and Dudhat, Amitkumar and Maulani, Giandari, "Use of Data Warehouse and Data Mining for Academic Data: A Case Study at a National University," *International Journal of Cyber and IT Service Management*, p 1,206–215,2021.

[32] Hou, Rong and Ye, Xu and Zaki, Hafizah Binti Omar and Omar, Nor Asiah Binti, "Marketing decision support system based on data mining technology," *Applied Sciences*, p 13,4315,2023.